OECD Digital Education Outlook 2026

Exploring Effective Uses of Generative AI in Education

OECD Digital Education Outlook
2026

EXPLORING EFFECTIVE USES OF GENERATIVE AI IN EDUCATION

OECD
BETTER POLICIES FOR BETTER LIVES

LAUNCH VERSION

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Member countries of the OECD.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Note by the Republic of Türkiye
The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union
The Republic of Cyprus is recognised by all members of the United Nations with the exception of Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Photo credits: Cover © Gerhard Richter 2025 (23122025).

Corrigenda to OECD publications may be found at: https://www.oecd.org/en/publications/support/corrigenda.html.

Editorial

The *OECD Digital Education Outlook* is the OECD's flagship publication presenting our latest analysis of emerging digital technologies in education.

This 2026 edition synthesises evidence and expert insights to show how generative AI has the potential to transform the quality and effectiveness of learning, as well as the productivity of education systems, provided its associated risks are carefully managed. Its applications include enhancing student learning, supporting teachers' performance while preserving professional autonomy, and strengthening education systems, as well as institutional and research capacities.

For students, generative AI can scale personalised learning through intelligent tutoring systems, including in low-infrastructure settings. Generative AI can also support knowledge acquisition by enabling collaborative learning and enhancing creativity.

However, evidence shows that overreliance on generative AI tools that provide direct answers can reduce students' active engagement, improving task performance without corresponding learning gains. When used as a shortcut rather than a learning tool, generative AI can displace cognitive effort and weaken the skills that underpin deep learning.

For example, a field experiment in Türkiye found that while access to GPT-4 improved short-term performance – by 48% with the standard interface, and by 127% with a tutoring version designed to support learning – students performed 17% worse once access was removed, showing that generative AI can undermine learning unless explicitly designed to support skill acquisition.

For teachers, generative AI can improve both productivity and teaching quality. Evidence cited in the report shows a 31% reduction in time spent on lesson and resource planning by secondary science teachers in England, and a 9-percentage-point increase in student pass rates when low-experience tutors used AI support, with smaller gains for more experienced tutors.

According to the OECD's 2024 Teaching and Learning International Survey, 37% of teachers already use generative AI for work-related tasks - such as learning about or summarising topics and supporting lesson planning - with substantial variation across countries.

At the same time, concerns persist that overreliance on AI could undermine teacher autonomy and professionalism, raise ethical risks, and, when used extensively for tasks such as marking, feedback or lesson planning, erode teachers' professional skills.

The report calls for a shift towards educational generative AI systems designed with teachers, enabling them to monitor students' interactions with generative AI and actively shape its use in learning.

At the system level, generative AI can improve the efficiency of education systems and school management by automating and supporting administrative and analytical processes. It can help develop standardised assessment items, review curricular alignment by analysing actual versus expected student workload, enhance study and career guidance, and support the classification of educational resources to name just a few. Generative AI can also have potentially transformative implications for education research, as in other fields.

To realise this potential, policymakers will need to mitigate and manage associated risks - such as those related to access, data privacy, ethics and bias – through sound policy frameworks and effective governance.

The OECD supports policymakers in making effective and responsible use of generative AI in education. This includes promoting approaches that place human judgement, feedback and oversight at the centre of AI use; strengthening teachers' capacity to engage with AI confidently and effectively; and providing clear, practical guidance on the appropriate use of generative AI in education. The OECD can also foster international co-operation and the exchange of good practices, enabling peer learning across jurisdictions, so that generative AI delivers on its full potential for better learning and more effective education systems.

Mathias Cormann

Secretary-General, OECD

Acknowledgments

Table of contents

TABLES

Executive Summary

Generative artificial intelligence (GenAI) is rapidly entering education systems worldwide, raising expectations of more personalised learning, enhanced teaching practices, and more efficient system management. The *OECD Digital Education Outlook 2026* draws on the best available empirical research, design experiments, and expert insights to explore where GenAI shows promise, and how education stakeholders can steer its effective and responsible adoption.

Evidence shows that GenAI can scale personalised learning support, enhance feedback quality, and automate parts of assessment. But this convenience can come at a cost. When students depend too heavily on GenAI, metacognitive engagement – the mental processes and effort that turns answers into understanding – drops. This results in a misalignment between task performance and genuine learning (chapters 1 and 2).

While some studies show both improved student outputs and learning, others do not, particularly when tools provide direct solutions rather than supporting true learning processes. Effectively integrating GenAI into teaching and learning may require that teachers encourage student agency and emphasise process, such as how students think and learn, rather than student output. Hybrid systems that combine GenAI with explicit pedagogical models, such as structured tutoring strategies or evidence-centred assessment design, show more promise than general-purpose chatbots (chapter 2).

Enhancing student learning with generative AI

One of the most striking uses for GenAI is tutoring. Unlike the rigid dialogue trees of traditional AI tutors, GenAI can hold flexible, personalised conversations, adapting explanations and language to individual learners' needs. Some AI tutors use methods like Socratic questioning to develop subject knowledge, critical thinking and reflection. The evidence is still emerging, but prototypes show promise (chapter 3).

Beyond one-on-one tutoring, GenAI is supporting collaborative learning. Studies identify four main roles: acting as an information hub, generating personalised materials to support group work, providing feedback to teachers, and acting as a peer contributor in group tasks. While evidence so far is limited, some studies find small-to-medium improvements in subject learning and large ones in critical thinking and teamwork (chapter 4).

GenAI may also support creativity. Evidence suggests it is most beneficial when used slowly, to support iterative exploration and reflection as opposed to churning out instant content (chapter 5). In this sense, it can also undermine creativity by reducing original thought. Importantly, GenAI has the potential to support students in places with limited digital infrastructure. A large-scale experiment in rural Brazil showed that even with intermittent connectivity and minimal equipment, AI could provide feedback and guidance. Small language models running offline on mobile devices represent a promising avenue for GenAI to bridge digital divides, despite their technical limitations (chapter 6).

Augmenting teachers' performance with generative AI

GenAI promises to drastically change the way teachers work in other ways too, including boosting productivity and the quality of teaching. It can already quickly write summaries, design exercises and even offer real-time tutoring support. But there is a risk that overreliance on GenAI could lead to the loss of skills and teaching expertise. A conceptual framework on how humans and AI can work together offers three paths: replacement, complementarity and augmentation. Replacement of some tasks should be assessed carefully to avoid loss of teacher-student interactions. Complementarity is better, pairing human judgment with machine efficiency. But the most effective approach is augmentation through

collaborative engagement. In this model, teachers and AI work in tandem, critiquing and refining each other's outputs. This iterative process offers the greatest potential for improved instructional quality while preserving professional judgement (chapter 7).

One of the key issues at the moment is that most tools are designed for general use. Off-the-shelf chatbots rarely align with curricula. That is why some argue for purpose-built educational GenAI systems. These tools can be co-created with teachers and students, giving educators control over how machines behave and how students interact with them (chapter 8). For example, this could enable teachers to set the level of "hallucinations" of the tools and give feedback on their student GenAI interactions.

Several GenAI tools are already being used to support teachers, especially in the higher education context. For example, some AI teaching assistants can help teachers, teaching assistants, and students across a wide range of instructional tasks while allowing human oversight. Students rated one such tool as comparable to human teacher assistants in clarity, accuracy and professionalism, though weaker in motivation and developmental guidance (chapter 9).

Other early evidence suggests that educational GenAI tools can improve online tutoring quality, especially for less experienced teachers. Research also highlights the benefits of AI-generated teaching materials and analytics for effective classroom dialogue. Yet motivation, relationships, and social-emotional learning remain inherently human responsibilities (chapter 10).

Improving system and institutional management

GenAI is also streamlining system and institutional management, enabling new forms of classification and recommendations. At the institutional level, GenAI is already reshaping administrative tasks. Embedding-based models can map equivalencies between courses and programmes, making tasks like admissions, career guidance and curriculum analytics faster and more accurate. Large-scale pilots demonstrate high predictive accuracy and efficiency gains, although human AI collaboration remains a must (chapter 11).

Beyond feedback, high-stakes standardised assessment is another field where GenAI promises changes. It can generate exam items at scale and design more authentic tasks, such as interactive writing and speaking tasks that mimic real-life communication. By teaming up with AI, teachers can achieve significant productivity gains (chapter 12). GenAI's impact on research is also notable. In natural sciences, it accelerates everything from hypothesis generation to experimental design. The technology is already changing how education research is performed and will potentially improve education systems' outcomes (chapter 13). For example, AI-generated synthetic datasets simulating real education datasets could expand research possibilities and feedback into policy and practice.

Ultimately, when designed with strong pedagogy and a human-centred approach, GenAI can do far more than help students complete tasks. It has the potential to deepen student learning, improve teaching practice and streamline institutional management and research. But these benefits come with risks. Overreliance risks turning students into passive consumers and teachers into supervisors. To unlock GenAI's full potential, education must move beyond generic chatbots towards purpose-built tools for education. The thoughtful integration of general-purpose GenAI tools will be essential – for realising the full learning benefits of GenAI and developing students' GenAI literacy for their future careers. The challenge for policymakers is to ensure that GenAI is a learning partner and not a learning shortcut.

1

Exploring effective uses of generative artificial intelligence in education:
An overview

This chapter presents an overview of the findings of the *OECD Digital Education Outlook 2026*. After a presentation of generative AI (GenAI) and of its uptake in society and education, the chapter shows how research and development on GenAI can inform policy and practice in education. It argues that general-purpose GenAI carries risks for learning, and that it must be used with pedagogical purpose or redesigned as specific educational GenAI tools. A number of educational GenAI tools and their functionalities are presented as examples. GenAI can also support educational workflows within education institutions and systems and present new opportunities for educational research.

This report examines generative AI (GenAI), a transformative technology that brought artificial intelligence into the public spotlight, including for students and education policymakers, following the launch of OpenAI's ChatGPT in 2022. Unlike earlier educational AI systems, GenAI is available and used by students outside of educational institutions, with or without the blessing of teachers, school leaders and policymakers. This presents both significant opportunities and complex challenges for education. After clarifying what is meant by GenAI, this chapter gives an overview of the uptake of GenAI among OECD populations, including students and teachers. It then provides a summary of the knowledge and information in this *OECD Digital Education Outlook 2026*: research evidence on the effects of GenAI on student learning, examples of what educational GenAI could look like, and possible uses to improve workflows at the institution and system levels.

Generative AI in education

What is generative artificial intelligence?

GenAI is a subset of AI focused on producing new content such as text, pictures, videos, songs, mathematical equations, computer programmes, typically in response to a question or command ("prompt").[1] These outputs are generated based on large volumes of training data. To do this, GenAI relies on advanced machine-learning techniques, such as neural networks based on transformers (notably Generative Pre-trained Transformer (GPT)), embeddings, tokens, etc. Most people have experienced GenAI via chatbots based on large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, Microsoft's Co-pilot, Anthropic's Claude, Mistral's LeChat or Deepseek's Deepseek.

In contrast, non-generative AI systems mainly produce predictions, classifications, recommendations, and ratings, for example for movies, books or other products and services. While they may use similar techniques as GenAI, their primary goal is to identify patterns and relationships in vast amounts of data, rather than create new content. Those AI systems are sometimes referred to as "rule-based", "predictive", or "good old-fashioned" AI. Despite often being less visible to end users, these systems are still powerful and have a variety of uses, including in education. They are

embedded in assistive technologies, for example for students with special needs, used to adapt learning to personal needs within intelligent tutoring systems, to score assessments or to predict whether students are at risk of dropping out (OECD, 2021[1]).

An important distinction should be made between AI tools that are general-purpose and those that are specialised (in our case, mainly educational): general-purpose systems are versatile and designed to serve many purposes, including educational ones, whereas specialised educational tools are designed for educational purposes only (see Table 1.1).

Table 1.1. Examples of different categories of AI

	Non-generative	Generative
General-purpose	Speech-to-text, Text-to-speech Note-taking tools (image-to-text) AI translation software	Chatbots (e.g. ChatGPT, Deepseek, Gemini) Image, video or sound generators
Educational	Intelligent Tutoring Systems (e.g. Assistments, Lalilo, PILA) Early Warning Systems Simulations (AR/VR)	GenAI tutors (e.g. Gauth, Khanmigo, Question AI, Socratic Playground) AI Teacher Assistants (e.g. JeepyTA, Coteach, CoTutor)

What is so special about general-purpose GenAI tools?

General-purpose GenAI tools often provide pertinent and contextualised answers to questions, with the ability to clarify and ask follow-up questions. These capabilities were not possible with earlier (non-generative AI) natural language processing. They are trained on massive data sets that exceed what humans could retrieve manually. Moreover, they are flexible and can be applied to many different subjects. Contrary to most educational AI, general-purpose GenAI tools usually offer free versions, enabling students and teachers to use them even if they are not provided by universities or schools, assuming they have an adequate device and connectivity. Even offline, small language models can run, albeit with lower performance (Isotani, 2026[2]).

A series of well-known shortcomings are also specific to current GenAI systems and inherent to its technology. Because they are based on probabilistic models, they can "hallucinate", that is, produce a plausible but wrong answer or fabricate details of an output. They do not generate consistent results over time. For example, repeating the same task several times will yield (at least slightly) different answers or productions, which is sometimes a problem. This is due to regular system updates and to their probabilistic nature. As they are trained on available datasets, their answers and other productions tend to reproduce the views and perspectives represented in those datasets, which are overwhelmingly based on English-speaking (and Western) cultures. For example, unless prompted otherwise, they will typically use Western names or examples in their production. In addition, despite appearing intelligent, GenAI tools do not "understand" the input they process or the content they generate. As a result, their outputs typically require human supervision and scrutiny, often more than specialised, non-generative AI systems.

While beyond the scope of this report, GenAI also comes with a series of societal challenges. Many observers are concerned by its environmental footprint, though this is still difficult to measure and compare with other digital technology. The dissemination of AI-generated information and data may decrease the quality of future generated content (as they enter their training datasets) and amplify some current limitations of our knowledge. This will make critical thinking and the development of metacognitive and higher-order thinking skills even more important than before. The full impact of how GenAI might transform societies, labour markets and economies is still emerging.

What is the general uptake of GenAI?

Most people experience GenAI through chatbots based on large language models (LLMs), such as OpenAI's ChatGPT, Google's Gemini, Microsoft Copilot, Anthropic's Claude, Mistral's LeChat, and Deepseek's Deepseek-R1. As of April 2025, based on website traffic data, chatbots dominated public use of GenAI tools, accounting for 95% of monthly traffic to the top 60 GenAI platforms. ChatGPT alone represented about 78% of the monthly visits, down from 89% in April 2023 (Liu, Huang and Wang, 2025[3]). Image-generating tools accounted for 2.4% of GenAI websites' traffic, video and audio tools for 1.9%, and productivity and business tools for less than 0.5%. While these shares remain small, the use of these

systems has grown significantly since 2023, in line with the overall growth of GenAI use. Competition is also mounting across platforms, with newcomers such as Deepseek and Perplexity gaining market share since 2023. Liu, Huang and Wang (2025[3]) show that the use of GenAI tools has both expanded and intensified. For example, between 2024 and 2025, the number of unique users of ChatGPT grew by 42%, visits per user increased by 50% and the average session duration doubled from 7 to 15 minutes – resulting in the doubling of its traffic (113% growth).

Most of this growth has been driven by users in high-income countries. In 2025, they accounted for 60% of GenAI use (compared to 55% in 2024), against 39% for middle-income countries and less than 1% for low-income countries (see Liu and Wang (2024[4]) for 2024 data). This reflects strong uptake in OECD members as well as accession and key partner countries such as Brazil, China and India. However, it also points toward a widening digital divide based on an adoption and use gap. Part of this gap might be due to measurement issues, as users in low-connectivity regions may not be able to access platforms via the Internet and use versions running offline on their device. Figure 1.1 presents the share of Internet users that used ChatGPT in 2025 and 2024 and thus provides an estimate of the uptake of GenAI tools across populations, acknowledging that averages mask higher usage among younger generations.

Figure 1.1. Increase of ChatGPT users as a share of Internet users, 2024-2025

The share of Internet users accessing ChatGPT has increased in OECD, accession and key partner countries
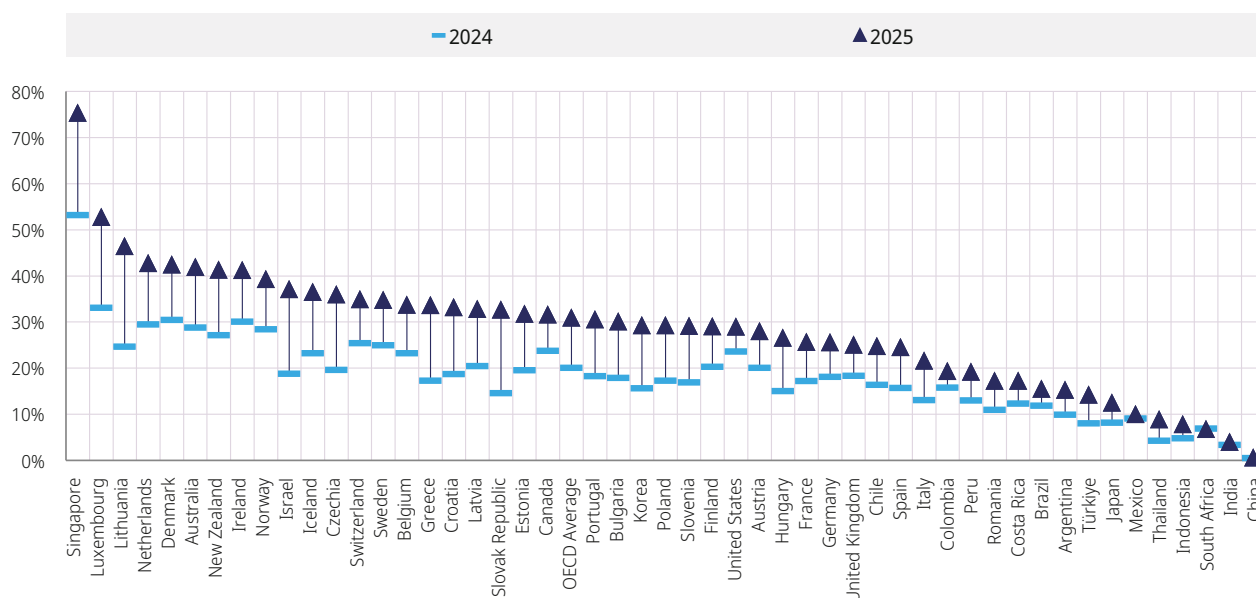
Note: ChatGPT is not generally accessible in China. While ChatGPT remained by far the largest GenAI chatbot service, local alternative chatbots tend to be more popular in their countries/regions of origin. The figure highlights the growth of the use of GenAI chatbots in almost all countries.

Source: Liu, Yan; Huang, Jingyun; Wang, He (2025). *Who on Earth Is Using Generative AI? Global Trends and Shifts in 2025* (English). Policy Research Working Paper; Digital; Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/099856110152535288

Is it common for students to use GenAI?

While there is currently no authoritative comparative data on the use of GenAI by students at different levels of education, several domestic and international surveys provide an initial picture of how widely students use these tools and for what educational purposes.

In Switzerland, a 2024 statistically representative survey of 8-18 year-old students points to a steep difference in use depending on age. Around 8% of primary students stated they used GenAI tools at least once a week, 30% in lower secondary, about half in general upper secondary education, and 40% in vocational education. Use in the home followed a similar age pattern (roughly 9%, 33%, 54% and respectively) (Oggenfuss and Wolter, 2024[5]). Including uses less frequent than at least once a week, about 70% of Swiss general upper secondary students use GenAI, and other Swiss pupils use it with a similar age/school pattern as intensive users.

In Estonia, a national survey of about 16 000 students found that 74% of lower secondary students and 90% of upper secondary students reported using AI tools to support their studies in 2024, with ChatGPT by far the dominant tool (70% of students use it) (Granström and Oppi, 2025[6]). Beyond national case studies, a cross-country European survey of more than 7 000 12-17 year-olds across seven countries (Germany, Greece, Portugal, Romania, Spain, Türkiye, and United Kingdom) saw high use of Generative AI by students. For example, 48% declared having used ChatGPT in 2024, with almost half of them instructed to do so by their teachers (Vodafone Foundation, 2025[7]).

The use of GenAI for higher education students seems to align with the age pattern mentioned above, although statistically representative surveys providing information on this are not yet available. Still, a few studies have surveyed a large number of higher education students (and reweighed their answers to make them more representative). In France, a 2023 study of about 4 500 students reported that 55% of higher education students used GenAI tools (Compilatio, 2023[8]). In 2025, the share had increased to 82% (Pascal et al., 2025[9]). In Germany, a survey of over 23 000 higher education students found 94% used AI in 2025, including 65% daily or weekly (Hüsch, Horstmann and Breiter, 2025[10]). A 2024 international survey of 3 000 higher education students in 16 countries also found that 86% used AI in their studies, including 54% daily or weekly (Rong and Chun, 2024[11]).

Evidence suggests that student use of generative AI has moved rapidly from marginal to mainstream since 2022. This is illustrated by looking at the trends among US upper-secondary students – the United States being one of the few places where several surveys were conducted over time. Surveys conducted in 2023 already indicated widespread exposure to GenAI, with around 25-33% of secondary students reporting having used GenAI for schoolwork (Center for Digital Thriving, Common Sense Media and Hopelab, 2024[12]). In 2024, comparable surveys suggest a marked acceleration, with close to 50% of middle and high school students reporting some use of AI tools, particularly for homework support, idea generation and explanations of difficult concepts (Impact Research, 2024[13]). In 2025, about 68% of teenagers aged 15-17 reported using AI chatbots such as ChatGPT (Pew Research Center, 2025[14]).

The above-mentioned increase in GenAI engagement between 2024 and 2025 (Liu, Huang and Wang, 2025[3]) is also likely driven by younger age groups. In 2024, compared to general Internet users, younger age groups and more educated people drove a substantial share of traffic to these tools, signalling early and concentrated use among teenagers and young adults (Liu and Wang, 2024[4]). There is no reason to believe that their contribution to this share decreased. It is possible or even likely that the early experimenters of 2024 might have transitioned to routine users in 2025.[2]

In short, students do use GenAI – a small extent in primary education, a moderate share in lower secondary education, but a majority seem to use it regularly in upper secondary and higher education. While student uptake of GenAI varies by country, the overall trends suggest student use is broadly growing across OECD countries.

What do students use GenAI for?

Many students are clearly turning to GenAI tools for academic purposes. However, their primary motivations often center on convenience and efficiency rather than deeper learning. When asked why they use GenAI, according to a number of studies, students typically responded they wanted "cognitive support", such as information, explanations and summaries, or "production support", such as idea generation, drafting, and, perhaps more problematically, solution generation.

In Estonia, for instance, grade 6-12 students most often reported using GenAI to achieve better scores, make educational tasks easier, and save time. These uses typically do not support student learning. Common uses include answering homework questions and generating ideas. Lower secondary students more often reported fact-checking, while upper secondary students tended to report summarising specific topics and creating visuals for presentations (Granström and Oppi, 2025[6]). In most of these cases, the primary motivation was efficiency and convenience (rather than deeper learning).

Similarly, in the seven-country European survey mentioned earlier, the most common out-of-school, non-instructed learning uses are obtaining information (56%) and getting explanations of terms and concepts (45%). Nearly one-third (31%) report using AI to provide complete solutions to tasks, while fewer (20%) use it for self-regulatory functions such as structuring personalised learning plans or tracking progress (Figure 1.2). These patterns align with findings from in-depth qualitative interviews with Dutch pupils (Topali, Ortega-Arranz and Molenaar, 2026[15]).

Figure 1.2. How do European students use AI to study? (2024)

What do you currently use AI applications for when learning outside of school and not being instructed by your teachers?
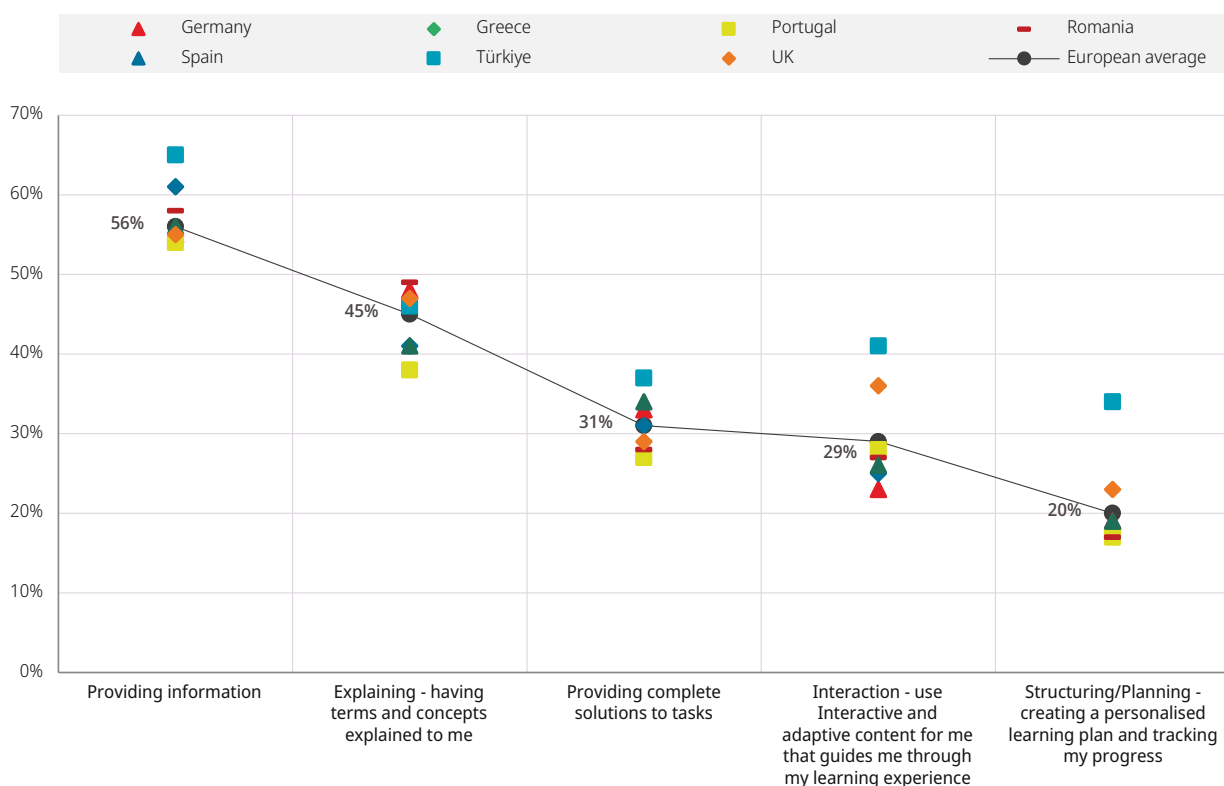
Note: Base: All participants; n = 7 000; shown without don't know / prefer not to answer. Multiple answers possible

Source: Vodafone Foundation (2025), *AI in European Schools: A European Report Comparing Seven Countries*, https://skillsuploadjr.eu/docs/contents/AI_in_European_schools.pdf

In higher education, students seem to mainly use GenAI tools to search for information, as well as for linguistic tasks such as editing, summarising, paraphrasing, and to a lesser extent drafting (Rong and Chun, 2024[11]). Hüsch, Horstmann and Breiter (2025[10]) provide the most detailed categories of use and show a largely similar picture, with students primarily using it for general search, idea generation and literature research on the "cognitive" side and for summarising and drafting on the "production" side (with about 22% of regular users) (see Figure 1.3). Interestingly, about 33% of students use GenAI as a "learning partner".

Taken together, available evidence suggests that a growing number of students use GenAI for general searches, comprehension and drafting, including as a shortcut to complete tasks and homework. The uses do not seem to be very different in higher education and upper secondary education and tend to reflect the study expectations for students at these different levels.

How do teachers use GenAI?

The OECD Teaching and Learning International Survey (TALIS) 2024 provides representative, comparative information on how lower secondary education teachers use AI (OECD, 2025[16]). On average, across OECD countries, 36% of lower secondary teachers' report having used AI in their work in the 12 months prior to the 2024 survey, with very large cross-country variations. Around 75% of teachers in Singapore and the United Arab Emirates report using AI compared to fewer than 20% of teachers in France and Japan.

Figure 1.3. Germany: Purpose and frequency of higher education student use of AI for their studies (2025)

Results from the CHE University Ranking's Student Survey 2025
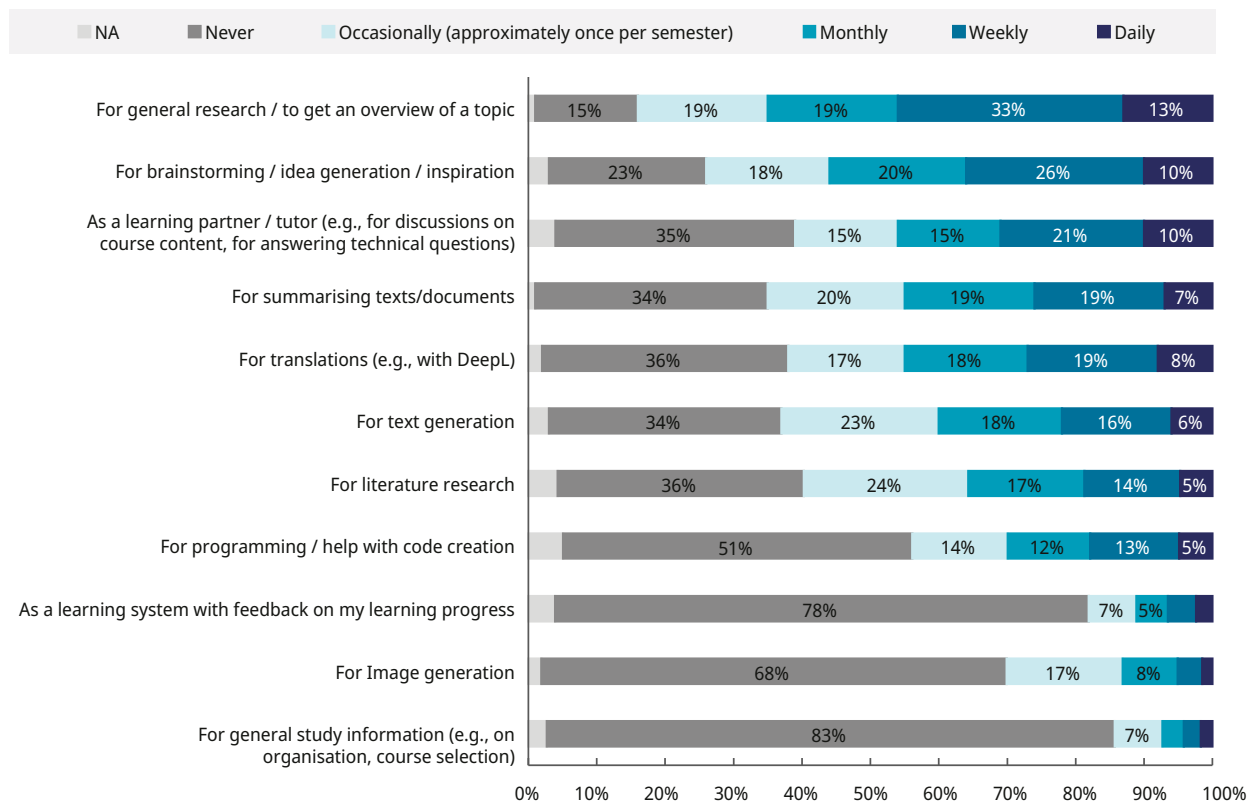
Note: N=23 288. The survey includes students enrolled in engineering subjects, as well as in psychology, educational science, German studies, and Romance studies. The students were distributed across 171 different higher education institutions (universities, universities of applied sciences, and cooperative education institutions), including six Austrian universities Students in undergraduate programs from the third semester up to and including two semesters beyond the standard period of study were surveyed. See more details on the CHE website. Ranked by descending order of weekly + daily use.

Source: *Studierendenbefragung für das CHE Hochschulranking*. Hüsch, Horstmann and Breiter (2025[10]).

While the survey does not ask whether they used AI (all kinds) or GenAI, the tasks that teachers report suggest that most uses involve GenAI tools. Teachers primarily use AI for preparation and productivity tasks: on average, 68% report they use it to efficiently learn about and summarise topics they teach, and 64% use it to generate lesson plans. Among AI users, on average 25% report using it to review data on student participation or performance and 26% use it to assess or grade student work (Figure 1.4).

In addition, 40% of teachers "agree" or "strongly agree" that AI helps them support students individually, on average. Around 50% agree that AI assists in creating or improving lesson plans, though agreement ranges from as low as 18% in France to as high as 91% in Viet Nam. Seven in ten teachers, on average, believe AI could enable students to misrepresent others' work as their own. Around four in ten teachers agree that AI may amplify biases, reinforce student misconceptions, or compromise data privacy and security.

As for the teachers who have not used it, they report feeling overwhelmed by the growing expectation to integrate digital tools in education, which they see as a barrier to using AI in their teaching. This varies markedly across systems, from fewer than 20% in Brazil, Chile, Costa Rica, Italy, Morocco, Türkiye and the United Arab Emirates, to over 50% in Croatia, the Flemish Community of Belgium, Japan and Serbia. On average, three in four teachers report that they lack the knowledge or skills to teach using AI. About half of these teachers also believe that AI should not be used in teaching. In terms of school policy, one in ten teachers reported that their school prohibits the use of AI in teaching.

Figure 1.4. Teachers' use of and opinions about AI in teaching (2024)

Percentage of lower secondary teachers who agree with the following statements
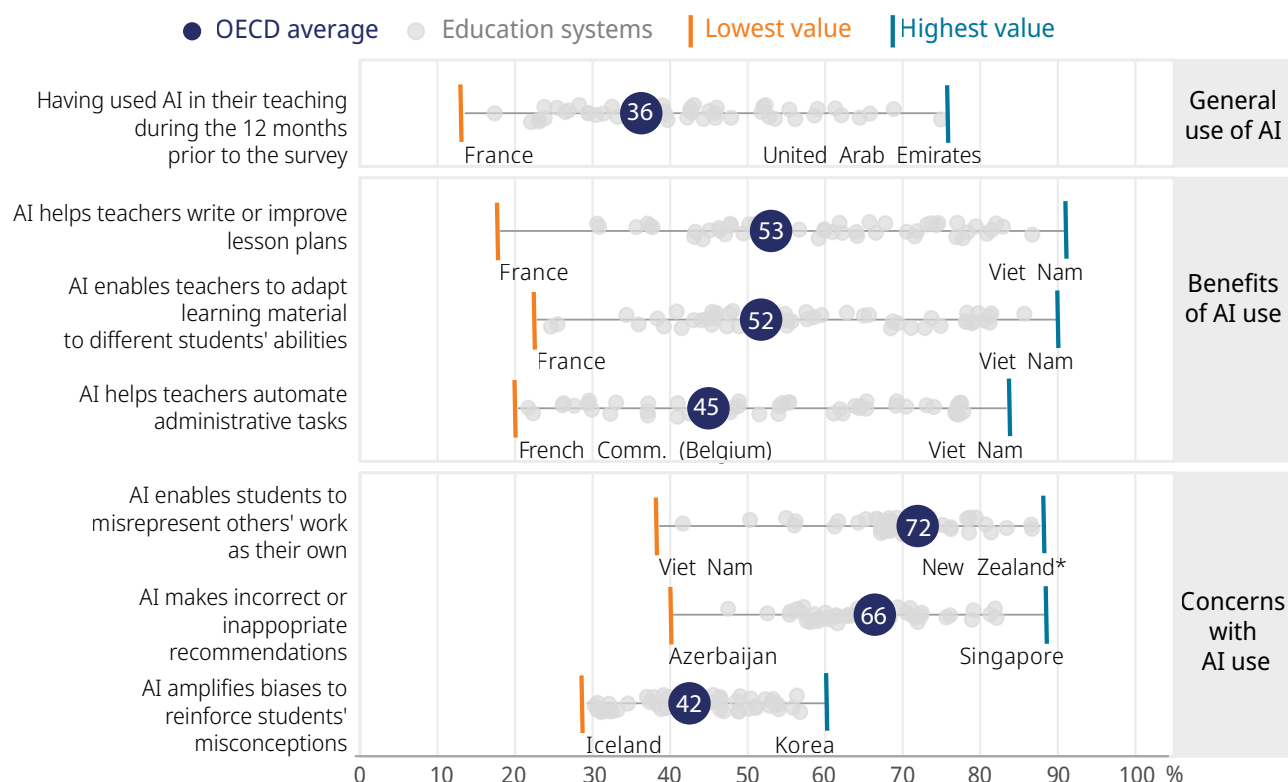
Note: *Estimates should be interpreted with caution due to higher risk of non-response bias.

Source: OECD, TALIS 2024 Database, Tables 1.59, 1.61 and 1.62.

What do we know about the use of GenAI tools by teachers in other levels of education? TALIS 2024 allows for comparisons of teachers' use of AI at other levels in a limited number of participating countries. In these education systems, primary teachers are generally less likely to report using AI than their lower secondary counterparts, with particularly large gaps observed in Australia and the Flemish Community of Belgium. Teachers who use AI in primary education are often more likely to apply it to targeted pedagogical practices, notably to support students with special education needs and to adapt the difficulty of lesson materials to students' learning needs, with especially large differences observed in France and the Netherlands. This may suggest that they use assistive AI technologies (for students with special needs) and adaptive learning systems, which may or may not involve GenAI (OECD, 2021[1]).

Domestic surveys and country studies largely confirm the picture drawn by TALIS in terms of GenAI uses. For example, in Estonia, a survey of about 4 000 teachers found that 53% reported using AI tools in their work, with higher reported use among primary and lower-secondary school teachers (66%) than among upper secondary teachers (50%) (Granström and Oppi, 2025[17]). Teachers who use generative AI mainly do so to increase efficiency in their work, such as preparing materials, supporting planning, or streamlining routine tasks, rather than for deeply transformative or individualised pedagogical practices. The strongest predictors of use are teachers' self-reported readiness, including confidence, access to tools, and institutional support, and their belief that GenAI provides clear practical benefits for teaching. By contrast, non-use is largely explained by insufficient skills, lack of confidence, limited training opportunities, or uncertainty about pedagogical value. Age shows a small positive association with use, while years of teaching experience do not matter once readiness and perceived usefulness are taken into account, suggesting that GenAI adoption in Estonia is driven less by seniority than by capacity-building and perceived relevance to day-to-day teaching.

Studies in Australia (AHISA, 2023[18]; Collie and Martin, 2025[19]), Italy (INDIRE, 2025[20]), Slovenia (Licardo et al., 2025[21]) and the United States (Diliberti et al., 2024[22]) provide variations on the same types of usages mainly for preparation work (generative worksheets, lesson plans, activities). They also provide differences in the use across subjects: for example, in the United States, English and "social studies" teachers were more likely to use these tools, possibly because they are more likely to either design or adapt their lessons compared to some other subjects (Diliberti et al., 2024[22]).

The uptake of GenAI is likely much greater in higher education, given the widespread use of GenAI tools for research (Guellec and Vincent-Lancrin, 2026[23]). In France, a survey of 30 000 higher education students, teachers and academic staff found that 80% of higher education teachers had already used GenAI tools in 2025, mainly to help them draft and prepare their course (49%) or draft student evaluations (26%), and more rarely to support them to correct (13%) or mark (8%) student work (Pascal et al., 2025[9]). An international study of 1 700 teachers in 52 higher education institutions reported that 68% used AI in general. Among those, 75% used GenAI to create teaching materials and 24% to generate feedback to student work (Rong and Chun, 2025[24]). Interestingly, both surveys suggest that higher education teachers rarely use GenAI as part of their actual teaching or request their students to use GenAI. Unfortunately, data on the frequency of use remains limited.

In sum, teachers' use of GenAI varies by country and the education level they teach. While the use of GenAI seems more prominent in lower secondary than in primary when looking at TALIS 2024 data, research from Estonia shows that primary teachers there use these tools more than secondary teachers. Teachers seem to use these tools to the same or to a lesser extent than their students, but not more. In Estonia again, while 50% of upper secondary teachers used GenAI tools in 2025, 90% of their students do – and in the United States, the only study surveying students and teachers at the same time found similar levels of regular use (with slightly higher levels of use overall for undergraduate students) (Impact Research, 2024). In higher education, the share of regular users seems similar among teachers and students. Whatever the education level, teachers report mainly using GenAI tools for the same reasons: assisting them with the generation of lesson plans, pedagogical activities and teaching/learning materials, and sometimes with the provision of feedback to students.

When does GenAI improve learning outcomes?

Given the widespread use of GenAI tools by students, including outside school settings and without teacher guidance, and to a lesser extent by teachers themselves, important questions for the education sector include: how does the use of GenAI affect learning? And how can GenAI tools be used to enhance learning? The first part of the *OECD Digital Education Outlook 2026* addresses these questions, which are further explored throughout the report.

Does the use of general-purpose tools automatically enhance students learning?

Gasevic and Yan (2026[25]) provide an overview of the emerging research on teaching and learning with GenAI and highlight that, depending on their use, GenAI tools can undermine or enhance learning. In some cases, the use of GenAI can be deceiving. For example, GenAI systems may enhance the apparent quality of student work (that is, their performance at educational tasks) without improving their actual learning (their knowledge and skill acquisition) – a paradox illustrated in several studies.

A study of 1 000 high school students in Türkiye in mathematics in grades 9, 10 and 11 examined three practice conditions across six 90-minute sessions: 1) with their course notes and textbook (as usual); 2) with a general-purpose LLM chatbot ("GPT base"); 3) with an educational LLM chatbot ("GPT tutor"). Educational LLMs are configured (or finetuned) to avoid the provision of direct answers and support learning (but there is no indication in this case that it was an adaptive learning tool). This randomised controlled trial analysed the results of the students during practice and noted that the percentage of correct answers to the exercises was much higher for the students using GenAI tools compared to those working by themselves, with a much higher performance for those using the educational chatbot. However, when their knowledge was assessed in a closed-book environment, the performance gains vanished: students who used the general-purpose GPT scored lower than those studying on their own (Bastani et al., 2024[26]) (Figure 1.5). The students who used the educational chatbot performed about as well as their self-study peers. While they enhanced their GenAI skills, one would expect effective digital learning tools to enhance learning, not merely practice performance, which hints to the possible under-performance of self-declared "educational" GenAI tools.

Figure 1.5. Successfully performing a task with GenAI does not automatically lead to learning

Türkiye: Comparison of high school students' maths exam and practice results when practicing alone or with GenAI tools
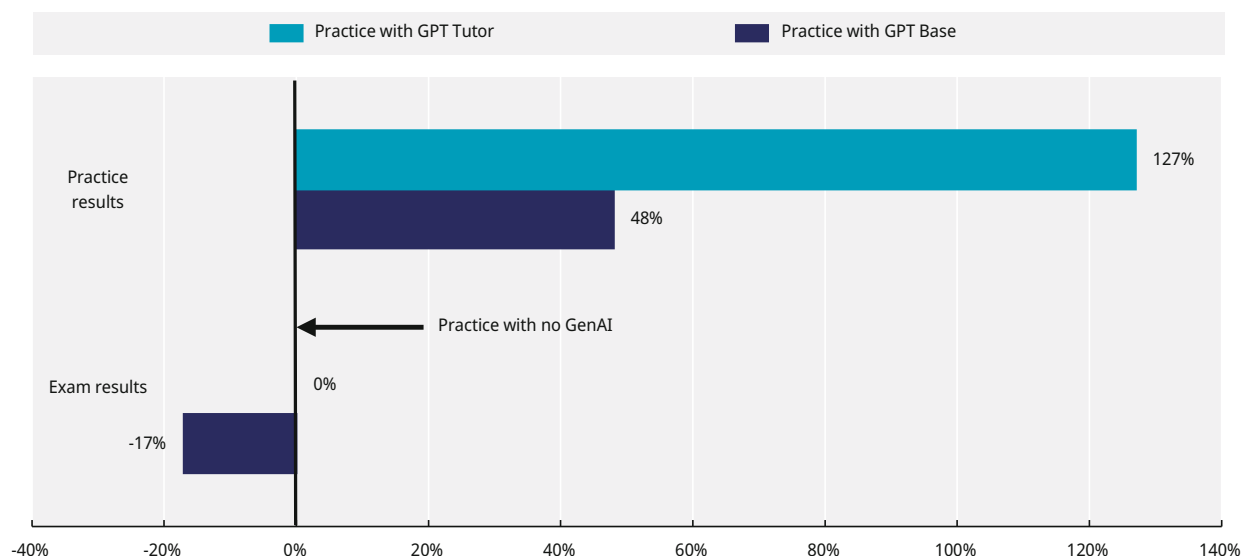
Note: In this randomised controlled trial, access to GPT-4 improved student performance when practising mathematics exercises - especially the tutoring version. However, during closed-book exams, students who had used the base (or general-purpose) version of GPT-4 performed 17% worse than those who never had access

Source: Bastani, H. O. Bastani, A. Sungu, H. Ge, O. Kabakcı and R. Mariman (2024), "Generative AI Can Harm Learning", The Wharton School Research Paper, Available at SSRN: http://dx.doi.org/10.2139/ssrn.4895486.

Other studies show similar results in different circumstances, although with less statistical power. Their purpose lies more in explaining why enhanced performance when using general-purpose tools does not necessarily lead to learning. Two studies analysed the metacognitive processes of Chinese higher education students using general-purpose LLMs to revise an essay they had written in English without using a GenAI tool. In a first randomised control trial (Fan et al., 2025[27]), students were assigned to revise their essay in 4 different ways: 1) alone; 2) with human expert advice; 3) using a check list (and digital tools); 4) with a general-purpose LLM. They found no statistically significant difference in motivation among the conditions, although the three groups with external support had slightly increased motivation. In terms of task performance, the group using the GenAI tool scored highest, but knowledge gains measured by a knowledge test did not improve. In terms of metacognitive processes or self-regulated learning, the group using GenAI performed less metacognitive tasks, especially evaluation and orientation. Another study compared differences between students who revised essays by asking for human advice or used a general-purpose LLM (Chen et al., 2025[28]). They found that those who interacted with human experts sought help in a linear way, following the models of "help seeking" theory: diagnosing what they need help with, asking for help, evaluating the help received, iterating, and then implementing the final help. They showed that when interacting with a general-purpose chatbot, some students tended to ask directly for help and implement immediately the solution received. They would often skip the diagnosis, evaluation and iteration stages. The authors refer to this as "metacognitive laziness", a way of describing "cognitive offloading".

Another study is worth mentioning as it casts neuroscientific light on the above findings (Kosmyna et al., 2025[29]). In the United States, students from 5 universities were asked to write a 20-minute essay under three working conditions: writing by themselves ("brain-only"), with a search engine, or with a general-purpose GenAI tool (ChatGPT). Afterwards, within one hour of the actual essay writing, only 12% of the LLM group could quote something from their essay (exact recall) as opposed to 89% in the two other groups. Even though the essays of the LLM group were well rated, the LLM group also had a lower ability to summarise their essay's viewpoint, lower level of ownership, and more similar content across essays. Brain imagery suggested a shift in their executive tasks from generating content to supervising the AI-

generative content, with lower neural connectivity and involvement. They also showed that writing alone first and then using the LLM preserved higher levels of activation and recall. In contrast, those who started with an LLM and then continued alone had low levels of activation and recall. These findings suggest that initial cognitive activation is crucial before using GenAI.

This evidence highlights that a proportion of students using general-purpose LLMs may take shortcuts, avoiding the productive struggle and cognitive effort needed for learning and durable knowledge and skill acquisition. This aligns with a "fast" rather than more productive "slow" and iterative use of GenAI, as argued by Beghetto (2026[30]), and the notion of "lazy" use of GenAI (Vinchon et al., 2023[31]).

Is it possible to enhance learning outcomes using pedagogical GenAI tools?

While evidence suggests GenAI tools sometimes enhance student performance at the cost of lasting skill and knowledge development, that does not mean positive outcomes are impossible. In fact, several chapters of this *Outlook* highlight some promising results too (Gašević and Yan, 2026[25]; Li and Hu, 2026[32]; Strauß and Rummel, 2026[33]; Cukurova, 2026[34]; Topali, Ortega-Arranz and Molenaar, 2026[15]; Baker et al., 2026[35]). Two types of uses of educational GenAI should be considered. Some LLM-based GenAI tools are repurposed for education. They have either been "finetuned", that is to say, partly retrained based on education-relevant data, or "configured" through a series of instructions and prompts on how they should answer user requests.

For example, in the United States, a Stanford-developed GenAI tool called "Tutor CoPilot" finetuned GPT4 based on the ethnographic observations of good teachers providing feedback (Wang et al., 2024[36]). The tool was integrated in an online tutoring platform to assist in real time 900 tutors (rather than teachers) working with 1 800 underserved pupils in the United States. The intervention raised student pass rates by 4% on average, with the largest gains among less experienced tutors (9%) and those previously rated as lower quality (7%). The use of the tool made less difference for the more effective and experienced tutors. Given the effectiveness of tutoring as a learning strategy, this robust study shows promise for GenAI tools to support tutors, and perhaps also less experienced teachers. Other promising uses are also being developed to make teaching more effective (Demszky, 2026[37]).

Figure 1.6. Educational GenAI tutoring can outperform in-class learning

Results from a randomised controlled trial taking place over two weeks in a higher education introductory physics course, United States, 2023
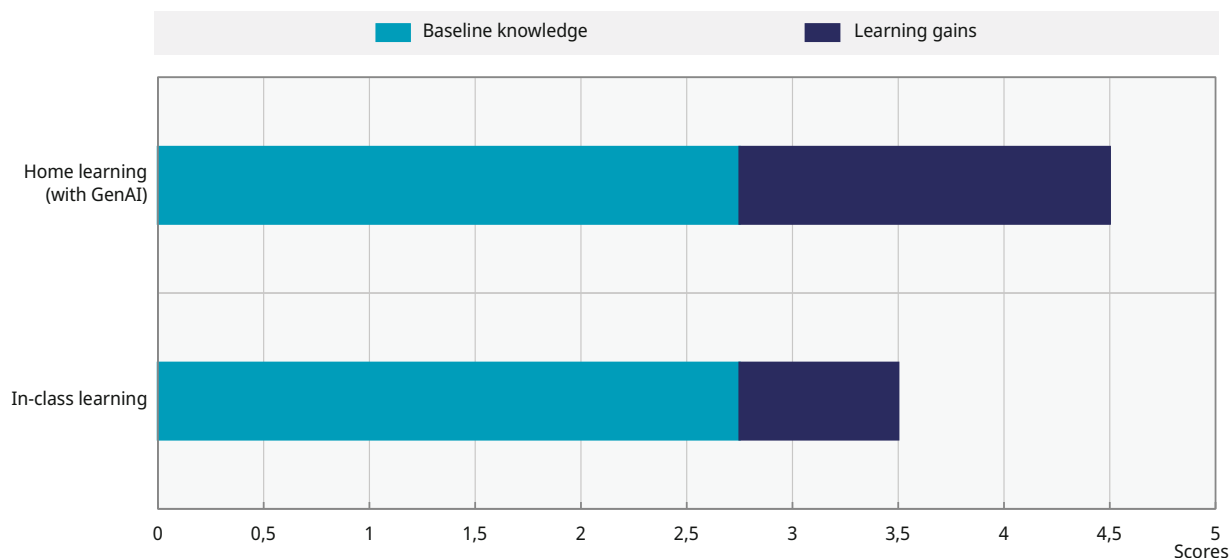
Note: N=316. These are the results of a randomised controlled trial comparing mean post-test performance between students taught with an active learning pedagogy, either in class or online with a GenAI tutor. All differences are statistically significant with a large effect size of 0.63 between the two groups. Students in the GenAI-tutoring group also reported significantly higher motivation and engagement than their peers in the in-class group.

Source: Kestin, G. et al. (2025[40]), "AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting", *Scientific Reports*, Vol. 15/1, https://doi.org/10.1038/s41598-025-97652-6.

Another randomised controlled trial in the United States compared learning in an undergraduate physics course at Harvard university delivered through in-person "active learning" versus an AI tutor based on GenAI that was configured to implement "active learning" pedagogical principles online. The idea was to have the same pedagogical practices in both conditions, only the delivery differed. The study found that students learnt significantly more in less time when using the AI tutor, and also felt more engaged and motivated (Kestin et al., 2025[38]) (Figure 1.6). In China, a GenAI tool configured to implement problem-based learning in reading led to enhanced reading performance and motivation compared to the conventional, less personalised approach (Huang et al., 2025[39]).

Strauß and Rummel (2026[33]) reviewed different studies that use LLMs to support collaborative learning. They noted that most studies do not use a general-purpose LLM but have instead usually configured the model to control the behaviour of the LLM and assign it some specific role and behaviour, following previously identified good practices in research on computer-assisted collaborative learning.

While evidence is limited and often lacks statistical power, these studies show that GenAI can enhance learning when tools have been configured to be educational and grounded in expert teaching practice and learning science.

Is it possible to enhance learning outcomes using general-purpose GenAI tools with pedagogical purpose?

A key question is whether we should move exclusively to educational GenAI tools that have been designed from the outset to support learning. While these tools are in principle more likely to support learning than general-purpose models, a pedagogical use of general-purpose GenAI tools can also develop students' knowledge and skills. In addition, the use of general-purpose GenAI tools will foster students' AI literacy in a broader way, allowing them to engage with some of the tools they will have to use in the labour market.

For example, in the United Kingdom, a study on creativity provides some lessons that could be applied to education (Doshi and Hauser, 2024[40]). The study measured how the use of a general-purpose LLM can enhance people's creative output as well as the quality of their writing (communication). Participants were asked to write an original eight-sentence creative story and assigned to three groups: 1) they had to write independently, with no GenAI support; 2) they could ask for one GenAI idea (the GenAI was configured to provide a 3-sentence idea); 3) they could ask for 5 GenAI ideas (same principle repeated). The groups that could brainstorm with GenAI outperformed those working alone both in terms of creativity and in terms of quality of the writing, with those receiving more ideas performing best (Figure 1.7). Importantly, compared to the studies presented before, in this case they all wrote the story without using GenAI so that the benefits did not stem from cognitive (and production) offloading. However, as in the "brain imagery" study mentioned earlier, the stories written from GenAI-assisted groups were more similar, pointing to a possible negative side effect and a drop in collective creativity.

A randomised controlled trial in nine secondary schools in Nigeria provided access to Microsoft Copilot based on GPT-4 in an after-school programme while the control group did not have access to it (De Simone et al., 2025[41]). The students in the treatment group worked in pairs and received teacher instructions on how to use Copilot, including prompts. While this was a general-purpose LLM, the students were pedagogically guided on how to use it effectively and put intentionally in a particular pedagogical scenario (peer learning). The study found a positive, medium-sized impact on learning.

In Indonesia, Darmawansah et al. (2025[42]) compared the effects of using ChatGPT rather than search engines as a support tool to prepare and implement argumentative speeches in English (as a foreign language). Far from being left alone, the ChatGPT group were provided with some initial training on prompting, and looked for initial information with ChatGPT. In a second phase, both groups were provided with a collaboration script guiding their collaborative learning. In a third phase, groups produced an argument mapping, with the help of ChatGPT in that group, before the performance of their argumentative speeches. The GenAI group outperformed the "search engine" group in learning gains on argumentation, and in self-reported levels of "critical thinking awareness" and "collaborative tendency". The GenAI group provided more "backing" (factual information) to its arguments and benefited from the "rebuttals" provided by the GenAI, with results dependent on the quality of student prompting and of their English proficiency. A study in Vietnam making students use ChatGPT and CharacterAI in a collaborative learning setting in English as a foreign language exhibited similar findings (Bui, Ngo and Mai, 2025[43]).

Other studies show that the use of general-purpose LLMs in a pedagogical manner can improve students' learning outcomes. For example, studies indicate better knowledge acquisition when using an LLM as a teachable agent in computer science (as opposed to studying alone) (Chen et al., 2024[44]).

Figure 1.7. Using GenAI can enhance human creativity and writing quality

Comparison of creative writing output between those with no support from GenAI and those receiving 1 or 5 GenAI ideas
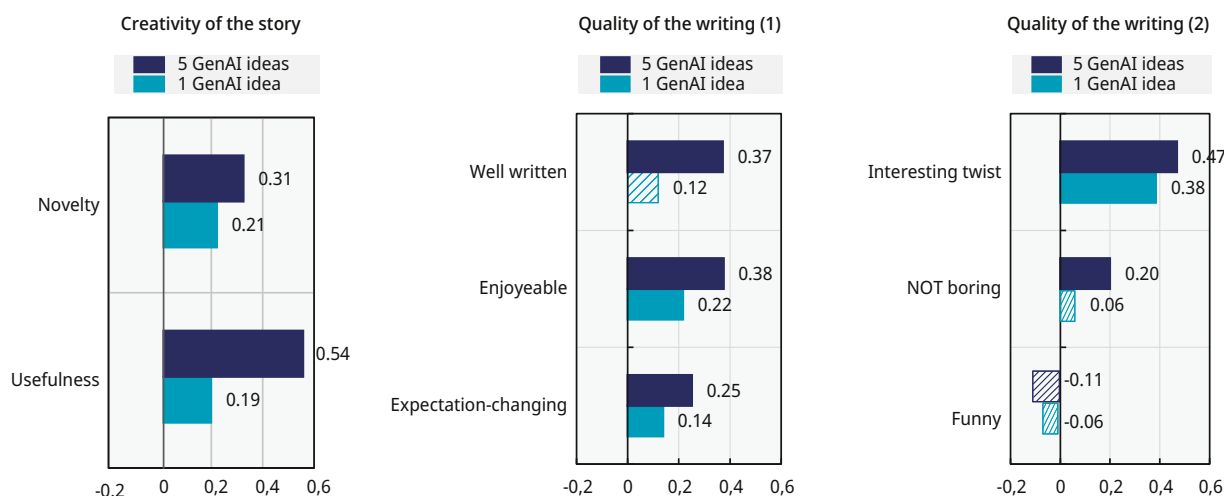
Note: Stories are written by humans after receiving no idea from GenAI, 1 idea or up to 5 ideas. By convention, an effect size below 0.2 indicates a small impact, and between 0.2 and 0.6 a medium impact. On the middle and right-hand panels, human assessors were asked whether they found the stories well written, whether they enjoyed reading it, whether it changed their future expectations about good stories, whether it included an interesting twist, whether they found it boring (reversed here to keep all results in the same direction) or funny. Results that are not statistically significant are marked with patterned colours.

Source: Doshi, A. and O. Hauser (2024), "Generative AI enhances individual creativity but reduces the collective diversity of novel content", *Science Advances*, Vol. 10/28, https://doi.org/10.1126/sciadv.adn5290.

Can GenAI improve learning outcomes by providing feedback to students?

GenAI holds strong promise for enabling the rapid and scalable generation of feedback to students. Good formative assessment relies on frequent, timely, targeted, and individualised feedback on student work. Given classroom sizes, teachers cannot always provide detailed, personalised feedback to all students, which makes AI-generated feedback a plausible driver of learning outcomes. Gašević and Yan (2026[25]) review the research literature on GenAI-generated feedback and argue that GenAI can support teachers to give better feedback, although it cannot replace human feedback.

The research literature comparing feedback generated by LLMs, usually configured based on marking rubrics and examples of good answers, finds that AI-generated feedback matches human feedback in quality, while acknowledging the shortcomings of human feedback. Heinrich et al. (2025[45]) don't find much difference on the grading of short answers in political science. Chevalier, Orzech and Stankov (2024[46]) found that students who received feedback based on GPT-4 had similar learning gains as those who always received feedback from human instructors, as is also shown by a meta-analysis of AI feedback covering all forms of AI (Kaliisa et al., 2025[47]). Dai et al. (2024[48]) compared the quality of human and GenAI-generated feedback on task (correctness), process (learning strategies), self-regulation (monitoring learning), and self (personal traits and motivation). On average, GenAI produces more readable and stylistically polished feedback than human educators on written essays. Human feedback tends to be more succinct. GenAI also slightly outperformed humans in the frequency of provision of feedback on process and self-regulation, which support deeper learning and learner autonomy. Despite similar performance on formative feedback, humans and GenAI can have low levels of agreement on the strength and weaknesses of student work (Dai et al., 2024[48]), and thus on the marking of student work (Chevalier, Orzech and Stankov, 2024[46]).

Does this equivalent performance imply that formative feedback on students' tasks should be delegated to GenAI? Not necessarily. Feedback is only effective if taken seriously and acted upon, which partly depends on its quality but also on its credibility and "motivational" dimension. As noted by Gašević and Yan (2026[25]), comparable performance does not imply pedagogical interchangeability. The growing research literature on the perception of AI feedback by students finds that students perceive human feedback as more credible and meaningful, which makes it more likely to influence their motivation, evaluative judgement, and trust – relational dimensions that remain difficult for GenAI systems to reproduce.

As a result, the consensus among experts points towards a hybrid approach to feedback. The differences in performance between GenAI and human instructors create new opportunities to complement and enhance the effectiveness of teachers on feedback provision. For example, GenAI could generate initial feedback on student work, which teachers can use as suggestions to enhance their own feedback drafts. GenAI could also be used to assess the quality of the feedback provided based on research on high quality feedback. To the extent that their work is done on digital platforms, GenAI could also provide feedback on the process of learning, an important form of feedback that is usually out of reach for teachers. Ultimately, getting this feedback vetted by humans is essential for it to be trusted and effective for students.

What are the initial take-aways on the effects of GenAI on teaching and learning?

In conclusion, using general purpose GenAI tools carries risks, particularly if students use them to avoid the cognitive effort that educational tasks are designed to elicit. However, learning science also recognises that learners may need "scaffolding", support that is gradually removed as students become more proficient. When used pedagogically, GenAI tools can serve as scaffolding (Strauß and Rummel, 2026[33]). This highlights the importance of teachers to develop pedagogical competences that include GenAI in their teaching and assignments.

Studies that found a positive effect in the use of general-purpose GenAI with pedagogical purpose included some development of students' GenAI literacy. Countries are starting to develop strategies to this effect (Box 1.1). The OECD and the European Commission have developed an AI literacy framework to support these efforts (European Commission and OECD, 2025[49]).

Even when teachers do not integrate GenAI as part of their teaching, they may still have to adapt their teaching practices as their students can use these tools independently. How should teachers adapt their teaching and assignments so that they continue to yield positive learning results, even for students who use GenAI as a productivity, rather than a learning, tool? While many of these attempts at pedagogical change are undocumented, a few studies offer some insights. For example, Kosar et al. (2024[50]) document how they redesigned their computer science course by: 1) changing their home assignments so they could not be directly completed by a GenAI chatbot; 2) using lab time for oral defences of the produced code, asking for "understanding" questions; 3) turning their mid-term exams into a paper and pencil exam focused on conceptual understanding. In these conditions, the groups with and without access to GenAI had similar learning outcomes. Documenting these efforts of pedagogical redesign internationally would support a more rapid sharing of knowledge around the use of GenAI.

Box 1.1. How do countries support GenAI literacy?

Most countries address students' GenAI literacy as part of AI literacy, which is often included in digital competence or transversal skills frameworks. Rather than introducing GenAI as a standalone subject, education systems integrate it across subjects and educational levels or, in some cases within a specific subject. Across countries, AI literacy increasingly encompasses understanding the limitations and risks of generative models, not only their use or the development of prompting.

Belgium (Flemish Community) integrates AI-related learning objectives into attainment targets, including prompt formulation and critical evaluation of digital sources. Czechia embeds generative AI within mandatory digital competence across subjects, supported by extensive methodological materials. In France, AI literacy related to generative tools is supported through the Pix platform, which includes modules on prompting, system functioning and data issues. In the United States, Arizona developed a roadmap aligning generative AI literacy goals with state academic standards, explicitly integrating generative AI into subject-level curricula rather than treating it as an add-on.

Some countries have also included GenAI literacy in the "computer science" part of their curriculum. Lithuania includes AI, including generative applications, within informatics for upper secondary education. In Ireland, GenAI-related topics will appear in the Junior Cycle Digital Media and Literacy Short Course (from 2026) and in optional senior-cycle computer science curricula.

In the 2025 European Schoolnet's survey, 19 out of 20 European countries cite teacher's GenAI literacy as one of their priorities (European Schoolnet, 2025[51]). GenAI literacy should be understood as the ability to effectively use GenAI to prepare and deliver their teaching as well as having a high-level understanding of GenAI technology. Nearly all systems addressing GenAI provide professional development through online courses, workshops, webinars and communities of practice, although TALIS 2024 results show big variations in the uptake among countries.

Finally, while GenAI hold promise for formative assessment and the quality of the feedback given to students, it may mainly be used to support human instructors rather than to fully automate feedback. The human relationship remains a core element of teaching and learning, and AI-generated feedback does not come with the same credibility and motivational drive as human feedback, even when its quality is equivalent.

While most research has so far focused on general-purpose GenAI tools, future studies should examine the effects of GenAI tools designed for education. These tools seem more promising. Nevertheless, their efficacy in terms of improving learning outcomes or pedagogical competence should be evaluated as a minimal requirement for their adoption. As shown in the reviewed evidence, educational GenAI tools can be both student-facing, teacher-facing, or both. In practice, many tools combine these roles.

What do educational GenAI tools look like?

While general-purpose GenAI tools can support learning when used with clear educational purpose, current evidence suggests that the development of education-specific GenAI tools may hold even greater promise for improving teacher practice and student learning outcomes. This raises an important question: what does an educational GenAI tool look like?

At a minimum, any educational GenAI tool should generate safe and age-appropriate content, respect users' privacy and data protection, be explainable and transparent, and mitigates algorithmic bias to the extent possible (OECD, 2023[52]; OECD, 2023[53]). Beyond these safeguards, it should also be "educational" and help teachers to teach more effectively and enable students to learn more or to catch up with peers. This report provides several examples of prototypes or early implementation of such educational GenAI tools.

How can GenAI tutors support personalised learning?

Adaptive learning systems are one of the most used digital learning tools in education systems. They illustrate the "personalisation" agenda associated with AI by providing learners with the possibility to practice and expand their knowledge. These systems usually assess students' initial knowledge and skills as well as their misconceptions, diagnose types of problems that students should be performing, and adjust the difficulty of problems depending on how students perform (OECD, 2021[1]). Impact evaluations show that these systems are overall effective for learning. Intelligent tutoring systems usually provide feedback and support students to learn, rather than simply telling them whether they were right or wrong. These rule-based AI tutors struggled with unanticipated student inputs or questions though, which limited the scalability and the richness of tutoring interactions. GenAI models, particularly LLMs, overcome these constraints, enabling more engaging and versatile tutoring experiences.

Recent research that compares legacy intelligent tutors to next-generation LLM-driven systems (e.g. Hu, Xu and Graesser, 2025[54]) help articulate a vision for GenAI pedagogical agents. Li and Hu (2026[32]) show how LLMs' capacity to dynamically generate fluent, contextually appropriate dialogue brings new opportunities to intelligent tutoring systems in terms of adaptability to learner profiles and ability to be applied to different subject matters. They allow for a more flexible tutoring experience, capable of addressing unforeseen questions or novel problem scenarios in real time. GenAI tutors can produce human-like explanations, ask clarification questions, and scaffold student thinking through multi-turn dialogue. Using techniques such as retrieval-augmented generation (RAG) or fine-tuning, they can incorporate up-to-date factual information into their tutoring.

Still, the big challenge is ensuring GenAI tutors are configured or fine-tuned to have interactions that are pedagogically sound, as was the case for effective intelligent tutoring systems not based on GenAI. Through mechanisms like conversation history or explicit memory modules, GenAI tutors can iteratively refine learners' profiles and adapt tutoring sessions accordingly, for example by adjusting difficulty or revisiting past misconceptions. Keeping the challenge level appropriate for learners is one of the key insights of learning science. GenAI tutors also have the potential to play different pedagogical roles, and to shift between these different roles according to circumstances. They can play the role of: mentors, providing academic guidance; coaches, providing motivational support; peers, with less formal interactions that resemble peer learning. Researchers are also exploring how they could become companions and support learners across a wide range of learning and subjects.

Often drawing on Socratic questioning and related strategies, GenAI tutors can provide scaffolded dialogue: guiding learners to develop their knowledge through carefully sequenced questions rather than delivering answers outright. This approach is rooted in Vygotskian "scaffolding" and "zone of proximal development" theory, where support is provided just beyond the learner's current ability and gradually withdrawn as competence grows (Vygotsky, 1978[55]). GenAI tutors are particularly well-suited to implementing Socratic questioning as they can generate an extensive range of questions and follow-ups and flexibly rephrase or adjust the difficulty of questions based on learner responses. Li

and Hu (2026[32]) take the Socratic Playground as a case study to illustrate the possibilities of GenAI tutors, as well as presenting the underlying AI models that make it possible.

How can GenAI support collaborative learning?

Improved learning outcomes come from sound pedagogical principles. When it comes to collaborative learning, Strauß and Rummel (2026[33]) emphasise that general purpose GenAI systems are unlikely to support effective collaborative learning, just as random human beings, even if knowledgeable, would not automatically make collaboration successful. To be effective, GenAI tutors should play different roles and target group interactions differently. Research identified several roles GenAI played when integrated into collaborative learning environments: 1) it served as a repository of information for the group; 2) it collected information about the group, its collaboration, or its results; 3) it generated additional learning material for the group to use in their reflection, for example a contrasting case; 4) it encouraged active participation as a facilitator positioned as "outside of the group"; 5) it aimed at developing domain-specific knowledge as a dialogue partner, in the Socratic spirit mentioned above; 6) it brought a specific expertise to the group as an artificial group member. By playing these possible formal roles, GenAI chatbots can target different aspects of the collaboration: the cognitive part, by providing knowledge and expertise; the social part, by making sure that students contribute equally and that all voices are heard; the metacognitive part, to encourage students to reflect on the collaborative process or encourage them to make some steps in their reflections.

The small number of studies on GenAI collaborative learning show small- to medium-sized positive results, but importantly, GenAI does not orchestrate collaborative learning on its own – researchers and practitioners do. All the possible roles mentioned above must be assigned by researchers or developers by configuring or fine-tuning the GenAI tools used, building on the accumulated knowledge of computer-assisted collaborative learning. One cannot assume that, by itself, general-purpose LLMs can spontaneously take on different roles to make collaborative learning successful. However, the rise of GenAI is opening new avenues for computer-assisted collaborative learning that show promise if they remain aligned with existing pedagogical knowledge and research.

How can GenAI support teachers and preserve their agency?

As mentioned earlier, teachers often use general-purpose GenAI to support their work, for example the generation of lesson plans and learning materials. Some studies have shown that these activities can bring productivity gains. For example, a randomised controlled trial examined the use of GenAI among 259 teachers across 68 secondary schools in England and found that teachers who received practical guidance to use GenAI reduced their lesson and resource planning time by an average of 31 percent, reducing their weekly average planning time from 81.5 to 56.2 minutes, without compromising the quality of their lesson plans and resources (Roy et al., 2024[56]). However, such uses may come with risks similar to those observed with students, of "cognitive offloading" or "metacognitive laziness". Teachers and education systems will have to explore and define how, when and in which cases it is appropriate to use these tools. If teachers offload too many of their tasks to GenAI they may stop their professional development and harm their relationship with students. For example, evidence shows that students prefer to receive feedback from human teachers, sometimes even if they rated the feedback from GenAI tools as superior (Gašević and Yan, 2026[25]; Cukurova, 2026[34]). A related question is what design features ensure educational GenAI tools for teachers improve teaching quality, while maintaining teacher agency and autonomy?

Cukurova (2026[34]) proposes a conceptual framework for human-AI interactions and defines three different types: replacement (or full automation), when the AI tool accomplishes a task or sub-task for the teacher; complementarity, when the AI tool amplifies a teacher's capabilities while the teacher remains actively involved; augmentation, when the human-AI system accomplishes a task with an improved output that outperforms what either the human or the AI system could have performed alone. While replacement may boost teachers' productivity by saving time, it may also come at a cost in terms of teacher professional development and autonomy but also impoverish human relationships in teaching and learning. Which tasks one would want to automate requires thorough reflection. Complementarity gives a boost to teachers' productivity while the teacher is still in control, but typically without enhancing teachers' competence. Augmentation requires interactions during which both teachers and AI evaluate and critique each other's suggestions and propositions to move towards a shared understanding and mutual development while solving a problem.

Cukurova (2026[34]) presents different examples of human-AI interactions and argues that GenAI tools provide new possibilities for augmenting teachers' professional abilities. For example, Reza et al. (2024[57]) developed a GenAI tool through a co-creation process with 10 mathematics teachers who worked on content creation for adaptive learning platforms. Using a prompting tool, teachers could quickly see how small changes affected the feedback being provided to students. In this case, the human-AI iterations were iterative and evaluative, but reduced the teachers' perceived workload by 50% and shortened the content development process from several months to a few hours. While GenAI holds promise for more reciprocal exchanges thanks to its dialogic functionalities, GenAI tools with the capacity to push

back against inappropriate or suboptimal pedagogical decisions, to question teachers' assumptions, or to propose alternative perspectives grounded in educational theory and evidence remain to be developed. It will require advances in both the cognitive modelling of teaching expertise and the design of teacher-AI interaction interfaces.

The process to design educational GenAI systems also matters. For example, Topali, Ortega-Arranz and Molenaar (2026[15]) provide an example of the different steps of the human-centred design approach for educational GenAI tools for teachers. By involving teachers and students in the design from the start, first by eliciting teachers' and students' uses and expressed needs and suggestions, the developers involved them at different stages of creating a prototype. This involvement ensured the development of tools that align with teacher and student needs but also recognised their autonomy and agency, both during the design and use stages of the GenAI tool.

The exemplar prototype co-designed through this process allows teachers to monitor student-AI interactions and to set the GenAI behaviour to some extent. Teachers have the possibility to easily set the "percentage" of hallucinations of the tool, depending on how much critical thinking they want students to exercise while interacting with it. From the student perspective, this is a general-purpose GenAI chatbot that they use for education. Teachers valued the potential of the tool to increase their insights into student progress and to personalise their feedback, while maintaining pedagogical control. However, some pointed to the risk of added complexity and workload. Meaningful AI integration in teaching requires intentional design of autonomy, where teachers define pedagogical approaches, have settings allowing them to define AI behaviour, and retain responsibility for educational interpretation. GenAI then serves as an assistant within teacher-defined parameters.

Some educational GenAI tools that meet these requirements are already developed and deployed to support teachers, teacher assistants and students in higher education. Baker et al. (2026[35]) provide several examples in Czechia and India, and present an AI teaching assistant that they developed in the United States as a case study: JeepyTA. JeepyTA supports teaching assistants, students and teachers at various tasks: 1) answering logistics questions related to the course (admission, requirements, dates, etc.); 2) providing feedback to student essays based on specific pedagogical goals and rubrics designed by the instructors; 3) responding to student reflections and questions on the course readings and lectures, offering additional clarification, prompting further thinking, and connecting ideas across course materials; 4) providing debugging support to programming code (in courses where programming is not the learning goal); 5) generating a discussion prompt to start weekly discussions among the class and summarising conversations on the discussion forum, sometimes to provide information to teachers, sometimes to make it visible to learners; 6) suggesting ideas for essays and supporting student brainstorming; and acting as different personas during a course.

In order to maintain teachers' autonomy and control of the tool, instructors or teaching assistants can modify the tool settings so that responses are automated (and go directly to students) or reviewed by them first. This helps prevent the provision of incorrect or misleading information, a particular issue in subject areas where misconceptions are highly present on the web and therefore also in the LLM knowledge base. Instructors or teaching assistants can review flagged responses. If a response is inaccurate, they can discard it and reply directly. If the response is mostly correct but needs refinement, they can edit it before posting. When a response is accurate and well-structured, it can be approved with no changes. This additional layer allows JeepyTA to provide timely support while making sure students receive information that is accurate, relevant, and aligned with the course objectives.

While researching the efficacy of JeepyTA and similar GenAI use cases is still underway, existing studies show that JeepyTA reduced median response times to students from around 7 hours to approximately 2 hours, while keeping human oversight in place. It brought some productivity gains leading to a better student experience. Students rated JeepyTA as comparable to human teacher assistants in clarity, accuracy and professionalism, though it was weaker in motivating students or offering higher-level developmental guidance. When used for feedback on essays, JeepyTA raised the proportion of students achieving top grades on essay assignments from roughly 64% to 95%, illustrating that careful prompt design and alignment with instructor expectations can significantly improve revision quality. However, risks include homogenisation of ideas (seen in brainstorming tasks), overreliance, and the temptation for institutions to reduce human teaching agency.

Together, these empirical findings confirm the potential of GenAI to amplify teacher productivity, improve instructional quality, while keeping human oversight and agency central to system design.

How could GenAI enhance the effectiveness of education systems and institutions?

GenAI tools can also support education systems and institutions in ways that do not immediately impact learning outcomes. Like other sectors, GenAI offers opportunities to streamline workflows and improve the operational efficiency

of educational institutions and systems. This *Outlook* focuses on this aspect through three lenses. It shows how GenAI techniques can already support the improvement of "back office" processes in higher education institutions, such as degree recognition, educational programme design, and support for study advisers (Pardos and Borchers, 2026[58]); the development of standardised assessment items (von Davier, 2026[59]); and new opportunities and challenges for scientific research, which are also largely similar for educational research (Guellec and Vincent-Lancrin, 2026[23]).

How can GenAI support more effective pathways within and across institutions?

One of the challenges for higher education institutions is articulating their programmes with those offered by other institutions, domestically and abroad. Recognising equivalences between them is key to ensuring that students can change study paths without losing time and to support student mobility across institutions, for example. Some systems address these issues by designing national degree structures or using common credit systems that ease pathways across institutions and subjects for students. Articulation agreements and exchange programmes between institutions play a similar role. In practice, however, transfer decisions and articulation agreements are still largely human made and time-consuming, requiring faculty and admission officers to review syllabi and programmes and make a judgment on the equivalence between institutions. Simplifying these processes can enhance the systems' efficiency but also their completion and attainment rates, thanks to more flexible pathways for students. School systems can face similar issues when, typically in secondary education, they start offering more tracks and choices of courses to meet students' interests.

Pardos and Borchers (2026[58]) show that models underpinning GenAI can support these equivalency processes by making relationships between courses, within and across institutions, apparent. To this effect, the AI model processes the text description of the courses and/or the past enrolment patterns to provide administrators with better clarity on which courses within their own institutions are close or similar or well-articulated, and which external programmes (and courses) can be considered equivalent to theirs. Current research shows that GenAI tools can be very similar to human judgment at identifying equivalences (and can also uncover new equivalence possibilities), but adoption depends on trust and the way information is displayed to the final decision makers. These models could also provide advice to students about the next steps of their studies, for example by recommending institutions or programmes.

In the same spirit, work on how GenAI can support study advisers is emerging. In education systems and higher education institutions, study advisers typically provide advice on course selection and career pathways. Lekan and Pardos (2025[60]) developed a GPT-driven model that asks first-year university students about their course preferences and career goals, and then gives recommendations with justifications for advisors to review before in-person meetings. Academic advisors rated the suggestions of the GenAI tool favourably, fully agreeing with the GenAI-generated major recommendation 33% of the time, saving them time while maintaining both their professional autonomy and student relationships.

Other functionalities using similar techniques can be mentioned, such as the automatic tagging of learning content according to changing taxonomies. Tagging learning resources such as open educational resources, or the learning content included in an adaptive learning system, is essential for their discoverability and ensuring that they match local curricula.

How can GenAI support the development of standardised assessment?

GenAI is increasingly used in the development and delivery of standardised assessments, including in high stakes contexts. Pardos and Borchers (2026[58]) and von Davier (2026[59]) highlight how LLMs can automate the creation of multiple-choice and short-answer assessment items, particularly when anchored in existing curricular material or when models are initially designed by experts, possibly assisted by GenAI tools. For example, Bhandari et al. (2026[61]) report that ChatGPT-generated algebra items demonstrated comparable psychometric properties as those created by humans. Notably, the LLM-generated items exhibited slightly stronger differentiation between high- and low-ability respondents, suggesting that GenAI can produce assessment content of similar or even superior quality under controlled conditions. This represents time and productivity gains for both national authorities, test developers and potentially instructors to design assessments for students. This can also address known limitations in traditional test banks, for example the persistent challenge of overexposure of items (which makes them less effective). The limitation is that until GenAI stops hallucinating, instructional staff will still need to check every problem before it is seen by a student.

GenAI can also be used to innovate standardised assessments. For example, von Davier (2026[59]) shows how the Duolingo English Test introduced two new types of writing and speaking assessments that would not have been possible without the use of GenAI. One is an interactive writing task, where a chatbot provides mid-task feedback to the test-takers as they are writing a short essay in English, suggesting additional directions and revisions. Another one is an automated assessment of their oral speaking during a dialogue in "natural language" with a GenAI-powered agent. In

the case of high stakes assessments, generative models are just a layer of a more complex architecture with other AI tools and humans.

Finally, GenAI models can be used to evaluate and calibrate the quality of standardised items. Liu et al. (2025[62]) demonstrate that multi-agent AI models, which bring together ensembles of LLMs acting as "synthetic" or "simulated" respondents, can produce response distributions with psychometric properties closely aligned to those of students. These augmentation strategies, such as adding LLM responses to even a small set of human respondent data, suggests that LLM-based calibration can complement limited student response data, reducing costs and accelerating item validation cycles in the development of standardised assessments.

How can GenAI strengthen educational research?

GenAI is having a significant impact on scientific research, with growing use by researchers to write academic papers and assist them at different stages of the research process. Guellec and Vincent-Lancrin (2026[23]) highlight these trends, providing examples of how GenAI intervenes in the research process in the natural sciences, including the elaboration of literature reviews, analysis of large datasets to generate new research questions or hypotheses, and cooperation with humans to generate and perform research and experiments. While GenAI has already had spectacular achievements, such as the generation of the 3D structure of 200 million proteins, it usually still involves intense human supervision and presents some risks for the research enterprise, such as reduced collective originality and the increasing impossibility to keep up with the collective research output.

Improving the quality of learning and effectiveness of education systems requires investment in high-quality research and evidence to inform policy and practice. While information on how education researchers use GenAI is not available, they likely leverage GenAI tools for tasks like writing and editing research papers and performing literature reviews. Beyond these tasks, three areas stand out as particularly promising for GenAI to support educational research. First, while education is a data-rich environment with lots of administrative data collected for the smooth operation of education systems and institutions, those are often under-analysed because of legitimate privacy concerns. GenAI applications can now easily generate synthetic privacy-preserving datasets that reproduce the statistical characteristics of a dataset with very little risk of privacy breach (as the dataset is entirely created). Second, the use of simulated data to augment real ones could also be tested and applied, albeit with caution and when the context makes it appropriate, to supplement real data where survey response rates are too low. Lastly, the rise of multi-agent models based on GenAI (also called "agentic AI") allows new possibilities for research that could be deployed in the case of education, where answering research questions often require interdisciplinary perspectives.

Concluding remarks

The emerging evidence on GenAI highlights its potential to improve the quality and effectiveness of education. It also demonstrates that it carries risks for student learning and for the professional development of teachers. GenAI appears more disruptive than non-generative AI because students have access to general-purpose GenAI platforms. They can and do use them at home to perform their educational assignments. Even if general-purpose GenAI tools are not used in education institutions, their availability outside of school would challenge current educational processes. As a result, education stakeholders should consider how education systems can leverage and/or adjust to GenAI tools.

Many countries have included digital skills, including GenAI literacy, as part of their curricular objectives. Students should acquire GenAI literacy over the course of formal education, mainly to prepare for the labour market and for societies where GenAI will likely continue to play an important role. In some domains, such as computer programming, students' employability depends on their ability to use GenAI to code, even though understanding the core concepts and principles of programming remains a must. Digital content will also increasingly incorporate AI-generated content, hence the importance for all to have some understanding of how GenAI works.

The acquisition of knowledge, skills and attitudes in various domains remains crucial to young people's education, including reasoning, critical thinking, creativity, empathy, curiosity and judgment. While these skills can be acquired with no technology, GenAI could be leveraged by teachers and students for this purpose. Current evidence shows that educational GenAI tools aligned with educational knowledge and science can lead to better learning, but also that teachers can use general-purpose GenAI effectively if they embed it in a clear pedagogical strategy. In that context, GenAI tools could be used at any stage of the educational process, provided the GenAI tools are well designed or used with sound pedagogical purpose.

GenAI tools can also be used or designed to support teachers and other educational staff such as teacher assistants or study and career advisers. While several studies show that they can improve their productivity, allowing them to spend less time on some tasks with the same quality of output, they face the same risks of cognitive offloading and lack of

learning as students. Research and development on educational GenAI is exploring how GenAI tool can maintain the users' autonomy, professional learning and sense of responsibility in the final output. This can take different avenues, from co-creation with end users following the "human-centred design" approach to ensuring that teachers or other educational staff can adjust the tools to their local context and objectives.

The provision of formative feedback to students, a crucial but time-consuming task for teachers, is a good point in case. Current research shows that feedback generated by GenAI, while often not as reliable and consistent as feedback provided by non-generative AI, matches or surpasses human feedback, given its own strengths and limitations. Still, most studies show that students find human feedback more meaningful, trustworthy and motivating. In this context, using GenAI to assist and complement teacher feedback, while they assume full responsibility for this feedback may be the way forward.

Understanding of what makes educational GenAI tools effective, and how this effectiveness compares to non-generative AI tools, is just emerging. Similarly, knowledge about the effective integration of general-purpose GenAI systems in teacher-designed learning scenarios is nascent. Several country initiatives will provide new knowledge on possible approaches (Box 1.2), highlighting the importance of international co-operation and educational research in this area.

Current research on the use of GenAI in education is still limited and could be strengthen by research investment and international co-operation. For example, most current results are based on very short interventions rather than a continuous, repeated use of GenAI tools over longer periods of time.

When considering how to effectively use GenAI in education, here are some key take-aways to consider:

- Successfully performing an educational task with GenAI does not automatically lead to learning;
- Acquiring and demonstrating foundational knowledge and skills in key subjects without the use of general-purpose GenAI remains key;
- GenAI tools, whether educational or general-purpose, should be used within learning scenarios intentionally designed by teachers to achieve specific learning goals;
- When using GenAI tools, teachers and other education staff must continue to exercise their professional judgment and remain responsible for the quality of the output by evaluating, modifying or endorsing the AI-generated output;
- GenAI developers should design education-specific GenAI tools based on educational research and pedagogical knowledge, and involve teachers as well as other stakeholders such as students or parents or teacher unions, as appropriate, in the design process;
- International co-operation on scientific research to assess the impacts of well-defined pedagogical uses of GenAI tools on students' and teachers' learning and well-being will help leverage these tools in an effective way.

Box 1.2. How do countries support the adoption of GenAI tools in education

Beyond issuing strategies and guidelines (see Annex 1.A.1), countries have a variety of innovative projects to support the adoption and use of GenAI in school. While information about the actual use of GenAI in school is still limited, expert guesses note that the adoption of GenAI in the classroom mainly takes places through pilot projects. These pilots are used to explore different uses of GenAI in school and test the pedagogical value of GenAI tools, identify their possible shortcomings and build institutional capacity before scaling.

These pilot projects typically involve collaboration processes, sometimes within country, across countries, or with companies. Across Europe, for instance, many education authorities participate in regional, national or international AI-related projects. They may also involve public-private partnerships. Two types of projects can be highlighted: some focus on the development of educational GenAI tools for education; others focus on the use of these tools in school (usually after some development).

As an example of the actual introduction of educational tools in school, Korea has licensed and made available GenAI-powered tutors to teachers and schools as AI digital learning materials that work like regular intelligent tutoring systems, allowing students to have adaptive practice in several subjects, and teachers to receive feedback on students' possible misconceptions. Such features include adaptive explanation, automated feedback and interactive dialogue aligned with the national curriculum.

With its AI Leap programme implemented in 2025-26, Estonia aims to explore the use of GenAI in upper secondary education, with a system-level approach combining infrastructure, curriculum development, teacher training and partnerships with technology providers. The programme has made available general-purpose GenAI tools to all teachers and will provide free access to high school students to LLM chatbots that are configured to be educational and follow education research principles (and also to stick to Estonian as the language of interaction).

In Greece, selected upper secondary schools pilot the use of ChatGPT Edu through the OpenAI for Greece partnership (launched in 2025). The project includes teacher training and monitoring of the pedagogical impacts of using GenAI.

Other countries focus on the introduction of teacher- and school-facing tools. Slovakia pilots AI assistants for lesson planning and assessment. Finland tests GenAI applications primarily for teacher support and feedback. Japan, Canada and Australia conduct subnational pilots focused on writing support, feedback generation and workload reduction. France develops a "sovereign AI" for education that will support teachers for lesson planning as well as a Chatbot that will provide generic answers on human resource management questions to its 1.3 million teachers, allowing humans to focus on individual cases.

Two examples can illustrate approaches to the development of appropriate GenAI tools for education.

In the United Kingdom (England), the Department for Education's "content store" consolidates curriculum guidance, lesson plans, and anonymised pupil assessments to support the training of AI models, enabling the development of accurate, high-quality, and legally compliant educational GenAI tools tailored for English schools. Safety expectations for GenAI tools were also developed, providing developers with a clear set of expectations that can facilitate adoption by schools.

In the Netherlands, the National Lab on Artificial Intelligence (NOLAI) co-designs and develops educational GenAI tools (among other AI tools) for the education system through a partnership between government, academia, industry and schools (Molenaar and Sleegers, 2023[63]; Topali, Ortega-Arranz and Molenaar, 2026[15]).

Note

1. The OECD Recommendation on Artificial Intelligence (AI) defines AI as follows: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment" (OECD-LEGAL-0449).

2. Comparing sources for students and the entire population gives the same idea. For example, while 55% of higher education students were estimated to use GenAI tools in France in 2023, a survey of the entire population estimated that 23% of the French population had ever used it in 2023 and 32% in 2024 (Ifop, 2024[74]).

References

AHISA (2023), *The use of generative AI in Australian independent schools*, https://www.ahisa.edu.au/AHISA/Advocacy/ Submission_Resources/Submissions_2023/Generative_AI.aspx (accessed on 21 December 2025). [18]

Baker, R., X. Liu, M. Shah, M. Pankiewicz, Y.J. Kim, Y. Lee and C. Porter (2026), "Generative AI as a Teaching Assistant", [35] in OECD (ed.), *OECD Digital Education Outlook: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı and R. Mariman (2024), *Generative AI Can Harm Learning*, https:// [26] hamsabastani.github.io/education_llm.pdf (accessed on 21 December 2025).

Beghetto, R. (2026), "Developing creativity with generative AI: A conversation with Ronald Beghetto", in OECD (ed.), [30] *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.

Boeskens, L. and K. Meyer (2025), "Policies for the digital transformation of school education: Evidence from the Policy [75] Survey on School Education in the Digital Age", *OECD Education Working Papers*, No. 328, OECD Publishing, Paris, https://doi.org/10.1787/464dab4d-en.

Bui, T., N. Ngo and D. Mai (2025), "Applying AI Tools to Enhance Argumentation Skills in Debate Activities for EFL [43] Students", *International Journal of AI in Language Education*, Vol. 2/3, pp. 83-100, https://doi.org/10.54855/ijaile.25236.

Center for Digital Thriving, Common Sense Media and Hopelab (2024), *Teen and Young Adult Perspectives on* [12] *Generative AI: Patterns of use, excitements, and concerns*, Blackwell Publishing Ltd, https://doi.org/10.1111/jcpp.13190.

Chen, A., Y. Wei, H. Le, and Y. Zhang (2024), "Learning-by-teaching with ChatGPT: The effect of teachable ChatGPT [44] agent on programming education", http://arxiv.org/abs/2412.15226.

Chen, A., M. Xiang, J. Zhou, J. Jia, J. Shang, X. Li, D. Gašević and Y. Fan (2025), "Unpacking help-seeking process [28] through multimodal learning analytics: A comparative study of ChatGPT vs Human expert", *Computers & Education*, Vol 226, p. 105198, https://doi.org/10.1016/J.COMPEDU.2024.105198.

Chevalier, A., J. Orzech and P. Stankov (2024), *Man vs Machine: Can AI Grade and Give Feedback Like a Human?*, [46] http://www.iza.org.

Collie, R. and A. Martin (2025), "Teachers' early uptake of genAI in teaching and learning: important questions and [19] answers", *Social Psychology of Education*, Vol. 28/1, p. 93, https://doi.org/10.1007/s11218-025-10052-6.

Compilatio (2023), *L'IA dans l'enseignement : résultats détaillés d'une enquête où étudiants et enseignants confrontent* [8] *leurs regards*, https://www.compilatio.net/blog/enquete-ia-enseignement-2023#enquete (accessed on 21 December 2025).

Cukurova, M. (2026), "A conceptual framework for teacher-AI teaming in education: Harnessing GenAI to enhance [34] teacher agency", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education,* OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.

Dai, W., Y.S. Tsai, J. Lin, A. Aldino, H. Jin, T. Li, D. Gašević and G. Chen (2024), "Assessing the proficiency of large [48] language models in automatic feedback generation: An evaluation study", *Computers and Education: Artificial Intelligence,* Vol. 7, p. 100299, https://doi.org/10.1016/j.caeai.2024.100299.

Darmawansah, D., D. Rachman, F. Febiyani, and G.J. Hwang (2025), "ChatGPT-supported collaborative argumentation: [42] Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills", *Education and Information Technologies*, Vol. 30/3, pp. 3803-3827, https://doi.org/10.1007/s10639-024-12986-4.

De Simone, M., F. Tiberti, M. Barron Rodriguez, F. Manolio, W. Mosuro, and E. Jolomi Dikoru (2025), "From [41] Chalkboards to Chatbots Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria", *Policy Research Working Paper,* No. 11125, World Bank, http://reproducibility.worldbank.org,.

Demszky, D. (2026), "Generative AI tools to support teachers: a conversation with Dorottya Demszky", in OECD (ed.), [37] *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.

Diliberti, M., H.L. Schwartz, S.Y. Doan, A. Shapiro, L.R. Rainey and J.L. Robin (2024), U*sing Artificial Intelligence Tools* [22] *in K-12 Classrooms,* RAND, https://www.rand.org/pubs/research_reports/RRA956-21.html (accessed on 21 December 2025).

Doshi, A. and O. Hauser (2024), "Generative AI enhances individual creativity but reduces the collective diversity of [40] novel content", *Science Advances*, Vol. 10/28, https://doi.org/10.1126/sciadv.adn5290.

European Commission and OECD (2025), *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*, https://ailiteracyframework.org.   [49]

European Schoolnet (2025), *Artificial Intelligence in School Education: An overview of policy priorities*, European Schoolnet.   [51]

Fan, Y., L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li, and D. Gašević (2025), "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance", *British Journal of Educational Technology*, Vol. 56/2, pp. 489-530, https://doi.org/10.1111/bjet.13544.   [27]

Gašević, D. and L. Yan (2026), "Generative AI for human skill development and assessment: implications for existing practices and new horizons", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [25]

Granström, M. and P. Oppi (2025), "Assessing teachers' readiness and perceived usefulness of AI in education: an Estonian perspective", *Frontiers in Education*, Vol. 10, p. 1622240, https://doi.org/10.3389/FEDUC.2025.1622240/BIBTEX.   [17]

Granström, M. and P. Oppi (2025), "Student engagement with AI tools in learning: evidence from a large-scale Estonian survey", *Frontiers in Education*, Vol. 10, p. 1688092, https://doi.org/10.3389/FEDUC.2025.1688092/BIBTEX.   [6]

Guellec, D. and S. Vincent-Lancrin (2026), "Generative AI and the transformation of scientific research", in OECD (ed.), OECD *Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [23]

Heinrich, T. S. Baily, K.W. Chen, J. DeOliveira, S. Park, S. and N.C.H. Wang (2025), "AI-assisted grading and personalized feedback in large political science classes: Results from randomized controlled trials", *PLOS ONE*, Vol. 20/8 August, https://doi.org/10.1371/journal.pone.0328041.   [45]

Huang, C., Y. Zhong, Y. Li, X. Wang, Z. Han, D. Zhang, and M. Liu (2025), "Enhancing student reading performance through a personalized two-tier problem-based learning approach with generative artificial intelligence", *Humanities and Social Sciences Communications*, Vol. 12/1, https://doi.org/10.1057/s41599-025-04919-4.   [39]

Hüsch, M., N. Horstmann and A. Breiter (2025), *Künstliche Intelligenz im Studium – die Sicht von Studierenden im Wintersemester 2024/25*, https://hochschuldaten.che.de/kuenstliche-intelligenz-im-studium-die-sicht-von-studierenden-im-wintersemester-2024-25/.   [10]

Hu, X., S. Xu and R. Graesser (2025), *Generative AI in Education: From Foundational Insights to the Socratic Playground for Learning*, https://arxiv.org/abs/2501.06682.   [54]

Ifop (2024), *Baromètre 2024 « Les Français et les IA génératives » Vague 2*, https://www.ifop.com/wp-content/uploads/2024/07/120717-rapport-reduit.pdf (accessed on 21 December 2025).   [74]

Impact Research (2024), *AI Chatbots in Schools: Findings from a Poll of K-12 Teachers, Students, Parents, and College Undergraduates*, https://static.waltonfamilyfoundation.org/bf/24/cd3646584af89e7c668c7705a006/deck-impact-analysis-national-schools-tech-tracker-may-2024-1.pdf (accessed on 30 December 2025).   [13]

INDIRE (2025), *L'intelligenza artificiale a scuola: come la usano i docenti? Risultati indagine Tecnica della Scuola e INDIRE*, https://www.tecnicadellascuola.it/lintelligenza-artificiale-a-scuola-come-la-usano-i-docenti-risultati-indagine-tecnica-della-scuola-e-indire.   [20]

Isotani, S. (2026), "AI in Education Unplugged: A Conversation with Seiji Isotani", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [2]

Kaliisa, R., K. Misiejuk, S. López-Pernas and M. Saqr (2025), "How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions", *Educational Psychology*, https://doi.org/10.1080/01443410.2025.2553639.   [47]

Kestin, G., K. Miller, A. Klales, T. Milbourne and G. Ponti (2025), "AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting", *Scientific Reports*, Vol. 15/1, p. 17458, https://doi.org/10.1038/s41598-025-97652-6.   [38]

Kosar, T., D. Ostojić, Y.D. Liu and M. Mernik (2024), "Computer Science Education in ChatGPT Era: Experiences from an Experiment in a Programming Course for Novice Programmers", *Mathematics*, Vol. 12/5, p. 629, https://doi.org/10.3390/math12050629.   [50]

Kosmyna, N., E. Hauptmann, Y.T. Yuan, J. Situ, X.-H. Liao, A.V. Beresnitzky, I. Braunstein and P. Maes (2025), *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*, MIT, https://arxiv.org/abs/2506.08872 (accessed on 21 December 2025).   [29]

Lekan, K. and Z. Pardos (2025), "AI-Augmented Advising", *Journal of Learning Analytics*, Vol. 12/1, pp. 110-128, https://doi.org/10.18608/jla.2025.8593. [60]

Licardo, M., E. Kranjec, A. Lipovec, K. Dolenc, B. Arcet, A. Flogie, D. Plavčak, M. Ivanuš Grmek, B. Bednjički Rošer, B. Sraka Petek, and M. Laure (2025), *Generativna umetna inteligenca v izobraževanju: Analiza stanja v primarnem, sekundarnem in terciarnem izobraževanju,* Univerza v Mariboru, Univerzitetna založba, https://doi.org/10.18690/um.pef.1.2025. [21]

Liu, Y., S. Bhandari and Z. Pardos (2025), "Leveraging <scp>LLM</scp> respondents for item evaluation: A psychometric analysis", *British Journal of Educational Technology*, Vol. 56/3, pp. 1028-1052, https://doi.org/10.1111/bjet.13570. [62]

Liu, Y., J. Huang and H. Wang (2025), "Who on Earth Is Using Generative AI? Global Trends and Shifts in 2025", *Policy Research Working Paper*, No. 11231, World Bank, http://reproducibility.worldbank.org,. [3]

Liu, Y. and H. Wang (2024), "Who on Earth Is Using Generative AI?", *Policy Research Working Paper*, World Bank, http://www.worldbank.org/prwp. [4]

Li, Y. and X. Hu (2026), "Learning with dialogue-based AI tutors: Implementing the Socratic method with Generative AI", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [32]

Molenaar, I. and P. Sleegers (2023), "Multi-stakeholder collaboration and co-creation: towards responsible application of AI in education", in OECD (ed.), *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [63]

OECD (2025), *Results from TALIS 2024: The State of Teaching, TALIS*, OECD Publishing, Paris, https://doi.org/10.1787/90df6235-en. [16]

OECD (2023), *Country Digital Education Ecosystems and Governance: A Companion to Digital Education Outlook 2023*, OECD Publishing, Paris, https://doi.org/10.1787/906134d4-en. [53]

OECD (2023), *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*, OECD Publishing, Paris, https://doi.org/10.1787/c74f03de-en. [52]

OECD (2021), *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots,* OECD Publishing, Paris, https://doi.org/10.1787/589b283f-en. [1]

Oggenfuss, C. and S. Wolter (2024), "Monitoring the digitalisation of education from the students' perspective: Supplementary report with results of the four surveys 2020–2024", *SKBF Staff Paper 26*, SKBF, Aarau, https://www.skbf-csre.ch/fileadmin/files/pdf/staffpaper/Staffpaper_26_e.pdf (accessed on 20 December 2025). [5]

Pardos, Z., S. Bhandari and I. Anastasopoulos (2026), "PromptHive: Demonstrating Collaborative, Human-Centered OER Creation with LLMs", https://doi.org/10.1007/978-3-032-03873-9_47. [61]

Pardos, Z. and C. Borchers (2026), "AI in institutional workflows: Learning from higher education to unlock new affordances for education systems and institutions", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [58]

Pascal, F., F. Taddéi, M. de Falco and É.-P. Gallié (2025), *IA et Enseignement Supérieur: Formation, Structuration et Appropriation par la Société, Ministère chargé de l'Enseignement Supérieur et de la Recherche*, https://mission-ia-sup.forge.apps.education.fr/#/ (accessed on 21 December 2025). [9]

Pew Research Center (2025), *Teens, Social Media and AI Chatbots 2025*, Pew Research Center, https://www.pewresearch.org/wp-content/uploads/sites/20/2025/12/PI_2025.12.09_Teens-Social-Media-AI_REPORT.pdf (accessed on 30 December 2025). [14]

Reza, M., I. Anastasopoulos and S. Pardos (2024), *PromptHive: Bringing subject matter experts back to the forefront with collaborative prompt engineering for educational content creation*, https://arxiv.org/abs/2410.16547. [57]

Rong, H. and C. Chun (2025), *Digital Education Council Global AI Faculty Survey 2025*, Digital Education Council, https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-faculty-survey (accessed on 21 December 2025). [24]

Rong, H. and C. Chun (2024), *Digital Education Council Global AI Student Survey 2024*, Digital Education Council, https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024 (accessed on 21 December 2025). [11]

Roy, P., H. Poet, R. Staunton, K. Aston and D. Thomas (2024), *ChatGPT in Lesson Preparation - evaluation report*, Education Endowment Foundation, http://www.educationendowmentfoundation.org.uk. [56]

Strauß, S. and N. Rummel (2026), "Fostering collaborative learning and promoting collaboration skills: What generative AI could contribute", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [33]

Topali, P., A. Ortega-Arranz and I. Molenaar (2026), "Transitioning from general-purpose to educational-oriented GenAI: maintaining teacher autonomy", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [15]

Vinchon, F., T. Lubart, S. Bartolotta et al. (2023), "Artificial Intelligence & Creativity: A Manifesto for Collaboration", **The Journal of Creative Behavior**, Vol. 57/4, pp. 472-484, https://doi.org/10.1002/jocb.597.   [31]

Vodafone Foundation (2025), *AI in European Schools A European report comparing seven countries*, Vodafone Foundation, https://skillsuploadjr.eu/docs/contents/AI_in_European_schools.pdf (accessed on 30 December 2025).   [7]

von Davier, A. (2026), "Generative AI for standardised assessments: A conversation with Alina von Davier", in OECD (ed.), *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education,* OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [59]

Vygotsky, L. (1978), *Mind in society: The development of higher psychological processes*, Harvard University Press.   [55]

Wang, R., A.T Ribeiro, C.D Robinson, S. Loeb and D. Demszky (2024), "Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise", *EdWorkingPaper: 24-1056*, Annenberg Institute at Brown University, https://doi.org/10.26300/81nh-8262.   [36]

Annex 1.A. Examples of country strategies and frameworks on generative AI in education

Since the public release of generative artificial intelligence (GenAI) tools in late 2022, education systems across OECD countries have expanded or updated their earlier strategies addressing artificial intelligence (AI) in general, with some recent policy documents or initiatives specifically about GenAI. Across countries, policy responses converge around three dimensions: the development or update of long-term strategies and of practical guidance and guardrails; the development of initiatives that address perceived GenAI challenges; and curriculum integration, literacy, and professional development. Examples of the two latter are included in Boxes 1.1 and 1.2. This Annex focuses on countries' strategies and guidance.

Most OECD countries entered the GenAI debate with pre-existing national AI or digital strategies. Since 2023, many have updated these frameworks or issued education-specific documents that explicitly address generative tools (OECD, 2023[52]) (Boeskens and Meyer, 2025[75]).

The most common immediate policy response to GenAI has been the development of national or system-level guidance. These documents typically focus on ethical and responsible use, academic integrity, data protection, and the roles and responsibilities of teachers and students.

Of the 23 countries that responded to a European survey in 2025, 10 OECD and accession countries (Belgium (Flemish Community), Croatia, Czechia, Finland, France, Hungary, Ireland, Italy, Norway and Türkiye) reported that generative artificial intelligence was formally addressed in their system's existing or planned strategies, and 9 were developing or planning to develop guidance or policies to address the use of generative artificial intelligence in education (Greece, Latvia, Lithuania, Luxembourg, the Netherlands, Portugal, Slovakia, Slovenia and Switzerland) (European Schoolnet, 2025[51]).

The report notes that, in the European Union, national strategies are increasingly aligning (and aligned) with the EU AI Act, even though its educational implications are still under review. Most countries have clarified accountability, human oversight and transparency requirements for GenAI use in education. Across systems, restrictions on GenAI tend to be targeted rather than generic. Some countries regulate specific tools or contexts (e.g. assessment), while others rely on device-use policies that indirectly limit GenAI access. Overall, guidance documents emphasise enabling informed professional judgement rather than imposing blanket bans (European Schoolnet, 2025[51]).

Specific GenAI strategies or practical guidelines

Australia: the Framework for Generative Artificial Intelligence in Schools highlights six principles, including diversity of perspectives, non-discrimination, privacy and data protection, and human oversight. Unlike earlier AI strategies, it addresses classroom uses of text- and image-generating systems.

Japan: the government provides guidance on generative AI in schools that explicitly warns against inputting personal or sensitive data into generative AI systems, reflecting early recognition of LLM-specific data reuse and retraining risks in education.

United Kingdom (Wales): the document on Generative artificial intelligence in education: opportunities and considerations is explicitly focused on the use of generative AI, complemented by safeguarding guidance addressing AI-specific risks such as deepfakes and synthetic media.

GenAI as part of AI strategies or guidelines

Most countries embed GenAI within broader digital education strategies.

Finland: the Finnish National Agency for Education issued Recommendations on the Use of Artificial Intelligence in Education, covering both practical guidance and policy reference for GenAI.

Lithuania is preparing Recommendations for the Safe Implementation of Artificial Intelligence in Schools, including a national register of approved AI tools and teacher competence frameworks.

France: the Cadre d'usage de l'intelligence artificielle en éducation (2025) (Framework on the use of AI in education) clarifies permitted uses of GenAI in school and sets conditions related to data protection, transparency and pedagogical responsibility.

Italy: the Linee guida per l'introduzione dell'intelligenza artificiale nelle istituzioni scolastiche (2025) (Guidelines on the adoption of AI in educational institutions) focus on safe and conscious adoption of generative tools to support teaching and organisational processes.

Ireland: the Guidance on Artificial Intelligence in Schools (2025) complements the revised National AI Strategy (2024) by providing education-specific principles for responsible GenAI use.

Latvia: the Guidelines for the Use of AI in Primary and Secondary Education explicitly address generative tools and critical thinking.

Luxembourg: the KI Kompass (AI Compass) serves as a national reference framework combining views on strategic orientation, pedagogical practice and professional exchange.

The Netherlands: a Vision on Generative Artificial Intelligence (2024), followed by the Dutch Digitalisation Strategy – Accelerating Together (2025) provides a cross-sectoral strategy, while education-specific guidance is provided by Kennisnet through the Guide to AI in Education and School Agreements on the Use of Generative AI.

Norway: the Strategy for Digital Competence and Infrastructure in Early Childhood Education and Schools addresses GenAI and is accompanied by national guidelines and competency development packages.

Slovakia: the Strategic Plan for the Integration of Artificial Intelligence into Education includes dedicated initiatives on AI assistants for teachers and personalised learning tools.

Spain: the Guide on the Use of Artificial Intelligence in Education (2024) covers GenAI among other AI tools, for example to discourage the use of non-compliant generative tools.

Türkiye: the Artificial Intelligence in Education Policy Document and Action Plan 2025–2029 explicitly covers the use and access of generative AI alongside ethics and capacity building considerations.

United States: at the federal level, the 2025 Executive Order on Advancing Artificial Intelligence Education for American Youth directs the establishment of a White House Task Force on Artificial Intelligence Education to coordinate federal efforts on integrating AI, including generative AI, into primary, secondary and post-secondary education. The order promotes appropriate integration of AI into curricula, comprehensive AI training for educators, and an annual AI Challenge to foster student and teacher innovation with generative AI models, while seeking public-private partnerships to develop online resources focused on foundational AI literacy and critical thinking skills. Further information on guidance at the State level can be found in (Teach AI 2025)

International guidelines on AI and GenAI

European Commission: In addition to its work on AI literacy, the European Commission published expert guidance on the use of AI and data in education (2022) as well as Ethical Guidelines on the Use of Artificial Intelligence and Data for Teaching and Learning for Educators (2022), whose revised version will come out in 2026 to respond to developments such as the emergence of GenAI.

UNESCO: UNESCO released in 2024 and 2025 some guidance on the adoption of GenAI in education (Guidance for generative AI in education and research, 2025) as well as advice more targeted towards teachers (AI competency framework for teachers, 2024).

UNICEF: In December 2025, UNICEF updated its guidance on AI and children to respond to rapid technological advances like generative AI and new concerns such as AI-generated harmful content. The framework outlines 10 principles to ensure AI technologies are safe and supportive of child wellbeing.

OECD and Education International: In 2023, the Secretariats of the OECD and Education International, the international federation of teacher unions, released a joint paper "Opportunities, Guidelines and Guardrails for Effective and Equitable Use of AI in Education" that covers all forms of AI, including generative AI, and provides a starting point for discussion between teacher unions and jurisdictions.

2

Generative AI for human skill development and assessment:
Implications for existing practices and new horizons

Dragan Gašević* and Lixiang Yan**

*Monash University, Australia
**Tsinghua University, China

Generative artificial intelligence (GenAI) is transforming the landscape of education by reshaping how skills are developed, assessed, and supported. This chapter synthesises recent empirical evidence on how GenAI technologies influence instructional practices, feedback, and assessment. It examines both the opportunities and limitations of using GenAI to provide personalised tutoring, enhance feedback quality, and automate assessment practices. The chapter argues for a careful balance between human skill development and AI-augmented performance, emphasising the need for pedagogically grounded integration of GenAI within intelligent tutoring and assessment frameworks. It concludes by outlining directions for research and policy that ensure GenAI strengthens, rather than substitutes, human learning and instructional expertise.

Introduction

The wide adoption of generative artificial intelligence (GenAI) – after the public release of ChatGPT in November 2022 – has triggered profound debates about their implications on education. GenAI can provide technologies that can support skill acquisition through personalised instruction and feedback, and enhance the efficiency and effectiveness of teaching practices (Giannakos et al., 2024[1]). However, GenAI poses ethical challenges and risks as well (Kofinas, Tsay and Pike, 2025[2]; Nikolic et al., 2023[3]). The developments in GenAI triggered educators, education leaders, and policymakers to engage with GenAI extensively, rethink pedagogical, assessment, and governance frameworks to harness GenAI's potential while mitigating its risks. Through these efforts, many education institutions have produced policies and guidelines to support staff and students in using generative AI (Jin et al., 2025[4]). Similarly, many government, intergovernmental, nongovernmental, and non-for-profit organisations have also produced documents that inform GenAI adoption, responsible practices, and frameworks for professional development of educators (Cardona and Ishmael, 2023[5]; Miao and Cukurova, 2024[6]; Miao and Holmes 2023[7]; OECD, 2023[8]). Equally so, the rapid developments in GenAI have also mobilised many researchers to study implications on education and human learning (Yan et al., 2023[9]; Zhang et al., 2024[10]).

This chapter aims to summarise recent evidence about the implications of GenAI in human skill development and assessment. The focus will be on human skill development and assessment as they are central to education and professional development programs. The analysis of the implications of GenAI on human skill development and assessment is particularly framed around two complementary perspectives. First, GenAI technologies offer some promising prospects for advancing our existing practices related to skill development and assessment. For example, GenAI can be used to provide interactive instructional support, provide personalised feedback at scale,

and automate the creation and implementation of assessments (Giannakos et al., 2024[1]; Yan et al., 2024[11]; von Davier, 2026[12]). Second, GenAI challenges our existing assumptions of our learning practices and calls for novel ways for assessment. For example, while GenAI can increase performance in certain situations, it can also limit human agency and result in overreliance on AI (Darvishi et al., 2024[13]; Fan et al., 2024[14]; Stadler, Bannert and Sailer, 2024[15]). Finally, we need to strengthen research methods that are used to study human skill development and assessment in the age of GenAI to avoid challenges recently noted in the literature, such as conflation of learning and performance (Weidlich, 2025[16]; Yan et al., 2025[17]).

This chapter is based on the analysis of empirical evidence published in the research literature. It offers a summary of the existing evidence about effectiveness of GenAI to support existing practices for instruction, feedback, and assessment given their central roles in education and professional development. It also describes the recent conceptualisation of hybrid human-AI skills that recognises the need to support development of human skills while enhancing task performance with the use of GenAI. The chapter concludes by providing implications for practice and policy grounded in existing evidence and promising directions for future research. Box 2.1 provides a glossary of the main terms and types of generative AI and associated techniques in the field of AI in education (AIED).

Box 2.1 Glossary

Generative Artificial Intelligence (GenAI)

A branch of artificial intelligence focused on creating new content, such as text, images, or audio, based on patterns learned from existing data. GenAI systems, like large language models, can generate human-like responses, creative works, and solutions in various formats.

Large Language Models (LLMs)

A type of machine learning model, often based on deep learning, that is trained on vast amounts of text data. LLMs are capable of generating coherent and contextually relevant text, answering questions, and performing a wide range of language tasks (e.g. GPT, BERT).

Diffusion Models

A class of generative models that create high-quality data (such as images or sound) by simulating a gradual process of adding noise and then reversing it to recover the original signal. They are particularly known for their ability to generate realistic, high-resolution images.

GPT (Generative Pre-trained Transformer)

A type of large language model that uses a transformer architecture. GPT models are pre-trained on vast datasets and fine-tuned for specific tasks, capable of generating human-like text across a variety of contexts.

BERT (Bidirectional Encoder Representations from Transformers)

A pre-trained transformer-based language model designed for natural language understanding tasks. Unlike GPT which is a unidirectional model, BERT processes text bidirectionally, improving its ability to understand context and meaning in sentences. Its "understanding" of text context has shown promising results for text analysis in education.

Prompt and Prompt Engineering

A prompt is the input text or instruction given to a GenAI model to produce a specific output. Prompt engineering involves designing and refining prompts to achieve the desired responses or behaviours from the GenAI model, often through trial and error.

Zero-shot Prompting

A method of prompting where the model is expected to perform a task without having been explicitly trained on that task.

First-shot Prompting

A variant of prompt engineering where the model is given a single example or instruction before performing the task.

Chain-of-thought Prompting

A prompting technique where the model generates intermediate reasoning steps before providing the final answer, helping improve accuracy on complex tasks.

Retrieval-Augmented Generation (RAG)

A hybrid approach combining information retrieval-based techniques with generative AI models. RAG models retrieve relevant information from large databases or knowledge sources and use that data to generate more accurate and contextually informed outputs.

Intelligent Tutoring Systems

AI-based educational software that provides personalised instruction and feedback to students. These systems simulate one-on-one tutoring by adapting content, pace, and difficulty based on the learner's needs and performance.

Learning Analytics

The measurement, collection, analysis, and reporting of data about learners and their contexts to understand and optimise learning and the environments in which it occurs. It leverages data to improve teaching practices and student outcomes.

Learning Analytics Dashboard

A visual tool that displays real-time data and insights about learners' progress, behaviour, and engagement. Dashboards typically provide instructors and students with actionable feedback to support decision-making and interventions.

Feedback

Feedback is "a process in which learners make sense of information about their performance and use it to enhance the quality of their work or learning strategies" (Henderson et al., 2019[18]).

Feedback Literacy

The ability to understand, interpret, and effectively use feedback in the learning process. Feedback literacy involves knowing how to respond to feedback, act on it, and integrate it into one's learning strategies.

Existing practices

Enhanced instructional support

Directly supporting students

Providing enhanced instructional support at scale is one of the most prominent areas for the use of GenAI in education (Yan et al., 2024[19]). This is grounded in the idea of making use of GenAI for developing systems that can offer personalised learning support. The idea of personalised learning support is grounded in Bloom's (1984[20]) "two-sigma problem" showing the significant benefits of one-to-one instruction over other forms of instruction. Before widespread use of GenAI, the effectiveness of personalised learning support has been long studied in the literature on artificial intelligence in education (du Boulay, Mitrovic and Yacef, 2023[21]), particularly focusing on intelligent tutoring systems (Graesser, Conley and Olney, 2012[22]; Molenaar, 2021[23]) and resulted in the development of many effective tutoring systems – e.g. SQL-Tutor (Mitrovic and Ohlsson, 2015[24]), MetaTutor (Azevedo et al., 2022[25]), and Cognitive Tutors (Koedinger and Aleven, 2015[26]). Especially relevant to today's attempts to provide personalised learning support are intelligent tutoring systems such as AutoTutor (Nye, Graesser and Hu, 2014[27]) and BEETLE

(Dzikovska et al., 2014[28]) that were already designed to provide tutoring through dialogue in natural language. This research also informed the development of many commercial tutoring systems such as MATHia based on Cognitive Tutor (Almoubayyed et al., 2023[29]) and ALEKS (Fang et al., 2018[30]). However, rapid development of such systems still remains a challenge.

GenAI offers promising approaches that can be used for rapid development of instructional systems for personalised support. Specifically, GenAI, through the use of large language models, can be leveraged to develop tutoring chatbots. A prominent example is Khan Academy's Khanmigo chatbot, which makes use of large language models to conduct scaffolded, Socratic-style tutoring across diverse subject areas (Khan, 2025[31]). As one of many emerging GenAI chatbots in education, Khanmigo illustrates how these technologies can scale personalised learning support and expand opportunities for learner autonomy and exploration. However, at the time of writing of this chapter, there have been no studies published that evaluated the effectiveness of Khanmigo on learning (although at least one pre-registered RCT is ongoing in Canada).[1]

Evidence on the effectiveness of GenAI to enhance instructional support is still emerging and offers mixed support. For example, a randomised controlled trial conducted at Harvard University showed the significant effects (0.73-1.3 standard deviations) of an AI Tutor – ChatGPT powered system – over those attending in-person active learning classes in an undergraduate physics course (Kestin et al., 2025[32]). The World Bank has recently reported the findings of a randomised controlled trial in nine secondary schools in Nigeria (De Simone, 2025[33]). In the trial, students were randomised in the treatment group that received access to Microsoft Copilot based on GPT-4 in an after-school programme and the control group who did not have access. The students in the treatment group received teacher instructions on how to use Copilot including the prompts and worked in pairs with other students. The results showed the positive effects of this intervention with the effect size of 0.31 standard deviations. However, this effect size was lower than the average effect size noted in the meta-analyses of the effectiveness for intelligent tutoring systems – i.e. 0.42–0.57 standard deviations according to Ma et al. (2014[34]) and 0.66 according to Kulik and Fletcher (2016[35]). This suggests that past AI tutors might be more effective, albeit the difference in context. Nevertheless, the World Bank study findings aligns with the range observed in promising computer-assisted learning interventions reviewed by Escueta et al. (2020[36]), who identified effect sizes between 0.18 and 0.63 standard deviations for personalised and adaptive programs, particularly in mathematics. The World Bank study also showed that the students with high prior academic performance (as also shown by Lehmann et al. (2025[37]) and high socio-economic status particularly benefited from the interventions. While these findings come from a context where socio-economic disparities are likely more pronounced than in most OECD countries, they still suggest that GenAI-based tutoring systems may disproportionately benefit certain groups of students. Future research in diverse educational settings is needed to corroborate this pattern.

The way how a GenAI-based tutoring system is configured and used may have profound implications on learning. In a large-scale field experiment in high school math classrooms, Bastani et al. (2024[38]) found that while GPT-4-based tutors improved performance during use (up to 127%), students who used a standard chatbot akin to ChatGPT performed worse (17% lower performance) than the control group once access to the chatbot was removed. The control group students did not use any GenAI-based instructional support in addition to the conventional classroom instruction. This negative effect of GenAI use was mitigated by a version designed with learning safeguards, suggesting that poorly configured systems may undermine long-term learning.

Similarly, Lehmann et al. (2025[37]) showed that a ChatGPT-based tutor for Python programming had no overall effect on learning, but its impact depended on usage patterns. Students who heavily relied on the ChatGPT-based tutor tended to cover a broader range of optics but developed shallower understanding, while those who used it to complement learning gained deeper understanding. The use of the ChatGPT- based tutor also widened performance gaps between students with high and low prior knowledge. In summary, the Bastani et al. (2024[38]) and Lehmann et al. (2025[37]) studies highlighted the importance of instructional strategies embedded in the design of GenAI-based instructional systems and the way how students use them are two key factors that need to be considered in research and practice. Future research should also examine how best to combine generative and conventional AI models, since, adding LLMs investigate effectiveness of different pedagogical approaches GenAI-based instructional systems use and factors (e.g. metacognitive skills) that explain different usage patterns of students. It is also important to examine how conventional and generative AI models can be effectively integrated. Such integration should enhance human learning by combining the strengths of each layer while minimising risks, including hallucinations that may compromise reliability (Qian et al., 2026[39]).

Effective instructional support from GenAI-powered systems requires alignment with the proven experience in developing intelligent tutoring systems. Although frequently mentioned in recent developments, GenAI-powered instructional systems do not follow a typical architecture of intelligent tutoring systems. At the

core of intelligent tutoring systems are *learner models, tutor models, domain models, and user interface* (Nkambou, Bourdeau and Mizoguchi, 2010[40]). Most GenAI–based instructional systems primarily make use of large language models to cover functions of all these four components. While user interfaces through natural language interaction can be quite advanced with LLMs, the support for the other three components is less obvious. Although the functions of the domain and tutor models can be performed by LLMs to some extent, there is presently limited research and evidence on how their quality can be assured. Specifically, due to the stochastic nature of LLMs, they cannot guarantee reliability of information covered in the domain model due to tendency to hallucinate (Ji et al., 2023[41]). Existing research shows that LLMs can easily be distracted and inconsistently comply with the instructions provided in the underlying prompts (Hwang et al., 2025[42]; Liu et al., 2025[43]; Zhao et al., 2024[44]). Therefore, future research is needed to assess the extent to which LLMs can consistently comply with a particular tutoring strategy to offer long-term effects. Moreover, future work is needed to develop effective computational approaches that can increase compliance of LLMs to promote effective tutoring strategies over time. Finally, there is very little evidence in the literature that existing GenAI-based instructional systems offer any learner models. They are precisely needed to understand individual student needs based on the tracing of their knowledge development (Abdelrahman, Wang and Nunes, 2023[45]) and learning approaches they take to provide adaptive and personalised support. Recent evidence by Borchers and Shou (2025[46]) shows that LLM-only tools offer only limited adaptivity compared to conventional intelligent tutoring systems. Future research is needed to address these critical architectural needs and identify effective ways of the integration of LLMs within tutoring architectures to enhance instructional effectiveness.

Supporting teachers

Enhancing instructional support does not necessarily need to be done through providing direct instruction to students. Teachers can also be beneficiaries of GenAI for tasks related to preparation for teaching and during the actual act of teaching. For example, GPTeach is an interactive teacher training tool that enables novice educators to practice teaching with GPT-simulated students. Evaluations of GPTeach have shown that it can enhance teachers' preparedness and confidence, offering valuable practice opportunities tailored to varied teaching scenarios (Markel et al., 2023[47]). Relatedly, Tutor Copilot is a GenAI-powered system that provides real-time, expert-like guidance to tutors during live tutoring sessions (Wang et al., 2025[48]). This approach is particularly relevant for supporting students by mobilising a less experienced workforce and addressing the issue of teacher shortages. In a randomised controlled trial, Wang et al. (2025[48]) evaluated the effectiveness of Tutor CoPilot. The study involved 900 tutors and 1 800 K-12 students from historically under-served communities. Results indicated that students whose tutors had access to Tutor CoPilot were 4 percentage points more likely to master topics, with the most significant benefits observed among students of lower-rated tutors, who experienced a 9 percentage point improvement. Additionally, tutors using Tutor CoPilot were more inclined to employ high-quality pedagogical strategies, such as asking guiding questions, and less likely to provide direct answers. Although the studies with GPTeach and Tutor CoPilot show much promise, future research is needed to understand the uptake and effectiveness of such systems in diverse educational and international contexts. Equally important is future research to understand how effectively tutoring practices supported by Tutor CoPilot are internalised by teachers over time as part of their professional development, and whether they may lead to overreliance on GenAI, potentially hindering the development of teachers' human teaching skills.

GenAI can support teachers in a range of tasks, with mixed results regarding their effectiveness and efficiency. Although lesson planning is frequently discussed as one of the key areas of teaching practice that can benefit from the use of GenAI (Moundridou, Matzakos and Doukakis, 2024[49]), evidence about its effectiveness is still emerging. For example, Dennison et al (2025[50]) evaluated Shiksha Copilot, an AI- assisted lesson planning tool deployed in schools in India. In a large-scale mixed-methods study, including interviews, surveys, and usage logs, the study found that teachers used Shiksha Copilot to meet administrative documentation needs and support their teaching. The use of the tool was associated with a reduction in lesson planning time, with small to large effect sizes (Cohen's d = 0.371 – 0.658), and lowered teaching-related stress (Cohen's d = 0.436), while promoting a shift toward activity-based pedagogy. However, systemic challenges, such as staffing shortages and administrative demands, constrained broader pedagogical change. In contrast, Selwyn et al. (2025[51]) conducted interviews with teachers about their experiences with GenAI tools for administrative tasks in Sweden and Australia, highlighting the significant work teachers self-report investing in reviewing, repairing, and reworking AI-generated outputs. Their findings suggest that the promise of time-saving in AI tools may overlook the complex professional judgments teachers must make regarding pedagogical appropriateness, social relations, and educational value. However, Selwyn et al.'s (2025[51]) findings are based on self-reports (i.e. interviews), which do likely not reliably estimate time spent on technology use (Parry et al., 2021[52]). Usage log analysis, on the other hand, offers a more accurate and less biased approach. In contrast to the Dennison et al. (2025[50]) study, which is grounded in usage log analysis to provide more reliable usage time estimates, the Selwyn et al. (2025[51]) study highlights the need to account for the hidden labour of teachers that may not be captured by usage logs. Yet, given that Dennison et al.

(2025[50]) compared GenAI-supported lesson planning with a non-GenAI baseline, some of this hidden labour may already have been reflected in their analysis. This highlights the importance of fair and contextually comparable evaluation frameworks that consider how GenAI tools are implemented and how teacher time use is measured across studies. Given the essential role teachers play in education, this underscores the importance of exploring design principles, organisational adoption strategies, and the broader implications of adopting GenAI technologies for teaching support.

Improving feedback practices

One of the most pressing areas of application for GenAI in education is the provision of automated feedback. Feedback represents a persistent challenge in higher education, where increasing student numbers are not matched by proportional increases in teaching resources (Paris, 2022[53]). It is also a challenge at school level in contexts where student/teacher ratios are high, or when teachers teach a subject with few curriculum hours (and thus many classes and students). This structural tension has made it difficult to offer timely, targeted, and individualised feedback at scale (Pardo et al., 2017[54]) that follows principles of effective (Hattie and Timperley, 2007[55]; Henderson et al., 2019[18]) and learner centred feedback (Ryan and Henderson, 2021[56]). Feedback can improve learning progression (Wisniewski and Zierer, 2020[57]) and support development of relationships between students and educators (Dai, Tsai and Gašević, 2025[58]; Henderson, 2015[59]). As shown in the remainder of this section, GenAI holds strong promise for enabling the rapid and scalable generation of feedback across multiple modalities, with the potential to enhance feedback scalability, quality, and even feedback literacy (Nieminen, 2023[60]; Baker, forthcoming (2026)[61])

Feedback generation and quality analysis

GenAI has been found to be promising to offer feedback on students' written products in higher education. In a recent study, Dai et al. (2024[62]) compared feedback generated by large language models to that provided by human tutors. Their study compared feedback on readability, similarity of positive and negative points identified, and levels feedback was provided on. The levels of feedback were grounded in Hattie and Timperley's (2007[55]) seminal framework that distinguishes feedback on task (correctness), process (learning strategies), self-regulation (monitoring learning), and self (personal traits and motivation). In this framework, higher-level feedback, particularly at the process and self-regulation levels, is widely recognised as more educationally valuable and a key indicator of feedback quality because it supports deeper learning and learner autonomy. The Dai et al. findings revealed that GenAI (i.e. GPT-3.5 and GPT- 4) tended to produce more readable and stylistically polished feedback with quite a large effect size (d = 1.79) than feedback produced by human educators. This finding was somewhat unsurprising, given that human assessors often operate under strict time constraints and offer rather succinct feedback. However, the study also revealed limited alignment between what GenAI produced feedback and human tutors identified as strengths and weaknesses in student work according to a rubric. Dai et al. (2024[62]) also showed that GPT (particularly GPT-4) models were able to produce feedback that offered guidance about future choice of learning strategies (i.e. process level feedback) in over 97% of feedback instances. Interestingly, this was higher than what was observed in human-provided feedback which was on process level in about 80% of feedback instances. However, GPT-4 was much less able to produce feedback on the self-regulation level, which was only in 17% of cases. However, even that was higher than that provided by human tutors who only offered self-regulation level feedback in 11% of cases. This just highlights the challenge of providing feedback on self-regulation levels where learners are guided to monitor their own learning. This challenge is particularly important in the age of GenAI as discussed later.

GenAI can also be used to generate feedback guiding students based on the insights of predictive modelling. Early prediction of students at risk of failing or dropping out have been at the core of research and practice in learning analytics for a long time (Molenaar, 2021[23]; Gašević et al., 2016[63]; Wang and Mousavi, 2022[64]). However, translating insights from predictive modelling to actionable feedback has received much less attention. A notable example with much success in improving student learning and experience was the OnTask system that allowed educators to manually write rules to generate personalised feedback based on student data (Pardo et al., 2017[54]). Although much more efficient than manual feedback writing at scale, it still could not translate granular insights of predictive modelling to actionable feedback (e.g. advice on which practice exercises to take). To address this challenge, Liang et al. (2024[65]) proposed an approach for transforming insights from predictive modelling to personalised feedback with the use of GPT-4 which was rated by experienced educators as "readily applicable to the course" (Liang et al., 2024[65], p. 75) and higher on readability, relational characteristics, and specificity than human-provided feedback. However, at present, future research is needed to assess the effectiveness of such personalised feedback on learning outcomes, student retention, and the extent to which students actually use and act upon the AI-generated feedback.

GenAI can also be used to check the quality of feedback to promote best practices at scale. Previous research demonstrates the potential of the use of conventional machine learning to recognise whether human produced feedback followed established models for feedback. For example, Osakwe et al. (Osakwe et al., 2022[66]) used a XGboost machine learning model trained on established linguistic features (e.g. cohesion or use of cognitive words) to identify self, task, and process levels of feedback with accuracy values of 0.87, 0.82, and 0.69, respectively. In a recent study, Aldino et al. (2024[67]) evaluated the performance of GPT-3.5 with zero-shot prompts to identify elements of learner-centred feedback on a large dataset of feedback messages (>16k) in higher education. GPT-3.5 showed some promising results with accuracy in the range of 0.53-0.97 across the seven attributes of learner-centred feedback. However, GPT-3.5 was consistently outperformed by conventional machine learning models (i.e. XGBoost and Random Forest) based on linguistic features (e.g. cohesion and word count), while BERT almost always performed reliably (accuracy 0.91-0.99) (see Box 2.1 for definition of BERT). Higher accuracy of traditional machine learning over ChatGPT was also shown in evaluation of the quality of peer feedback (Hutt et al., 2024[68]). Similarly, Dai et al. (2025[58]) showed that GPT-4o was able to identify nine out of 10 relational characteristics of feedback with an average accuracy exceeding 80%. For example, the model successfully recognised feedback that acknowledged students' strengths, offered balanced critical comments, and included actionable suggestions for improvement. Yet, they found no significant increase in the use of few-shot prompting strategies over zero-shot prompting. These findings suggest that while GPT prompting approaches offer a promising and accessible entry point due to their lower technical barrier, achieving consistently high accuracy still requires conventional machine learning methods and language models like BERT.

The differences in performance between GenAI and human educators create new opportunities to complement and enhance the effectiveness of human tutors. For example, GenAI can provide positive and negative points in feedback, which human educators can use as suggestions to enhance their own feedback drafts. This is also suggested by Lu et al. (2024[69]) who argue that GenAI can offer immediate and personalised feedback on lower-order concerns in written products such as grammar, vocabulary, and sentence structure. Their premise is that this may allow teachers to focus on higher-order thinking skills, content depth, and argumentation, where human judgment remains crucial. The results of the Dai et al. (2024[62]) study indicate that GenAI can help enhance human feedback with effective feedback practices (Hattie and Timperley, 2007[55]). This hybrid approach holds potential to enhance efficiency without compromising pedagogical judgment and future research and practice should evaluate its effectiveness. Despite all these promises, specialised tools that promote this hybrid approach for educators are in early days. For example, Feedback Copilot was developed to incorporate principles of co-design to create effective user interfaces that incorporate the use of GenAI (Pozdniakov et al., 2023[70]). Efficacy of Feedback Copilot is yet to be evaluated in practice and highlights the important research gap and direction for future research.

Effectiveness of GenAI feedback

A growing body of research has explored how students perceive and respond to feedback generated by GenAI, particularly in comparison to human feedback. Studies have shown that students tend to act more readily on feedback from human instructors than from GenAI tools (Zou et al., 2025[71]; Lu et al., 2024[69]). Students often found GenAI feedback to be specific and clear though, especially in technical tasks (Ouyang et al., 2024[72]; Roest, Keuning and Jeuring, 2024[73]; Zhang et al., 2024[74]). However, several studies also highlight concerns regarding the perceived usefulness and trustworthiness of GenAI-generated feedback (Escalante, Pack and Barrett, 2023[75]; Er et al., 2024[76]). Although these studies differ in focus and methodological design, ranging from quasi-experimental evaluations of learning outcomes (Escalante, Pack and Barrett, 2023[75]) and randomised controlled comparisons of instructor and AI feedback (Er et al., 2024[76]) to large-scale perception studies in higher education contexts (Nazaretsky et al., 2024[77]), they consistently point to lower perceived usefulness and trust in AI-generated feedback relative to human feedback.

Overall, current evidence suggests that while GenAI feedback can match human feedback in measurable learning outcomes, it does not replicate its pedagogical value or social credibility. For example, Escalante et al. (2023[75]) found no significant difference in learning outcomes between students receiving feedback from GPT-4 and those receiving tutor feedback, although participants were evenly split in their perceptions of usefulness. While this might appear to suggest functional equivalence, comparable performance does not imply pedagogical interchangeability. As shown in the recent meta-analysis by Kaliisa et al. (2025[78]) across 41 studies, AI-generated and human feedback yield statistically similar learning gains, yet students perceived human feedback as more credible and meaningful. This distinction points to the broader role of feedback in shaping motivation, evaluative judgment, and learner trust, dimensions that remain difficult for GenAI systems to reproduce even when outcome measures are equivalent. Similarly, Er et al. (2024[76]) reported that human feedback was

perceived as significantly more useful, and students who received it showed greater improvement in lab scores in Java programming. In a related study, Nazaretsky et al. (2024[77]) found that students' perceptions of feedback varied depending on the provider's identity. When the feedback source was unknown, students rated AI feedback more favourably; however, when the source was revealed, they placed greater trust in human feedback. Although highly relevant to trust, the effects of hallucinations in GenAI on feedback uptake has received little attention in the literature and warrant future attention. Perceptions of fairness have also been somewhat contradictory: while some studies found that GenAI feedback was rated as fair by students (McGowan, Anderson and Smith, 2024[79]; Rudolph et al., 2024[80]), other studies observed the opposite (Er et al., 2024[76]).

GenAI feedback has also shown potential to support important metacognitive processes. For instance, Tang et al. (2024[81]) demonstrated that structured GenAI feedback on writing tasks significantly improved students' self-assessment accuracy, which is a key skill for independent learning. However, this potential is not always realised. Jin et al. (2025[4]) found that students with low feedback literacy engaged only minimally with a GenAI-based support tool, often due to a mismatch between the tool's responses and their expectations. These findings suggest that the impact of GenAI feedback depends not only on its technical qualities but also on learners' readiness to interpret and apply it effectively. As Zhan and Yan (2025[82]) argue, fostering feedback engagement in a GenAI context requires the explicit development of students' feedback literacy, including skills in prompt engineering, evaluative judgment, and metacognition, to facilitate deeper and more meaningful interaction with GenAI in feedback practices. Future research should aim to (a) investigate the extent to which feedback literacy of students can be promoted to more effectively and critically engage with AI-generated feedback and (b) understand whether feedback literacy enables learners to improve their learning outcomes when using AI-generated feedback.

Beyond textual feedback

GenAI can support generation of feedback in different modalities that goes beyond textual feedback. For example, learning analytics offers dashboards as an alternative and cost-effective approach to provide feedback based on analysis of student data (Matcha et al., 2020[83]). However, learning analytic dashboards have not achieved their full potential (Kaliisa et al., 2024[84]). One of the main reasons for this is relatively limited data visualisation literacy of educators and students to understand and translate insights from different statistics and charts into action (Donohoe and Costello, 2020[85]; Pozdniakov et al., 2023[70]). To address the limitations in visualisation literacy, GenAI can offer two complementary approaches.

First, GenAI can provide a layer guiding educators and learners to improve their abilities to comprehend dashboards accurately. For example, Yan et al. (2024[19]) developed a tool called *VizChat*, which allows students and educators to interact with a chatbot to help them understand the data shown in the dashboard by asking questions (Figure 2.1). When configured in a proactive mode (i.e. used scaffolding questions), *VizChat* significantly enhanced the comprehension of learning analytic dashboard compared to both passive chat mode (i.e. responding to student queries a la ChatGPT) and standalone scaffolding (Yan et al., 2025[86]). Importantly, these benefits continued to persist even when the students did not have access to proactive VizChat. Building on these promising results, future research should investigate the extent to which learners and educators can transform insights obtained from learning analytic dashboards into effective learning and teaching practice thanks to GenAI.

Second, GenAI can be used to generate feedback in other forms than text, for example in the form of data comics. *Data comics* follow established principles of comic strip genres (e.g. Manga) and are generated by prompting multimodal language models to generate images based on analytic insights (Milesi et al., 2024[87]). Data comics were applied in simulation-based learning for healthcare professionals, where student nurses engage in highly collaborative learning scenarios in physical spaces (Martinez-Maldonado et al., 2023[88]). Data comics (see Figure 2.2) aim to present feedback in a more accessible, emotionally engaging format. Qualitative evidence suggests that data comics can improve student motivation and reflective engagement; some students even reported feeling seen or valued (Milesi et al., 2024[87]). However, some students in higher education found this approach as not sufficiently professional and potentially perpetuating biases (e.g. all nurses generated in data comics were women and the doctor was a man). Expanding positive aspects of GenAI-powered data comics, future research is needed to understand their effectiveness across different educational contexts and levels, while minimising potential negative effects. The same idea could also be applied to the AI generation of video clips based on the multimodal learning analytics.

Figure 2.1. VizChat – an LLM-based chatbot designed to enhance the ability of leaders and educators to interpret and understand visual learning analytics.
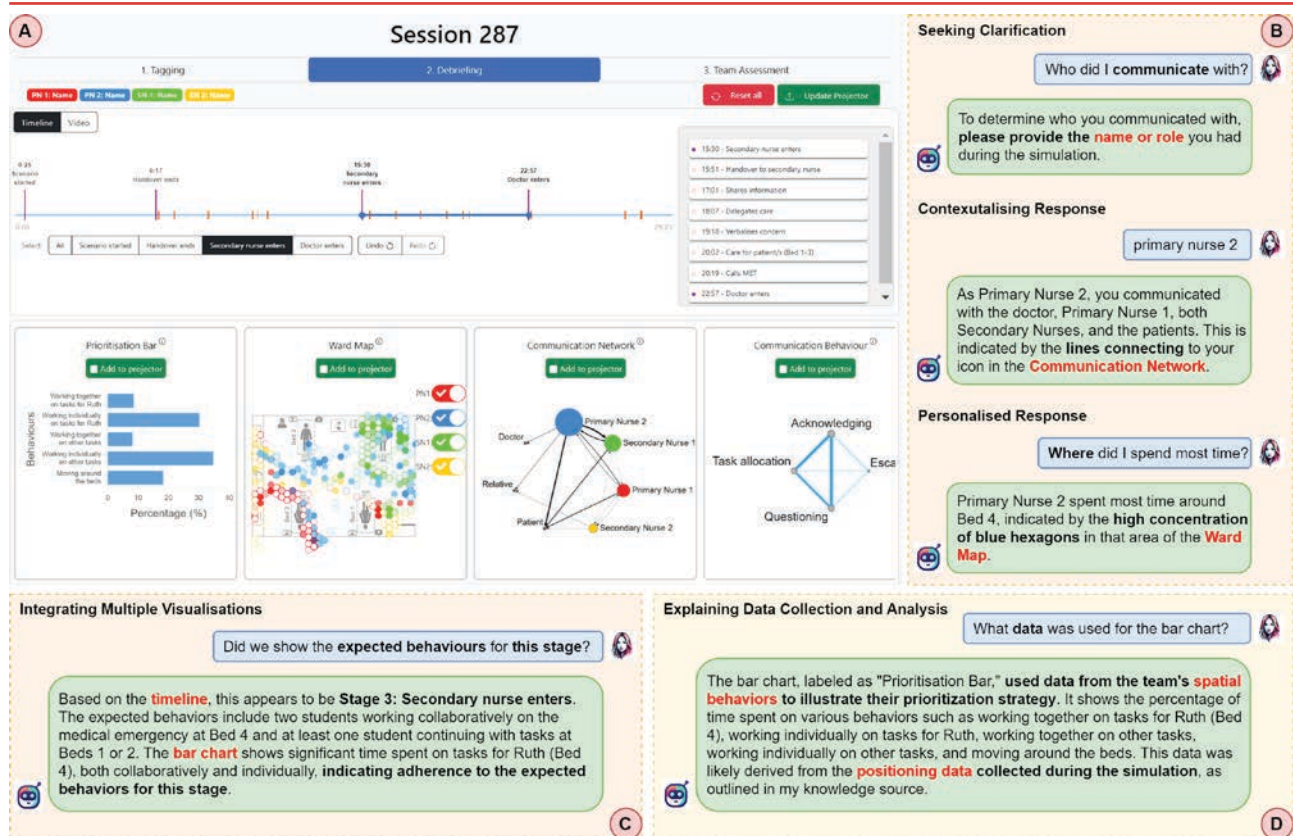
Source: Yan, L. et al. (2024[11]), "VizChat: Enhancing Learning Analytics Dashboards with Contextualised Explanations Using Multimodal Generative AI Chatbots", in *Lecture Notes in Computer Science, Artificial Intelligence in Education,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64299-9_13.

Figure 2.2. Data comics - using LLMs to generate visual feedback based on multimodal data about learning process
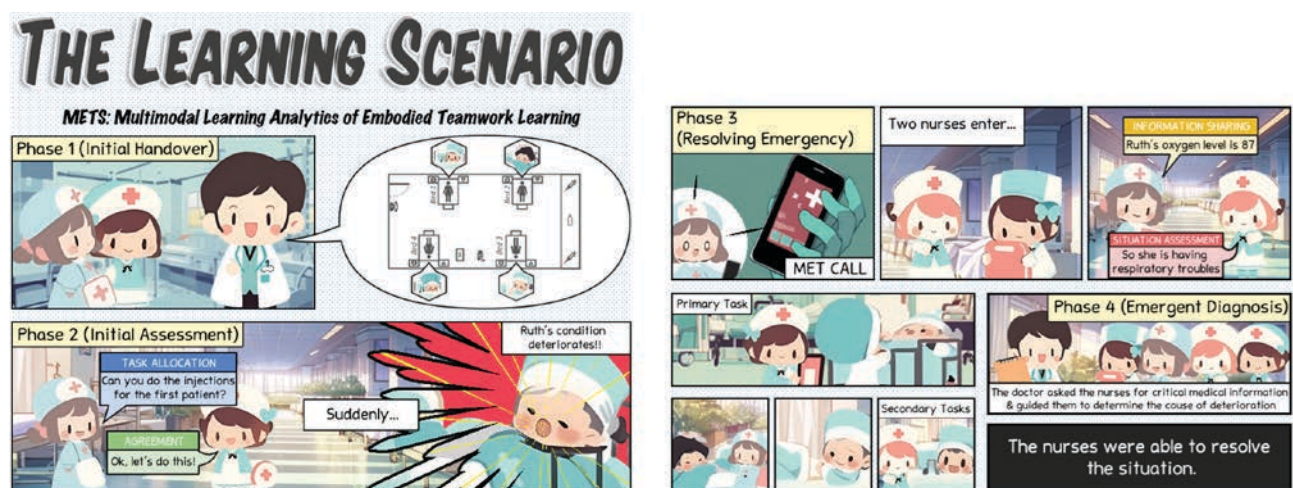
Source: Milesi, M. et al. (2024[87]), ""It's Really Enjoyable to See Me Solve the Problem like a Hero": GenAI-enhanced Data Comics as a Learning Analytics Tool", Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1-7, https://doi.org/10.1145/3613905.3651111

Process feedback

Feedback on learning processes (e.g. goal setting, strategy use, and self-monitoring) is underrepresented in existing literature on AI in education (Gašević, Greiff and Shaffer, 2022[89]). While learning is never fully transparent, learning analytics has made substantial progress in visualising and interpreting the otherwise invisible dynamics of the learning process, which can provide educators and learners with actionable insights into how learning unfolds. With advances in learning analytics, we can now analyse fine-grained trace data such as clickstreams, mouse movements, and other digital traces of student activity to identify cognitive, metacognitive, affective, and motivational processes (Molenaar et al., 2023[90]). Existing research in learning analytics has also shown that such approaches can offer insights into nuanced details about learning strategies learners used (Lämsä et al., 2025[91]). Moreover, existing research has shown that learning processes can explain more variance in student essay scores than linguistic essay properties (e.g. text cohesion) that are commonly used in automated essay scoring (Raković et al., 2022[92]). However, translating insights from the underlying representations of data analytic models – e.g. process maps, networks, or descriptive statistics – requires considerable data literacy, which can be a barrier for many educators and learners.

GenAI holds a strong potential to support feedback practices on learning process due to their ability to combine insights from data analytic models about learning processes with instructional information and subject matter content. By combining all these perspectives, GenAI can produce contextually relevant and personalised learning support – e.g. feedback or scaffolds – that aim to guide learners to improve their learning processes and performance (Thomann and Deutscher, 2025[93]; van der Graaf et al., 2023[94]). As outlined in Box 2.2, LLMs can be prompted with insights of real-time analytics of processes of self-regulated learning along with information about principles for effective feedback (Hattie and Timperley, 2007[55]) and relevant content information, to generate personalised scaffolds (Li et al., 2025[95]).

The potential of process feedback has profound implications in the age of GenAI. As students can now easily use GenAI tools to produce polished final products, it becomes increasingly important to assess how students *engage with the learning process*, rather than focusing solely on the end result. Moreover, process feedback can highlight important critical challenges - learners may face when using GenAI (e.g. metacognitive laziness and overreliance). The transformative potential of process assessment is further discussed below.

Box 2.2. Formative process assessment in the FLoRA platform: Using LLMs to transform real-time analytics into personalised feedback

The FLoRA platform is a suite of integrated tools that allow students to engage in a range of self-regulated learning activities (Li et al., 2025[95]): highlighting text, searching for information, taking notes, planning, using tools like ChatGPT, and more. All these interactions are logged and analysed using algorithms developed to map student actions onto cognitive, metacognitive, affective and motivational processes. These insights can then be used in real-time analytics to understand whether students use effective learning strategies, plan their learning, and regularly monitor how they progress with respect to task expectations. The insights from analytics can then be fed into LLMs to generate timely process feedback for learners. Such LLM-enabled feedback is more adaptive and context-sensitive than earlier rule-based systems (see the Instruction Panel in Figure 2.3). LLMs can interpret information about the learner, the task, and prior actions to generate personalised scaffolds in natural language, which learners can understand and act on. LLMs can also contextualise feedback to the topics and content learners are studying, make it easier to scale across different subject domains and tasks. Such process feedback can be delivered at predefined moments or triggered dynamically based on the observed learning processes. Li et al. (2025[96]) showed that students tend to follow these GenAI-produced scaffolds closely and thus improve their learning strategies, which were previously shown to be positively associated with improved performance in reading and writing tasks. Similarly, Alnashiri et al. (2026, under review[97]) showed that personalised feedback targeting particular processes of self-regulated learning (e.g. monitoring) is associated with improved writing performance and post-test knowledge in secondary education.

The success of systems like FLoRA demonstrates how GenAI can extend personalised learning support beyond individual classrooms and subjects. As GenAI tools become increasingly embedded in students' learning activities, concerns have emerged about overreliance and metacognitive laziness, where learners depend on AI-generated responses rather than actively monitoring and directing their own learning. Self-regulated learning is therefore central to maintaining students' lifelong adaptability and capacity to engage critically with AI-powered tools, yet teachers often struggle to observe and support these processes across diverse classrooms. Traditional feedback and scaffolding approaches are rarely scalable and tend to overlook subtle individual differences in learners' strategies and needs. By integrating real-time analytics with LLM-based feedback, GenAI systems can help bridge this gap, providing timely, context-aware support that complements teachers' work and sustains students' metacognitive engagement. Such scalable and individualized guidance is particularly important as education systems seek to foster independent and reflective learners in the age of AI.

Figure 2.3. Formative process assessment feedback on self-regulated learning in the FLoRA platform

Advancing assessment practices

Item generation and validation

GenAI holds the promise to generate assessment items. Although GenAI can produce a wide range of content, its use in standardised assessment requires generating items that meet psychometric standards of validity and reliability (Kaldaras, Akaeze and Reckase, 2024[98]) Emerging evidence suggests this is feasible (von Davier, 2026[12]). For example, Bhandari et al. (2024[99]) showed that ChatGPT can generate psychometrically sound items for Algebra, while Attali et al. (2022[100]) demonstrated similar success for reading tasks. The work by Attali et al. (2022[100]) underpins the automated item generation process used in the Duolingo English Test that is a widely recognized language proficiency exam. GenAI can be used to evaluate the quality of assessment items. Work at Duolingo emphasized the importance of the human- in- the-loop to perform item quality review and sensitive review as part of quality assurance and before checking for psychometric properties of the generated items (Hao et al., 2024[101]). This is also aligned with the recommendations by Moore et al. (2024[102]) combining human judgements with LLMs to produce high quality multiple choice questions and short answer questions.

Automated scoring

There is growing evidence of the potential of the use of GenAI in existing assessment practices. Existing research shows that the use of GenAI can be particularly effective when fine-tuned LLMs are used for automatic scoring of open-ended responses, demonstrating accuracy comparable or superior to models based on conventional machine and deep learning approaches. Latif and Zhai (2024[103]), for instance, showed that a fine-tuned version of GPT-3.5 significantly outperformed BERT in scoring multi-label and multi-class science education tasks, achieving up to a 10.6% accuracy improvement. Similarly, GPT-4 has shown strong alignment (Quadratic Weighted Kappa (QWK) over 0.8) with contemporary writing evaluation tools in high-stakes language assessment contexts for L2 English learners, especially when provided with a single calibration example for each rating category (Yancey et al., 2023[104]). However, Mansour et al. (2024[105]) showed that conventional approaches dramatically overperformed ChatGPT-3.5 Turbo and Llama2 (average QWK of 0.817 vs 0.313 and 0.201) on automatic essay scoring of English essays from the Automated Student Assessment Prize dataset, which contains essays written in English by U.S. secondary school students in grades 7–10 for whom English is the first language on persuasive, source-dependent, and narrative writing tasks. The results indicate that although LLMs can potentially be useful for some types of automated scoring tasks, they may not be for others. It is therefore important to extend the existing body of knowledge to understand the types of tasks LLMs can be effective for to inform educational practice and policy. Likewise, educators need to be careful in their choices of relying on GenAI for automatic scoring.

Several studies have examined the extent to which GenAI can automatically assess responses to open-ended questions in standardised assessments and identify effective prompting strategies. For example, Rodrigues et al. (2024[106]) evaluated GPT-4 across 738 open-ended questions drawn from high school Biology, Earth Science, and Physics tasks categorized by Bloom's taxonomy (BLOOM, 1984[20]). The model produced high-quality responses overall, though its performance declined on questions requiring factual recall or creative reasoning. Chan et al. (2025[107]) analysed LLMs in standardized STEM assessments and showed that chain-of-thought prompting significantly improved accuracy, particularly for reasoning-intensive problems. In higher education, Moore et al. (2022[108]) explored GPT-3's ability to evaluate student-generated short-answer chemistry questions in online college courses and found only modest alignment (32-40%) with expert judgments. Together, these studies show that while GenAI can complement human grading in structured educational contexts, its reliability still varies by domain, cognitive demand, and prompt design, highlighting the continued need for human oversight in both item generation and scoring.

Challenging assumptions and envisioning new horizons

The paradox of generative AI in skill acquisition

Despite the promise of generative AI, its integration into education raises fundamental questions that demand critical scrutiny. As the capacity of AI systems to automate cognitive tasks increases, it becomes imperative to interrogate not only what these technologies can accomplish but also what might be lost in the process. This section examines how prevailing assumptions about skill development and assessment are being disrupted and suggests a reorientation of educational priorities for the era of generative AI.

Distinguishing learning gains and performance in an AI-rich era

A central contention of this chapter is that educational systems must intentionally foster human capabilities even as they leverage GenAI's transformative potential. This imperative is not merely pedagogical; it is foundational to the cultivation of human skills that will enable individuals to thrive in rapidly evolving digital environments. According to Yan, Greiff, et al. (2025[17]), it is important to distinguish between two interrelated dimensions in human learning when using GenAI: AI-empowered performance and human skill development (Figure 2.4). The first dimension (vertical in Figure 2.4) is focused on development of human skills – i.e. human learning. This dimension has traditionally been covered in education including educational research in AI in education and the application of GenAI to support human learning are covered in the previous section on "Existing Practices". However, the ubiquitous presence of GenAI changes the context in which learning happens. This is why we also consider the second dimension (horizontal in Figure 2.4), which concerns the extent to which individuals use AI tools, such as large language models, to enhance task execution and produce high-quality outputs. In the remainder of this section, we consider the implications of these two dimensions on human skill development according to evidence emerging from the existing literature.

Figure 2.4. Balancing learning gains and performance with GenAI
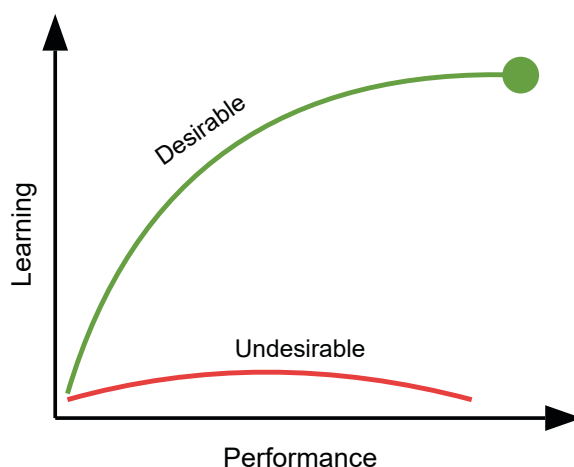
The mirage of "false mastery" in AI-augmented learning

The intersection of these dimensions defines the horizon toward which education should strive: learners who combine strong independent skills with the effective, reflective use of AI augmentation. However, a growing body of evidence suggests that this aspiration is not easily achieved and that the introduction of generative AI can create a "mirage of false mastery," where high-quality, AI-enabled output conceals underlying weaknesses in human skill – i.e. the undesirable curve in Figure 2.4 where task performance does not correlate with learning.

While generative AI has shown promise in supporting various educational tasks, its effectiveness in fostering long-term skill development remains uncertain. An important study in this space is conducted by Darvishi and colleagues (2024[13]) who investigated the extent to which an AI support tool could extend student ability to provide effective peer feedback. The GenAI tool was designed to support students in generating feedback more effectively, rather than directly improving the content of their responses. In a large-scale randomised controlled trial with approximately 1 600 students, Darvishi et al. (2024[13]) observed that while initial AI-supported gains in peer feedback quality were significant, these gains were not sustained once the tool was withdrawn. Students did not retain the feedback skills that appeared to have been acquired with AI support. Moreover, there was no robust evidence of synergistic development of human and AI-empowered skills; students generally exhibited strength in either AI-assisted performance or independent skill, but rarely both. These findings are echoed by a systematic review and meta-analysis by Vaccaro et al. (2024[109]) which analysed 106 experimental studies of human-AI collaboration. The meta-analysis found that, on

average, human-AI combinations performed worse than the best of either humans alone or AI alone, especially in decision-making tasks. This cautions against the assumption that human-AI synergy will naturally emerge. The risk in many educational contexts is that generative AI either simply augments current abilities or, more problematically in an education content, substitutes for human effort without fostering genuine skill development.

This pattern of substitution is often driven by external pressures. Research by Abbas et al. (2024[110]) revealed that university students were more likely to use ChatGPT when facing a high academic workload and time pressure. Their study, involving nearly 500 students, found that this utility came at a cost; increased use of ChatGPT was correlated with higher levels of procrastination, self-reported memory loss, and ultimately, diminished academic performance. Such findings suggest that students may turn to GenAI not as a partner for learning but as a tool to manage overwhelming demands, leading to unintended negative consequences. Furthermore, this substitution can foster an uncritical over-reliance on AI. In a systematic review, Zhai et al. (2024[111]) investigated how over-reliance on AI dialogue systems affects students' cognitive abilities. They define over-reliance as the uncritical acceptance of AI-generated recommendations, a tendency that arises when individuals struggle to assess the trustworthiness of the tool. Their findings indicate that this behaviour encourages the use of cognitive shortcuts, favouring fast, efficient answers over slow, effortful reasoning. This preference undermines the development of essential cognitive abilities, including decision-making, analytical reasoning, and critical thinking. It is crucial, therefore, to resist the temptation to conflate AI-augmented performance with authentic competence or deep learning.

Effects of GenAI use on learning processes

Another dimension of concern relates to the effects of the use of GenAI on learning processes. There is accumulating evidence that increased reliance on GenAI tools can suppress students' engagement in self-monitoring (defined as the ongoing process of checking, regulating, and adjusting one's understanding and strategies during learning), reflection, and evaluative judgement of one's learning processes, processes that are fundamental to autonomous learning (Molenaar, 2022[112]). When GenAI is used as a shortcut rather than as a scaffold that promotes learning, students may defer cognitive effort to technology, thereby weakening the very skills that underlie deep learning.

Empirical research has begun to quantify risks of GenAI on reduced human cognition and metacognition. In a study comparing the use of ChatGPT to traditional search engines for a scientific inquiry task, Stadler et al. (2024[15]) found that students using the large language model experienced a significantly lower cognitive load. However, this cognitive ease came at a cost: these students produced lower-quality reasoning and argumentation in their final recommendations compared to the group using the Google search engine. This highlights a critical trade-off, suggesting that while LLMs can reduce the cognitive burden of information gathering, they may not promote the deeper cognitive engagement necessary for high-quality learning. This finding is reinforced by a randomised experimental study by Fan et al. (2024[14]), which compared university students' writing processes when supported by ChatGPT, a human expert, a writing analytics tool, or no additional support. While the ChatGPT-supported group showed greater improvements in essay scores, these gains did not translate into deeper knowledge acquisition or transfer (as measured by knowledge transfer test on different topics). More importantly, the study found that learners in the AI-supported group demonstrated a marked reliance on the technology and were less likely to engage in metacognitive activities such as self-monitoring and reflection, a phenomenon the authors term *metacognitive laziness.*

The impact of AI on self-directed learning is further complicated by students' motivations for using these tools. A year-long longitudinal study by Xie et al. (2024[113]) examined how interaction frequency with chatbots affected learning autonomy. The results were nuanced: for learners seeking virtual companionship, the social presence fostered by the AI had a positive mediating effect on their learning autonomy. Conversely, for learners focused purely on knowledge acquisition, more frequent interaction with the chatbot was negatively correlated with both social presence and learning autonomy. This indicates that the effect of AI interaction is not uniform and that frequent use for instrumental purposes may undermine the development of independent learning habits.

These findings illustrate a crucial distinction: apparent improvements in performance enabled by generative AI may mask deficits in learners' underlying cognitive and metacognitive processes. However, this does not mean AI cannot play a productive role in learning. When structured intentionally within a collaborative learning environment, AI can act as a powerful scaffold. For instance, An et al. (2025[114]) studied student teachers using a mind-mapping tool integrated with GenAI. The groups using the AI tool not only outperformed the control groups on their collaborative tasks but also demonstrated a more sophisticated knowledge construction process, moving progressively from individual ideas to peer interaction and group synthesis.

Rethinking assessment

As students increasingly use GenAI tools for learning, traditional assessment models that focus solely on final outputs are becoming inadequate. When high-quality products can be produced with minimal engagement in the learning process, assessment risks measuring technological proficiency rather than human skill or understanding. To address this challenge, there is a pressing need to reorient assessment practices towards process-oriented approaches that evaluate not just what students produce, but how they engage with learning to create products. Assessments should aim to capture the processes students use to plan, monitor, and adapt their work, thereby revealing the authenticity and depth of their learning in GenAI-rich environments. Only by prioritising cognitive and metacognitive engagement alongside product quality can educational systems ensure that AI augments, rather than supplants, the development of meaningful human expertise.

One promising way to operationalise this shift is through evidence-centred assessment design (ECD) (Mislevy, 2006[115]). The ECD framework provides a principled model for linking assessment tasks, evidence, and inferences about learners' knowledge and skills. By moving beyond a narrow focus on final outputs, ECD enables the design of multidimensional assessments that capture both product and process evidence.

An illustrative example of this process-oriented approach comes from recent work in medical education, where clinical reasoning tasks have been redesigned to capture a more holistic view of learning (Tang et al., 2025[116]). Drawing on the ECD framework, this approach moves beyond assessing only the final diagnostic conclusion. Instead, it builds a multidimensional evidence model by collecting three streams of data as students interact with GenAI-powered virtual patients: product evidence (e.g. diagnostic accuracy), process evidence (e.g. conversation logs where students do history taking), and metacognitive evidence (e.g. clickstream data and interaction logs). Analysis of this rich data reveals that integrating all three evidence sources provides a significantly more reliable prediction of learner performance than relying on product-based measures alone. Notably, process data emerged as the strongest standalone predictor, underscoring the value of assessing the "how" of learning, not just the "what."

Figure 2.5. User interface of the FLoRA platform for formative process assessment of skills for history taking in medical education
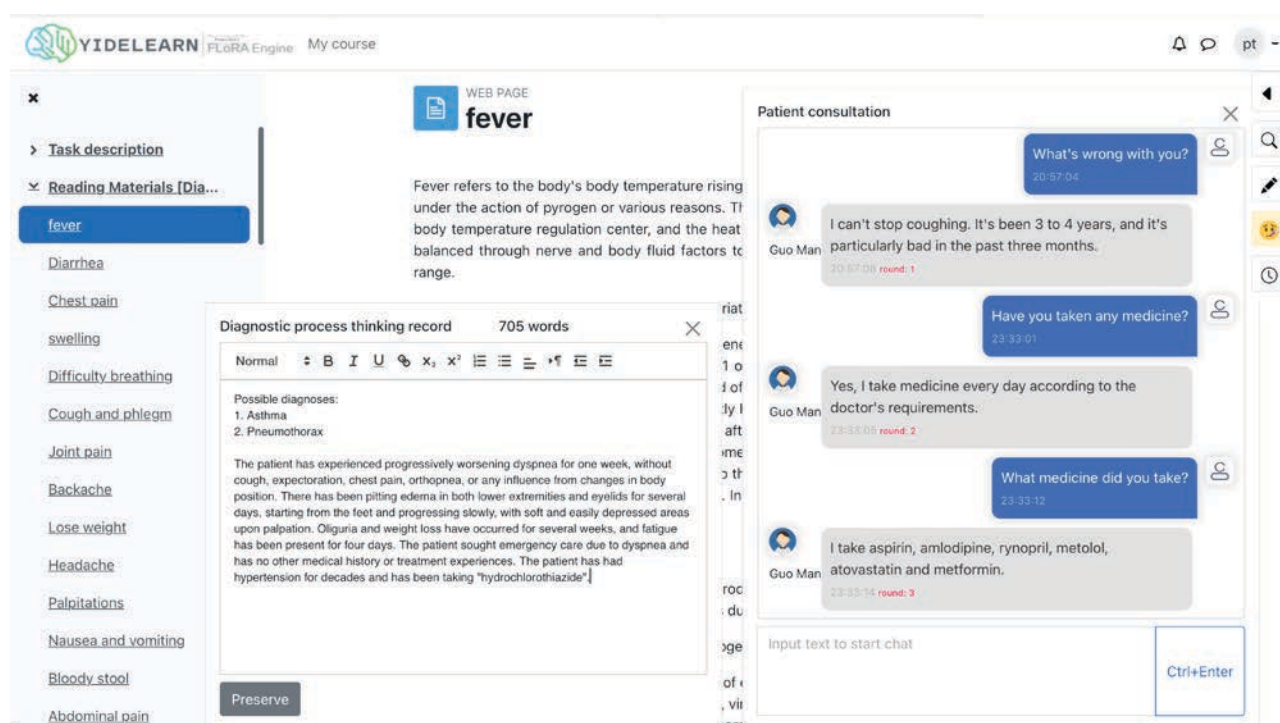
Note: (A) GenAI-based virtual standard patient; (B) the writing tool where the learners enter their diagnosis of the virtual standard patient. (C) assessment and feedback learners receive upon submitting their diagnosis of the virtual standard patient.

Building on this ECD foundation, predictive models are now being paired with Explainable AI (XAI) to make the assessment process not only accurate but pedagogically meaningful. Simply predicting performance with a "black-box" machine learning model is insufficient for supporting learning. To make insights actionable, the XAI layer identifies the key factors influencing a prediction (Khosravi et al., 2022[117]). These technical explanations are then translated by a GenAI system into structured, personalised, and pedagogically relevant feedback for the learner. This hybrid XAI–GenAI approach ensures that feedback is aligned with self-regulated learning principles, helping students understand not only their performance but also the cognitive and metacognitive strategies that shaped it. By grounding feedback in specific evidence from the learning process, this approach extends the ECD model beyond assessment design to feedback delivery, providing transparent, actionable guidance that fosters genuine skill development.

Figure 2.5 demonstrates how this assessment approach is implemented in the FLoRA platform for history-taking skills as part of the development of clinical reasoning in medical education (Tang et al., 2025[116]). Learners first interact with the virtual standard patients Figure 2.5 A, which are also based on GPT. Once the learner completes the interaction with the virtual standard patient and submits their diagnosis (Figure 2.5 B), the system applies the evidence model and generates personalised feedback (Figure 2.5 C).

Methodological rigour

The methodological rigour of research on generative AI in education is critical to produce quality evidence. If we are to make sound, evidence-informed decisions, it is essential to move beyond the commentaries and hype cycle and uphold high standards of empirical inquiry. As also indicated in the previous subsection and Figure 2.4, a central challenge in producing robust empirical evidence about effects of GenAI on human skills is the pervasive conflation of *performance* with *learning* (Yan et al., 2025[17]). Performance refers to the observed performance of a task, whereas learning involves an enduring change in knowledge and skills that is demonstrated through retention and transfer (Soderstrom and Bjork, 2015[118]). The distinction is essential; high performance, especially when mediated by a powerful tool, does not imply that learning has occurred.

A second, related but distinct issue is the *media/methods* fallacy (Clark, 1983[119]). For decades, researchers have cautioned against simplistic "media comparison studies" that attribute learning gains to a technology itself, rather than to the specific instructional methods it enables. Much of the nascent research on generative AI repeats this error, comparing an ill-defined "ChatGPT condition" with a control group and concluding that the technology "works". Such designs may demonstrate that a particular arrangement (e.g. students working with ChatGPT) can yield different outcomes than another (e.g. students working alone). However, because they attribute effects to the technology as a whole rather than to the specific instructional processes it affords, these studies provide limited insight into the underlying mechanisms. This limits their explanatory power and risks conflating performance support with genuine learning (Weidlich, 2025[16]).

A further methodological weakness, distinct from but often co-occurring with the media/methods fallacy, is the conflation of task performance with learning. For instance, meta-analyses claiming that ChatGPT enhances "academic performance" often measure immediate task achievement, not durable learning, and sometimes not even learning (Deng et al., 2024[120]). While students may produce a better essay or translation with AI assistance, this performance gain may mask a lack of underlying cognitive engagement and learning gains. As discussed earlier, offloading effort to AI can reduce cognitive load but also risks fostering "metacognitive laziness", thereby undermining the very processes required for deep skill development (Fan et al., 2024[14]; Stadler, Bannert and Sailer, 2024[15]). This problem is amplified by the "fast science" culture, where sensational claims, such as GPT-4 "acing" the MIT curriculum (Zhang et al., 2023[121]) gain traction despite significant methodological flaws, including data contamination and a lack of transparent verification (Chowdhuri and Koplow, 2024[122]). Even if such claims were accurate, they would have limited educational meaning, as GenAI outperforming standardised or benchmark tasks does not come with conceptual understanding and does not imply that the underlying processes of learning and transfer by humans should be abandoned. The danger lies in conflating technical proficiency with educational value, which can distort expectations and policy directions and fuel the kind of policy-practice misalignment that has characterised many AI-in-education debates (Lodge, 2025[123]).

To build a robust evidence base, the field must adopt a more rigorous research agenda. First, researchers must explicitly differentiate learning from performance by incorporating process-oriented assessments, such as delayed retention and knowledge transfer tests, into their designs (Yan et al., 2025[17]). Second, studies must move beyond media comparisons to isolate causal mechanisms, clearly defining the pedagogical function of the AI intervention, much like the decades of theory-driven research on Intelligent Tutoring Systems (Weidlich, 2025[16]). Finally, we must prioritise longitudinal research that tracks the durable effects of AI interaction on students' knowledge, skills, and dispositions over time.

For policymakers and funding organisations, this highlights a critical need to guide future investment. To build a robust evidence base, funding should prioritise longitudinal studies that track durable skills, demand that interventions clearly specify their pedagogical underpinnings and support the development of process-oriented assessments. Only by investing in research that distinguishes task performance from learning can we ensure that technology serves our ultimate goal: fostering deep and lasting human competence.

Conclusion

In conclusion, our findings underscore the significant promise of GenAI in enhancing educational practices related to learning and assessment. Specifically, we have demonstrated that GenAI-powered systems can directly support both students and educators, streamlining teaching activities and providing targeted assistance. However, despite these promising developments, our analysis also reveals several critical caveats that must be carefully considered when informing future practice and policy in this area. One such concern is the need for careful attention to the design of teaching practices that enable the effective use of GenAI, particularly in systems designed to directly support students, such as tutoring chatbots. For example, recent studies have shown that combining GenAI with established instructional methods (e.g. scaffolding), where GenAI agents guide students through step-by-step reasoning, can foster genuine learning and sustained performance enhancement even after removal of GenAI support (Yan et al., 2025[86]). By contrast, unguided "answer-giving" practices, where students simply request solutions from a chatbot, have been found to undermine reflection and suppress metacognitive engagement (Fan et al., 2024[14]; Stadler, Bannert and Sailer, 2024[15]). As emerging evidence suggests, not all students may benefit equally from these systems. Therefore, it is essential to consider how different student subpopulations, based on factors such as socio-economic status or prior academic achievement, interact with these technologies. By identifying the specific conditions under which these subgroups may benefit from GenAI-powered systems, we can mitigate potential inequalities and ensure that AI tools support a diverse range of learners.

Moreover, while GenAI offers the potential for rapid development of tools like tutoring chatbots, one should recognise that on average general-purpose *LLMs* (e.g. off-the-shelf GPT systems) do not yet match the effectiveness of traditional intelligent tutoring systems, when not designed or finetuned with adequate pedagogical knowledge (Borchers and Shou, 2025[46]). Hybrid systems that embed GenAI within educationally grounded frameworks may show more promise, but the evidence base remains limited. We still also need evidence that will compare how GenAI-powered tutors compare to their conventional counterparts in terms of their ability to provide sustained, long-term learning support. As a result, future research needs to investigate whether GenAI-based tutoring systems can effectively support learners over extended periods. In addition, there is considerable potential for GenAI to complement existing intelligent tutoring systems by enhancing their interactivity and enabling more natural language communication, which could ultimately create more personalised learning experiences. Future research should focus on exploring how GenAI can be integrated into intelligent tutoring systems, drawing on well-established educational principles to enhance these systems' overall efficacy.

Despite the promising capabilities of GenAI to generate high-quality feedback, research has shown that students' trust in AI-generated feedback varies considerably across contexts. In some studies, students respond positively and perceive such feedback as clear and useful, while in others they express scepticism about its accuracy or relevance. This variability in trust can influence whether learners engage with GenAI feedback, which in turn affects its potential impact on learning. To fully realise the benefits of GenAI in supporting feedback processes, future work should focus on developing teaching practices that help integrate AI-generated feedback effectively into classroom use. One promising direction is to use GenAI as a tool to help educators reflect on and refine their feedback (i.e. human or AI-generated) by checking whether it is clear, balanced, and aligned with established feedback principles before it reaches students.

GenAI shows promise in supporting educators with their daily teaching and administrative tasks. Although existing evidence grounded in more reliable measures of time spent with technology, such as usage log analysis, shows increases in efficiency for some tasks like lesson planning, qualitative studies highlight potential «blind spots» in these estimates that warrant further research. Specifically, the hidden labour educators must invest in reviewing and verifying the accuracy of AI-generated content may not be fully accounted for when relying solely on usage logs, unless they do the revisions online. All this emphasises the need for further research into how the GenAI tools can be designed to enhance, rather than complicate, teaching practices.

In the realm of assessment, GenAI offers valuable opportunities to streamline the creation of assessments and automate scoring processes. Its potential has been demonstrated in large-scale standardised tests, such as the Duolingo English Test, where GenAI can assists in generating items that meet psychometric standards. However,

this level of psychometric rigor is rarely required or feasible in everyday classroom assessments. While the widely available LLMs can help teachers design questions or tasks more efficiently, its outputs still require human review to ensure pedagogical relevance, truthfulness, fairness, and alignment with learning goals. Likewise, prompt-based approaches to automated scoring, although more accessible to non-technical users, remain less reliable than fine-tuned or conventional machine learning models. Educators should therefore treat GenAI tools as a complementary aid rather than a substitute for human judgment, validating its outputs for clarity and appropriateness before classroom use. Future research should focus on developing practical frameworks that help teachers integrate GenAI tools into formative and summative assessment processes responsibly, combining the efficiency of automation with the interpretive expertise of educators. To assist educators, future work needs to focus on developing classroom assessment strategies that incorporate GenAI in meaningful ways, expanding its applicability while also ensuring that it enhances the overall learning and teaching experience for both educators and students.

Box 2.3. Hybrid Human-AI Skills: Cultivating Competencies for a New Cognitive Ecosystem

As generative AI becomes an integral part of the cognitive ecosystem, the focus of skill development must shift from performing tasks that AI can automate to mastering the skills required to work with and through these technologies effectively. Hybrid Human-AI skills are not about replacing human intellect but augmenting it. They represent a suite of metacognitive, critical, and ethical competencies that enable individuals to leverage AI for enhanced performance without sacrificing the development of their own durable knowledge and judgment. Cultivating these skills is the primary defence against the "mirage of false mastery." Key components of the Hybrid Human-AI skillset include:

Strategic AI Engagement and Prompt Crafting:

This goes beyond simple "prompt engineering." It involves the ability to analyse a task, strategically select the appropriate AI tool (or choose not to use one), and design effective prompts that guide the model towards a desired outcome. It also includes the skill of iterative refinement, critically assessing AI output and adjusting prompts to improve results.

Critical Evaluation and Verification:

This is the capacity for intellectual scepticism when faced with AI-generated content. It involves actively fact-checking claims, recognising potential biases in the model's output, assessing the credibility of information (especially in Retrieval-Augmented Generation systems), and ultimately making an informed judgment about the trustworthiness and quality of the AI's contribution.

Creative Co-creation and Synthesis:

This higher-order skill involves using GenAI not as an answer engine, but as a creative partner. It is the ability to use AI for brainstorming, exploring counterfactuals, generating diverse perspectives on a problem, and synthesising AI-generated content with one's own knowledge to create novel work that is more than the sum of its parts.

Metacognitive Awareness and Self-Regulation:

This is the crucial ability to monitor one's own learning and cognitive processes while using AI. It includes recognising the onset of over-reliance or "metacognitive laziness," making conscious decisions about when to delegate a task to AI versus when to engage in effortful thinking for the sake of learning, and reflecting on how AI use is impacting one's own skill development.

Ethical and Responsible Use:

This involves understanding and navigating the ethical aspects of GenAI. It includes respecting data privacy, upholding academic and professional integrity, providing appropriate attribution for AI contributions, and considering the broader societal implications of how these powerful technologies are deployed.

Finally, our study highlights a critical risk: the uncritical adoption of GenAI may inadvertently undermine the development of key human skills such as critical thinking, metacognition, and evaluative judgment, all of which are foundational to genuine expertise. This could result in what we describe as the "mirage of false mastery," where the impressive outputs generated by AI mask the underdevelopment of essential skills, including hybrid human-AI skills (Box 2.3). The path forward, therefore, is not a rejection of technology, but a commitment to pedagogical intentionality and methodological rigor. Rather than simply asking whether GenAI "augments students' task performance," we must focus on how it can be used to foster deep, meaningful, and durable learning. This means reorienting our focus from GenAI-driven products to human-centred processes, ensuring that GenAI tools are designed to scaffold rather than supplant human thinking. By prioritising the development of durable, transferable skills and integrating metacognitive awareness into both learning and assessment, we can unlock the transformative potential of GenAI, creating an educational future that is not only more efficient but also authentically human.

Note

1. https://www.socialscienceregistry.org/trials/13519

References

Abbas, M., F. Jam and T. Khan (2024), "Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students", *International Journal of Educational Technology in Higher Education*, Vol. 21/1, https://doi.org/10.1186/s41239-024-00444-7. [110]

Abdelrahman, G., Q. Wang and B. Nunes (2023), "Knowledge Tracing: A Survey", *ACM Computing Surveys,* Vol. 55/11, pp. 1-37, https://doi.org/10.1145/3569576. [45]

Aldino, A., Y. Tsai, R. Mello, D. Gašević and G. Chen (2024), "Enhancing Feedback Quality at Scale: Leveraging Machine Learning for Learner-Centered Feedback", *Computers and Education: Artificial Intelligence*, Vol. 7, p. 100-332, https://doi.org/10.1016/j.caeai.2024.100332. [67]

Almoubayyed, H., R. Bastoni, S. Berman, S. Galasso, M. Jensen, L. Lester, A. Murphy, M. Swartz, K. Weldon, S. Fancsali, J. Gropen and S. Ritter (2023), "Rewriting Math Word Problems to Improve Learning Outcomes for Emerging Readers: A Randomized Field Trial in Carnegie Learning's MATHia", in *Communications in Computer and Information Science, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-36336-8_30. [29]

Alnashiri, H (2026, under review), *Analytics of Personalised Scaffolding Conditions for Self-Regulated Learning in Writing Tasks: Associations with Learning Outcomes in Secondary Students.* [97]

An, S., S. Zhang, T. Guo, S. Lu, W. Zhang and Z. Cai (2025), "Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment", *Computers & Education,* Vol. 227, p. 105-227, https://doi.org/10.1016/j.compedu.2024.105227. [114]

Attali, Y., A. Runge, G. LaFlair, K. Yancey, S. Goodwin, Y. Park and A. von Davier (2022), "The interactive reading task: Transformer-based automatic item generation", *Frontiers in Artificial Intelligence,* Vol. 5, https://doi.org/10.3389/frai.2022.903077. [100]

Azevedo, R., F. Bouchet, M. Duffy, J. Harley, M. Taub, G. Trevors, E. Cloude, D. Dever, M. Wiedbusch, F. Wortha and R. Cerezo (2022), "Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning With an Intelligent Tutoring System", *Frontiers in Psychology,* Vol. 13, https://doi.org/10.3389/fpsyg.2022.813632. [25]

Baker R.S., X. Liu, M. Shah, M. Pankiewicz, Y.J. Kim, Y. Lee and C. Porter (2026), "Generative AI as a teaching assistant", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [61]

Bhandari, S., Liu, Y., Kwak, Y. and Z. Pardos (2024), "Generative AI Can Harm Learning", *The Wharton School Research Paper,* Elsevier BV, https://doi.org/10.2139/ssrn.4895486.   [38]

Bhandari, S., Y. Liu, Y. Kwak and Z. Pardos (2024), "Evaluating the psychometric properties of ChatGPT-generated questions", *Computers and Education: Artificial Intelligence,* Vol. 7, p. 100-284, https://doi.org/10.1016/j.caeai.2024.100284.   [99]

Bloom, B. (1984), "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring", *Educational Researcher*, Vol. 13/6, pp. 4-16, https://doi.org/10.3102/0013189x013006004.   [20]

Borchers, C. and T. Shou (2025), "Can Large Language Models Match Tutoring System Adaptivity? A Benchmarking Study", in *Lecture Notes in Computer Science, Artificial Intelligence in Education,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-98417-4_29.   [46]

Cardona, M. and R. Ishmael (2023), *Artificial intelligence and the future of teaching and learning: Insights and recommendations,* https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf.   [5]

Chan, K., F. Ali, J. Park, K Buckingham Sham, E. Tan, F. Chong, K. Qian and G. Sze (2025), "Automatic item generation in various STEM subjects using large language model prompting", *Computers and Education: Artificial Intelligence,* Vol. 8, p. 100-344, https://doi.org/10.1016/j.caeai.2024.100344.   [107]

Chowdhuri, R. and N. Koplow (2024), *No, GPT4 can't ace MIT. Flower-Nutria-41d on Notion,* https://flower-nutria-41d.notion.site/No-GPT4-can-t-ace-MIT-b27e6796ab5a48368127a98216c76864 (accessed on 25 November 2025).   [122]

Clark, R. (1983), "Reconsidering Research on Learning from Media", *Review of Educational Research*, Vol. 53/4, pp. 445-459, https://doi.org/10.3102/00346543053004445.   [119]

Dai, W., Y. Tsai and D. Gašević (2025), "Designing relational feedback: A rapid review and qualitative synthesis", *Assessment & Evaluation in Higher Education,* Vol. 50(1), pp. 16-30.   [58]

Dai, W., Y. Tsai, J. Lin, A. Aldino, H. Jin, T. Li, D. Gašević and G. Che. (2024), "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study", *Computers and Education: Artificial Intelligence,* Vol. 7, p. 100-299, https://doi.org/10.1016/j.caeai.2024.100299.   [62]

Darvishi, A., H. Khosravi, S. Sadiq, D. Gašević and G. Siemens (2024), "Impact of AI assistance on student agency", *Computers & Education*, Vol. 210, https://doi.org/10.1016/j.compedu.2023.104967.   [13]

De Simone, M (2025), "From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria", World Bank Group, https://documents.worldbank.org/en/publication/documents-reports/documentdetail/en/099548105192529324.   [33]

Deng, R., M. Jiang, X. Yu and Y. Liu (2024), "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies", *Computers & Education,* Vol. 105224.   [120]

Dennison, D., B. Ahtisham, K. Chourasia, N. Arora, R. Singh, R. Kizilcec, A. Nambi, T. Ganu and T. Vashistha (2025), *Teacher-AI Collaboration for Curating and Customizing Lesson Plans in Low-Resource Schools.*   [50]

Donohoe, D. and E. Costello (2020), "Data Visualisation Literacy in Higher Education: An Exploratory Study of Understanding of a Learning Dashboard Tool", *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 15/17, p. 115, https://doi.org/10.3991/ijet.v15i17.15041.   [85]

du Boulay, B., A. Mitrovic and K. Yacef (eds.) (2023), *Handbook of Artificial Intelligence in Education,* Edward Elgar Publishing, https://doi.org/10.4337/9781800375413.   [21]

Dzikovska, M., and N. Steinhauser, E. Farrow, J. Moore and G. Campbell, G (2014), "BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics", *International Journal of Artificial Intelligence in Education*, Vol. 24/3, pp. 284-332, https://doi.org/10.1007/s40593-014-0017-9.   [28]

Er, E., G. Akçapınar, A. Bayazıt, O. Noroozi and S. Banihashem (2024), "Assessing student perceptions and use of instructor versus AI-generated feedback", *British Journal of Educational Technology,* Vol. 56/3, pp. 1074-1091, https://doi.org/10.1111/bjet.13558.   [76]

Escalante, J., A. Pack and A. Barrett (2023), "AI-generated feedback on writing: insights into efficacy and ENL student preference", *International Journal of Educational Technology in Higher Education*, Vol. 20/1, https://doi.org/10.1186/s41239-023-00425-2.   [75]

Escueta, M., A. Nickow, P. Oreopoulos and V. Quan (2020), "Upgrading Education with Technology: Insights from Experimental Research", *Journal of Economic Literature*, Vol. 58/4, pp. 897-996, https://doi.org/10.1257/jel.20191507. [36]

Fang, Y., Z. Ren, X. Hu and A. Graesser (2018), "A meta-analysis of the effectiveness of ALEKS on learning", *Educational Psychology*, Vol. 39/10, pp. 1278-1292, https://doi.org/10.1080/01443410.2018.1495829. [30]

Fan, Y., L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li and D. Gašević (2024), "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance", *British Journal of Educational Technology*, Vol. 56/2, pp. 489-530, https://doi.org/10.1111/bjet.13544. [14]

Gašević D., S. Dawson, T. Rogers and D. Gasevic (2016), "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success", *The Internet and Higher Education*, Vol. 28, pp. 68-84, https://doi.org/10.1016/j.iheduc.2015.10.002. [63]

Gašević, D., S. Greiff and D. Shaffer (2022), "Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond", *Computers in Human Behavior*, Vol. 134, https://doi.org/10.1016/j.chb.2022.107304. [89]

Giannakos, M., R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, S. Järvelä, M. Mavrikis and B. Rienties (2024), "The promise and challenges of generative AI in education", *Behaviour & Information Technology*, Vol. 44/11, pp. 2518-2544, https://doi.org/10.1080/0144929x.2024.2394886. [1]

Graesser, A., M. Conley and A. Olney (2012), "Intelligent tutoring systems.", in *APA educational psychology handbook*, Vol 3: Application to learning and teaching., American Psychological Association, Washington, https://doi.org/10.1037/13275-018. [22]

Hao, J., A. von Davier, V. Yaneva, S. Lottridge, M. von Davier and D. Harris (2024), "Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI", *Educational Measurement: Issues and Practice*, Vol. 43/2, pp. 16-29, https://doi.org/10.1111/emip.12602. [101]

Hattie, J. and H. Timperley (2007), "The Power of Feedback", *Review of Educational Research*, Vol. 77/1, pp. 81-112, https://doi.org/10.3102/003465430298487. [55]

Henderson, M. (2015), "Video-based feedback on student assessment: Scarily personal", *Australasian Journal of Educational Technology*, Vol. 31(1), pp. 51–66. [59]

Henderson, M., M. Phillips, T. Ryan, D. Boud, P. Dawson and E. Molloy (2019), "Conditions that enable effective feedback", *Higher Education Research & Development*, Vol. 38(7), pp. 1401–1416. [18]

Hutt, S., A. DePiro, J. Wang, S. Rhodes, R. Baker, G. Hieb, S. Sethuraman, J. Ocumpaugh and C. Mills (2024), *Feedback on feedback: Comparing classic natural language processing and generative AI to evaluate peer feedback*. [68]

Hwang, Y., Y. Kim, J. Koo, T. Kang and H. Jung. (2025), *LLMs can be easily Confused by Instructional Distractions*. [42]

Jin, Y., L. Yan, V. Echeverria, D. Gašević and R. Martinez-Maldonado (2025), "Generative AI in higher education: A global perspective of institutional adoption policies and guidelines", *Computers and Education: Artificial Intelligence*, Vol. 8, https://doi.org/10.1016/j.caeai.2024.100348. [4]

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. Chan, A. Madotto and P. Fung. (2023), "Survey of Hallucination in Natural Language Generation", *ACM Computing Surveys*, Vol. 55/12, pp. 1-38, https://doi.org/10.1145/3571730. [41]

Kaldaras, L., H. Akaeze and M. Reckase (2024), "Developing valid assessments in the era of generative artificial intelligence", *Frontiers in Education*, Vol. 9, https://doi.org/10.3389/feduc.2024.1399377. [98]

Kaliisa, R., Misiejuk, K. and S. López-Pernas (2025), "How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions", *Educational Psychology*. [78]

Kaliisa, R., K. Misiejuk, S. López-Pernas, M. Khalil and M. Saqr (2024), "Have Learning Analytics Dashboards Lived Up to the Hype? A Systematic Review of Impact on Students' Achievement, Motivation, Participation and Attitude", *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 295-304, https://doi.org/10.1145/3636555.3636884. [84]

Kestin, G., K. Miller, A Klales, T. Milbourne and G. Ponti (2025), "AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting", *Scientific Reports*, Vol. 15/1, https://doi.org/10.1038/s41598-025-97652-6. [32]

Khan, S. (2025), *Brave New Words, How AI Will Revolutionize Education (and Why That's a Good Thing)*, Penguin Random House LLC. [31]

Khosravi, H., S. Buckingham Shum, G. Chen, C. Conati, D. Gašević, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq and Y. Tsai (2022), "Explainable artificial intelligence in education", *Computers and Education: Artificial Intelligence,* Vol. 3.   [117]

Koedinger, K. and V. Aleven (2015), "An Interview Reflection on "Intelligent Tutoring Goes to School in the Big City"", *International Journal of Artificial Intelligence in Education,* Vol. 26/1, pp. 13-24, https://doi.org/10.1007/s40593-015-0082-8.   [26]

Kofinas, A., C. Tsay and D. Pike (2025), "The impact of generative <scp>AI</scp> on academic integrity of authentic assessments within a higher education context", *British Journal of Educational Technology,* Vol. 56/6, pp. 2522-2549, https://doi.org/10.1111/bjet.13585.   [2]

Kulik, J. and J. Fletcher (2016), "Effectiveness of Intelligent Tutoring Systems", *Review of Educational Research,* Vol. 86/1, pp. 42-78, https://doi.org/10.3102/0034654315581420.   [35]

Lämsä, J., S. de Mooij, O. Aksela, S. Athavale, I. Bistolfi, R. Azevedo, M. Bannert, D. Gašević, I. Molenaar and S. Järvelä (2025), "Measuring secondary education students' self-regulated learning processes with digital trace data", *Learning and Individual Differences,* Vol. 118.   [91]

Latif, E. and X. Zhai (2024), "Fine-tuning ChatGPT for automatic scoring", *Computers and Education: Artificial Intelligence,* Vol. 6, https://doi.org/10.1016/j.caeai.2024.100210.   [103]

Lehmann, M. and F. P.B. Cornelius and F.J. Sting (2025), *AI Meets the Classroom: When Do Large Language Models Harm Learning?.*   [37]

Liang, Z., L. Sha, Y. Tsai, D. Gašević and G. Chen (2024), "Towards the Automated Generation of Readily Applicable Personalised Feedback in Education", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64299-9_6.   [65]

Li, T., D. Nath, Y. Cheng, Y. Fan, X. Li, M. Raković, H. Khosravi, Z. Swiecki, Y. Tsai and D. Gašević (2025), "Turning Real-Time Analytics into Adaptive Scaffolds for Self-Regulated Learning Using Generative Artificial Intelligence", *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 667-679, https://doi.org/10.1145/3706468.3706559.   [95]

Liu, W., Z. Guo, Z., Xie, M., Xu, J., Huang, Z., Tian, M., . . . Zheng, X. a. (2025), *RECAST: Strengthening LLMs' Complex Instruction Following with Constraint-Verifiable Data.*   [43]

Li, X., Y. Fan, T. Li, M. Raković, S. Singh, J. Van der Graaf, L. Lim, J. Moore, I. Molenaar, M. Bannert and D. Gašević (2025), "FLoRA Engine", *Journal of Learning Analytics,* Vol. 12/1, pp. 391-413, https://doi.org/10.18608/jla.2025.8349.   [96]

Lodge, J. (2025), *Research, policy, and pedagogy: The three conversations on AI in education,* https://www.linkedin.com/pulse/research-policy-pedagogy-three-conversations-ai-education-lodge-p12ce/.   [123]

Lu, Q., Y. Yao, L. Xiao, M. Yuan, J. Wang and X. Zhu (2024), "Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing?", *Assessment and Evaluation in Higher Education,* Vol. 49/5, pp. 616-633, https://doi.org/10.1080/02602938.2024.2301722.   [69]

Mansour W.A., S. Albatarni, S. Eltanbouly and T. Elsayed (2024), Can Large Language Models Automatically Score Proficiency of Written Essays? in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* https://aclanthology.org/2024.lrec-main.247/   [105]

Markel, J., S. Opferman, J. Landay and C. Piech (2023), "GPTeach: Interactive TA Training with GPT-based Students", *Proceedings of the Tenth ACM Conference on Learning @ Scale,* pp. 226-236, https://doi.org/10.1145/3573051.3593393.   [47]

Martinez-Maldonado, R., V. Echeverria, G. Fernandez-Nieto, L. Yan, L. Zhao, R. Alfredo, X. Li, S. Dix, H. Jaggard, R. Wotherspoon, A. Osborne, S. Buckingham Shum and D. Gašević (2023), "Lessons Learnt from a Multimodal Learning Analytics Deployment In-the-Wild", ACM Transactions on Computer-Human Interaction, Vol. 31/1, pp. 1-41, https://doi.org/10.1145/3622784.   [88]

Matcha, W., N. Uzir, D. Gasevic and A. Pardo (2020), "A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective", *IEEE Transactions on Learning Technologies,* Vol. 13/2, pp. 226-245, https://doi.org/10.1109/tlt.2019.2916802.   [83]

Ma, W., O. Adesope, J. Nesbit and Q. Liu (2014), "Intelligent tutoring systems and learning outcomes: A meta-analysis.", *Journal of Educational Psychology*, Vol. 106/4, pp. 901-918, https://doi.org/10.1037/a0037123.   [34]

Miao, F. and M. Cukurova (2024), *AI competency framework for teachers,* United Nations Educational, Scientific and Cultural Organization UNESCO, https://doi.org/10.54675/ZJTE2084.   [6]

Miao, F. and Holmes, W. (2023), *Guidance for generative AI in education and research*. UNESCO Publishing, https://doi.org/10.54675/ewzm9535.   [7]

Milesi, M., R. Alfredo, V. Echeverria, L. Yan, L. Zhao, Y. Tsai and R. Martinez-Maldonado (2024), "It's Really Enjoyable to See Me Solve the Problem like a Hero": GenAI-enhanced Data Comics as a Learning Analytics Tool", *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems,* pp. 1-7, https://doi.org/10.1145/3613905.3651111. [87]

Mislevy, R. (2006), *Evidence-Centered Assessment Design,* Routledge. [115]

Mitrovic, A. and S. Ohlsson (2015), "Implementing CBM: SQL-Tutor After Fifteen Years", *International Journal of Artificial Intelligence in Education*, Vol. 26/1, pp. 150-159, https://doi.org/10.1007/s40593-015-0049-9. [24]

Molenaar, I. (2022), "The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning", *Computers and Education: Artificial Intelligence*, Vol. 3, p. 2369-2384, https://doi.org/10.1016/j.caeai.2022.100070. [112]

Molenaar, I. (2021), *"Personalisation of learning: Towards hybrid human-AI learning"*, OECD Publishing, Paris, https://doi.org/10.1787/589b283f-en. [23]

Molenaar, I., S. Mooij, B Azevedo, M Bannert, S. Järvelä and D. Gašević (2023), "Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data", *Computers in Human Behavior,* Vol. 139, https://doi.org/10.1016/j.chb.2022.107540. [90]

Moore, S., N. Bier and J. Stamper (2024), "Assessing Educational Quality: Comparative Analysis of Crowdsourced, Expert, and AI-Driven Rubric Applications", *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing,* Vol. 12, pp. 115-125, https://doi.org/10.1609/hcomp.v12i1.31606. [102]

Moore, S., H. Nguyen, N. Bier, T. Domadia and J. Stamper (2022), "Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3", in *Lecture Notes in Computer Science, Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption,* Springer International Publishing, Cham, https://doi.org/10.1007/978-3-031-16290-9_18. [108]

Moundridou, M., N. Matzakos and S. Doukakis (2024), "Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans", *Computers and Education: Artificial Intelligence,* Vol. 7, https://doi.org/10.1016/j.caeai.2024.100277. [49]

Nazaretsky, T., P. Mejia-Domenzain, V. Swamy, J. Frej and T. Käser (2024), "AI or Human? Evaluating Student Feedback Perceptions in Higher Education", in *Lecture Notes in Computer Science, Technology Enhanced Learning for Inclusive and Equitable Quality Education,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-72315-5_20. [77]

Nieminen, J. (2023), "Feedback literacy: A critical review of an emerging concept", *Higher Education,* Vol. 85(6), pp. 1381–1400. [60]

Nikolic, S., S. Daniel, R. Haque, M. Belkina, G. Hassan, S. Grundy, S. Lyden, P. Neal and C. Sandison (2023) "ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity", *European Journal of Engineering Education,* Vol. 48/4, pp. 559-614, https://doi.org/10.1080/03043797.2023.2213169. [3]

Nkambou, R., J. Bourdeau and R. Mizoguchi (2010), "Introduction: What Are Intelligent Tutoring Systems, and Why This Book?", in *Studies in Computational Intelligence, Advances in Intelligent Tutoring Systems,* Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-14363-2_1. [40]

Nye, B., A. Graesser and X. Hu (2014), "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring", *International Journal of Artificial Intelligence in Education*, Vol. 24/4, pp. 427-469, https://doi.org/10.1007/s40593-014-0029-5. [27]

OECD (2023), *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem,* OECD Publishing, Paris, https://doi.org/10.1787/c74f03de-en. [8]

Osakwe, I., G. Chen, A. Whitelock-Wainwright, D. Gašević, A. Pinheiro Cavalcanti and R. Ferreira Mello, (2022), "Towards automated content analysis of educational feedback: A multi-language study", *Computers and Education: Artificial Intelligence,* Vol. 3, https://doi.org/10.1016/j.caeai.2022.100059. [66]

Ouyang, F., M. Guo, N. Zhang, X. Bai and P. Jiao (2024), "Comparing the Effects of Instructor Manual Feedback and ChatGPT Intelligent Feedback on Collaborative Programming in China's Higher Education", *IEEE Transactions on Learning Technologies*, Vol. 17, pp. 2173-2185, https://doi.org/10.1109/tlt.2024.3486749. [72]

Pardo, A., J. Jovanovic, S. Dawson, D. Gašević and N. Mirriahi (2017), "Using learning analytics to scale the provision of personalised feedback", *British Journal of Educational Technology,* Vol. 50/1, pp. 128-138, https://doi.org/10.1111/bjet.12592. [54]

Paris, B. (2022), "Instructors' Perspectives of Challenges and Barriers to Providing Effective Feedback", *Teaching and Learning Inquiry*, Vol. 10, https://doi.org/10.20343/teachlearninqu.10.3. [53]

Parry, D., B. Davidson, C. Sewall, J. Fisher, H. Mieczkowski and D. Quintana (2021), "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use", *Nature Human Behaviour,* Vol. 5/11, pp. 1535-1547, https://doi.org/10.1038/s41562-021-01117-5. [52]

Pozdniakov, S., R. Martinez-Maldonado, Y. Tsai, V. Echeverria, N. Srivastava and D. Gasevic (2023), "How Do Teachers Use Dashboards Enhanced with Data Storytelling Elements According to their Data Visualisation Literacy Skills?", *LAK23: 13th International Learning Analytics and Knowledge Conference,* pp. 89-99, https://doi.org/10.1145/3576050.3576063. [70]

Qian, K., S. Liu, T. Li, M. Raković, X. Li, R. Guan, I. Molenaar, S. Nawaz, Z. Swiecki, L. Yan, L. and D. Gašević (2026), "Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support", *Computers & Education,* Vol. 240, https://doi.org/10.1016/j.compedu.2025.105448. [39]

Raković, M., S. Iqbal, T. Li, Y. Fan, S. Singh, S. Surendrannair, J. Kilgour, J. van der Graff, L. Lim, I. Molenaar, M. Bannert, J. Moore and D. Gašević (2022), "Harnessing the potential of trace data and linguistic analysis to predict learner performance in a multi-text writing task", *Journal of Computer Assisted Learning,* Vol. 39/3, pp. 703-718, https://doi.org/10.1111/jcal.12769. [92]

Rodrigues, L., F. Dwan Pereira, L. Cabral, D. Gašević, G. Ramalho and R. Ferreira Mello (2024), "Assessing the quality of automatic-generated short answers using GPT-4", *Computers and Education: Artificial Intelligence,* Vol. 7, https://doi.org/10.1016/j.caeai.2024.100248. [106]

Roest, L., H. Keuning and J. Jeuring (2024), "Next-Step Hint Generation for Introductory Programming Using Large Language Models", *Proceedings of the 26th Australasian Computing Education Conference*, pp. 144-153, https://doi.org/10.1145/3636243.3636259. [73]

Rudolph, E., H. Seer, C. Mothes and J. Albrecht (2024), "Automated feedback generation in an intelligent tutoring system for counselor education", *Annals of Computer Science and Information Systems, Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS),* Vol. 39, pp. 501-512, https://doi.org/10.15439/2024f1649. [80]

Ryan, T. and M. Henderson (2021), "Designing learner-centred text-based feedback: A rapid review and qualitative synthesis", *Assessment & Evaluation in Higher Education*, Vol. 46(6), pp. 894–912. [56]

Selwyn, N., M. Ljungqvist and A. Sonesson (2025), "When the prompting stops: exploring teachers' work around the educational frailties of generative AI tools", *Learning, Media and Technology,* Vol. 50/3, pp. 310-323, https://doi.org/10.1080/17439884.2025.2537959. [51]

Soderstrom, N. and R. Bjork (2015), "Learning Versus Performance", *Perspectives on Psychological Science*, Vol. 10/2, pp. 176-199, https://doi.org/10.1177/1745691615569000. [118]

Stadler, M., M. Bannert and M. Sailer (2024), "Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry", *Computers in Human Behavior,* Vol. 160, https://doi.org/10.1016/j.chb.2024.108386. [15]

Tang, L., K. Shen, N. Chen and D. Gašević (2025), *Advancing formative assessment with generative and explainable AI: A multidimensional evidence model for measuring and supporting learning..* [116]

Tang, L., K. Shen, H. Le, Y. Shen, S. Tan, Y. Zhao, T. Juelich, X. Li, D. Gašević and Y. Fan (2024), "Facilitating learners' self-assessment during formative writing tasks using writing analytics toolkit", *Journal of Computer Assisted Learning,* Vol. 40/6, pp. 2822-2839, https://doi.org/10.1111/jcal.13036. [81]

TEQSA (2024), *Gen AI strategies for Australian higher education: Emerging practice. Tertiary Education Quality and Standards Agency,* https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/gen-ai-strategies-australian-higher-education-emerging-practice (accessed on 27 November 2025). [124]

Thomann, H. and V. Deutscher (2025), "Scaffolding through prompts in digital learning: A systematic review and meta-analysis of effectiveness on learning achievement", *Educational Research Review,* Vol. 47, https://doi.org/10.1016/j.edurev.2025.100686. [93]

Vaccaro, M., A. Almaatouq and T. Malone (2024), "When combinations of humans and AI are useful: A systematic review and meta-analysis", *Nature Human Behaviour,* Vol. 8/12, pp. 2293-2303, https://doi.org/10.1038/s41562-024-02024-1. [109]

van der Graaf, J., M. Raković, Y. Fan, L. Lim, S. Singh, M. Bannert, D. Gašević and I. Molenaar (2023), "How to design and evaluate personalized scaffolds for self-regulated learning", *Metacognition and Learning*, Vol. 18/3, pp. 783-810, https://doi.org/10.1007/s11409-023-09361-y. [94]

von Davier, A. (2026), "Interview with Alina von Davier," in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [12]

Wang, Q. and A. Mousavi (2022), "Which log variables significantly predict academic achievement? A systematic review and meta-analysis", *British Journal of Educational Technology,* Vol. 54/1, pp. 142-191, https://doi.org/10.1111/bjet.13282. [64]

Wang, R., A. Ribeiro, C. Robinson and S. Loeb (2025), *Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise.* [48]

Weidlich, J. (2025), *ChatGPT in education: An effect in search of a cause, Center for Open Science,* https://doi.org/10.31234/osf.io/t6uzy_v2. [16]

Wisniewski, B. and K. Zierer (2020), "The power of feedback revisited: A meta-analysis of educational feedback research", *Frontiers in Psychology*, Vol. 10/487662. [57]

Xie, Z., X. Wu and Y. Xie (2024), "Can interaction with generative artificial intelligence enhance learning autonomy? A longitudinal study from comparative perspectives of virtual companionship and knowledge acquisition preferences", *Journal of Computer Assisted Learning*, Vol. 40/5, pp. 2369-2384, https://doi.org/10.1111/jcal.13032. [113]

Yancey, K., G. Laflair, A. Verardi and J. Burstein (2023), "Rating Short L2 Essays on the CEFR Scale with GPT-4", *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023),* https://doi.org/10.18653/v1/2023.bea-1.49. [104]

Yan, L., S. Greiff, J. Lodge and D. Gašević (2025), "Distinguishing performance gains from learning when using generative AI", *Nature Reviews Psychology,* Vol. 4/7, pp. 435-436, https://doi.org/10.1038/s44159-025-00467-5. [17]

Yan, L., S. Greiff, Z. Teuber and D. Gašević (2024), "Promises and challenges of generative artificial intelligence for human learning", *Nature Human Behaviour*, Vol. 8/10, pp. 1839-1850, https://doi.org/10.1038/s41562-024-02004-5. [19]

Yan, L., R. Martinez-Maldonado, Y. Jin, V. Echeverria, M. Milesi, J. Fan, L. Zhao, R. Alfredo, X. Li and D. Gašević (2025), "The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics", *Computers & Education,* Vol. 234, https://doi.org/10.1016/j.compedu.2025.105322. [86]

Yan, L., L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin and D. Gašević (2023), "Practical and ethical challenges of large language models in education: A systematic scoping review", *British Journal of Educational Technology,* Vol. 55/1, pp. 90-112, https://doi.org/10.1111/bjet.13370. [9]

Yan, L., L. Zhao, V. Echeverria, Y. Jin, R. Alfredo, X. Li, D. Gašević and R. Martinez-Maldonado (2024), "VizChat: Enhancing Learning Analytics Dashboards with Contextualised Explanations Using Multimodal Generative AI Chatbots", in *Lecture Notes in Computer Science, Artificial Intelligence in Education,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64299-9_13. [11]

Zhai, C., S. Wibowo and L. Li (2024), "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review", *Smart Learning Environments,* Vol. 11/1, https://doi.org/10.1186/s40561-024-00316-7. [111]

Zhang, S., S. Florin, A. Lee, E. Niknafs, A. Marginean, A. Wang, K. Tyser, Z. Chin, Y. Hicke, N. Singh, M. Udell, Y. Kim, T. Buonassisi, A. Solar-Lezama and I. Drori. (2023), *Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models.* [121]

Zhang, X., P. Zhang, Y. Shen, M. Liu, Q. Wang, D. Gašević and Y. Fan (2024), "A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education", *Frontiers of Digital Education*, Vol. 1/3, pp. 223-245, https://doi.org/10.1007/s44366-024-0028-5. [10]

Zhang, Z., Z. Dong, Y. Shi, T. Price, N. Matsuda and D. Xu (2024), *Students' Perceptions and Preferences of Generative Artificial Intelligence Feedback for Programming.* [74]

Zhan, Y. and Z. Yan (2025), "Students' engagement with ChatGPT feedback: implications for student feedback literacy in the context of generative artificial intelligence", *Assessment and Evaluation in Higher Education,* pp. 1-14, https://doi.org/10.1080/02602938.2025.2471821. [82]

Zhao, Y., L. Yan, W. Sun, G. Xing, S. Wang, C. Meng, Z. Cheng, Z. Ren and Dawei Yin (2024), *Improving the Robustness of Large Language Models via Consistency Alignment.* [44]

Zou, S., K. Guo, J. Wang and Y. Liu (2025), "Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: a comparison study", *Computer Assisted Language Learning,* pp. 1-30, https://doi.org/10.1080/09588221.2024.2447279. [71]

Part 1

Enhancing student learning with generative AI

3 Learning with dialogue-based AI tutors:
Implementing the Socratic method with Generative AI

Yuheng Li and Xiangen Hu
The Hong Kong Polytechnic University, China

This chapter presents the new affordances that generative artificial intelligence (GenAI) offers compared to previous AI-powered pedagogical agents. Taking the Socratic Playground system as an example, it highlights the different roles that GenAI-powered agents can take and emphasises how they can materialise a "pedagogy first" approach. After reviewing some of the evidence and proposing a framework for efficacy studies, it points to possible future directions for the development of educational GenAI systems. An annex presents some of the technical aspects of educational GenAI agents.

Introduction

Early AI-powered pedagogical agents in education were often limited to pre-scripted, rule-based tutoring systems, exemplified by platforms like *AutoTutor* that simulated tutor-learner dialogue through fixed scripts (Graesser et al., 2005[1]). These systems demonstrated that conversation-based learning with AI could mimic one-on-one human tutoring, yet they operated within tightly controlled dialogues and anticipated student responses. With the emergence of large language models (LLMs) such as OpenAI's GPT-4 in the 2020s, a new generation of generative AI (GenAI) agents has begun to transform this landscape (Alrobai and Cristea, 2025[2]; Zhang et al., 2024[3]). Unlike their scripted predecessors, GenAI agents can produce contextually relevant and linguistically coherent responses on the fly, allowing much deeper and more natural interactions. This chapter explores how LLM-powered pedagogical agents can transition from static virtual characters to adaptive, conversational partners by leveraging GenAI.

A central example of the chapter will be the Socratic Playground (SPL) system (Hu, Xu and Graesser, 2025[4]), which extends prior research on dialogue-based tutoring with generative AI and was developed by one of the co-authors. The SPL is a working prototype Intelligent Tutoring System (ITS) that integrates GPT-4 to deliver open-ended Socratic questioning and personalised feedback in real time (Hu, Xu and Graesser, 2025[4]). This introduction frames two key questions guiding our inquiry: *How do human interactions with GenAI agents differ psychologically and pedagogically from earlier scripted AI tutors, and what new affordances does GenAI agents bring to personalised learning?* From a psychological perspective, GenAI agents can emulate human-like conversational nuances and even emotional sensitivity, rather than the mechanical turn-taking of older systems (Córdova-Esparza, 2025[5]). Pedagogically, LLM-driven GenAI agents are capable of tailoring their prompts and explanations to each learner's inputs in ways that static decision-tree tutors could not, enabling a form of adaptive tutoring that was previously aspirational (Alrobai and Cristea, 2025[2]). For illustration, whereas a traditional AI tutor may offer a fixed hint regardless of students' affections reflected in their input, a GenAI agent may try to identify and predict these emotional states and tailor its responses

accordingly (e.g. when signs of frustration are detected, the GenAI agent can adopt a more encouraging tone to support the learner). These new affordances – ranging from possible on-demand personalisation at scale to the ability to engage in open-ended Socratic dialogue – hold promise for more effective and engaging learning experiences.

At the same time, a generative approach raises critical considerations that reaffirm the enduring centrality of pedagogy. Prior research emphasises that technology's power must be harnessed in service of sound instructional methods, not as a substitute for them (Koehler and Mishra, 2009). In other words, even the most advanced LLM-based tutor will fall short unless guided by established theories of learning and thoughtful educational design. Therefore, our investigation keeps this *pedagogy-first* perspective at the forefront. We position generative models as catalysts that can deepen an agent's persona and dialogue capabilities, provided that these models are integrated into robust teaching frameworks. The Socratic Playground prototype exemplifies how pedagogy can remain central – drawing on well-established tutoring approaches to foster critical thinking and learner reflection – while simultaneously integrating the generative capacities of cutting-edge AI within authentic educational contexts. By revisiting the evolution from rule-based systems like AutoTutor to today's generative Socratic tutors, this introduction sets the stage for a detailed examination of how these new agents function and how they can be used responsibly to improve learning. The chapter draws on recent research that compares legacy intelligent tutors to next-generation LLM-driven systems (e.g. Hu and Graesser, 2025[4]) and findings from the literature on AI-powered pedagogical agents in the age of generative AI. The ultimate goal is to articulate a vision for generative learning companions that not only move beyond the limitations of traditional AI tutors but also remain grounded in meaningful pedagogy and human empowerment.

Generative AI meets traditional, AI-powered pedagogical agents

Traditional ITS were largely rule-based expert systems that relied on predefined if-then rules and domain knowledge to emulate human tutors. These systems could provide step-by-step problem-solving support and feedback, but their behaviour was entirely scripted in advance (i.e. a pre-scripted avatar). Studies showed that such ITS, when carefully engineered, could approach the effectiveness of human one-on-one tutoring in certain domains (Graesser et al., 2005[1]; Paladines and Ramirez, 2020[6]). However, building these systems was labour-intensive: crafting the expert rules, questions, expected answers and feedback messages required extensive domain expertise. Each new subject domain meant starting a new rule base from scratch. Most importantly, the rigidity of rule-based tutors also meant they struggled with unanticipated student inputs or questions, constraining the practical scalability and the richness of tutoring interactions of such systems.

By contrast, recent advances in generative LLM like OpenAI's GPT series can generate fluent, contextually appropriate dialogue dynamically, bringing new opportunities to ITS implementation that excels in generalisability and adaptability compared to traditional pre-programmed tutors (Giannakos et al., 2024[7]; Hu, Xu and Graesser, 2025[4]). It is envisioned that generative AI agents can enable a more flexible tutoring experience, capable of addressing unforeseen questions or novel problem scenarios in real time (Kovari, 2025), something earlier rule-based systems often struggled to achieve.

Advances in dialogue and adaptability

This section further unpacks how the paradigm shift from rule-based intelligent tutoring systems to neural network-driven generative agents in education implicates educational practice and research.

Modern LLMs have achieved a level of conversational fluency and understanding that enables digital pedagogical agents to engage learners in open-ended discussions (Giannakos et al., 2024[7]; Yan et al., 2024[8]). GPT-4, for instance, has demonstrated the capacity to produce human-like explanations, ask clarification questions, and scaffold student thinking through multi-turn dialogue (Zhang et al., 2024[3]). Such models leverage vast pre-trained knowledge and contextual reasoning abilities that far surpass the pattern-matching techniques of earlier ITS (Frank et al., 2024). An LLM-based agent can "improvise" follow-up questions or hints based on a student's last response, rather than selecting from a fixed menu of replies. Research by Hu et al. (2025[4]) introduced the concept of a Socratic Playground as an exemplar next-generation ITS implementation which uses a GPT-4 core precisely to achieve this kind of dynamic adaptability. In pilot implementations, the generative approach led to significant improvements in the fluidity and personalisation of tutoring dialogues, compared to the more scripted interactions of legacy systems (Liu et al., 2024[9]; Zhang et al., 2024[3]). The agent demonstrated sound abilities to interpret nuanced or partially correct answers with high accuracy and generate new prompts or scenarios accordingly to address the learner's needs. These capabilities underscore how LLMs empower agents to navigate beyond the anticipated paths charted by designers, making the tutoring experience more responsive to individual learners.

In contrast, previous systems often faced issues when a student's input did not match any pre-programmed

expectation; the dialogue could stall or the agent might give a generic response. Generative models mitigate this by generalising from their training data to handle a variety of inputs, even those not foreseen by developers. They also bring knowledge grounding potential – via retrieval-augmented generation (RAG) or further fine-tuning on downstream, task-specific materials, so that an agent can incorporate up-to-date factual information into its tutoring (Modran et al., 2024[10]). Additionally, LLM-based agents can maintain a form of memory over the tutoring session (often retrain as the context of the session), tracking what concepts have been covered and what misconceptions the student has shown. This is done through mechanisms like conversation history or explicit memory modules that older systems lacked. For illustration, an agent can remember that a student struggled with a concept earlier and later revisit it with additional practice or questions, aligning with Vygotsky's notion of assisting within the learner's Zone of Proximal Development (Vygotsky, 1978[11]). By establishing and iteratively refining the learners' profiles, a GenAI agent can adapt the tutoring session accordingly (e.g. by adjusting difficulty), keeping the challenge level appropriate for the learner – a capability much closer to what a skilled human tutor would do than what earlier scripted tutors could manage.

In practice, prompt engineering techniques are oftentimes adopted to guide the LLM's behaviour toward an educational role. Developers design prompts that instruct the model to behave like a Socratic tutor, sometimes including structured guidelines or even standardised schemas for representing data (JSON schemas) to enforce pedagogical logic. Such prompt-based control, combined with the model's generative capacity, enables multimodal responsiveness as well – some agents can now produce not only textual explanations but also formulae, code, or even images on demand to aid understanding (Yan et al., 2024[8]). The ability to generate varied representations (e.g. analogies, examples, visualisations) helps address different modalities of learning, which was difficult for text-bound, pre-authored systems. In summary, GenAI has equipped digital pedagogical agents with a toolkit of affordances that include real-time dialogue generation, deep language understanding, context retention, and content creation. These affordances allow agents to personalise instruction and engage learners more flexibly than earlier systems that were constrained by pre-scripted logic. The remainder of this chapter will further examine what these changes mean for the roles agents can play, how their interactions can be designed, and how to ensure that this technological leap is anchored in effective pedagogy.

Enhanced agent roles and capabilities

Generative AI has expanded the pedagogical capabilities of artificial agents, allowing them to move beyond the conventional role of previous AI tutors dispensing knowledge and providing feedback. In the Socratic Playground (SPL) and similar systems, agents can fluidly adopt roles such as *mentor*, peer or emotional *coach* depending on the context and learner needs. This section explores these expanded roles and the new capabilities that enable them, illustrating how an LLM-driven agent can shift from being a mere content deliverer to a multifaceted educational partner.

Mentors and coaches

GenAI agents can act as mentors that guide learners through open-ended problems or projects. In SPL's essay-writing scenario, for instance, the agent does not simply provide facts or correct answers; instead, it mentors the student in critical thinking by asking probing why and how questions about the student's essay arguments. This approach, grounded in the Socratic method, aims to foster learner reflection and reasoning in a manner analogous to the guidance of a human mentor. Since the agent can dynamically generate follow-up questions based on the student's previous response, the dialogue feels tailored and intellectually challenging. Empirical observations suggest that SPL's Socratic agents effectively scaffold deeper reflection – students are prompted to explain their reasoning, consider counterpoints, and refine their ideas, rather than passively receiving information (Zhang et al., 2024[3]). This form of question-driven scaffolding, adaptive to individual learners' interactions, marks a shift in capability: the agent moves beyond the role of lecturer to that of a personalised cognitive coach, guiding learners in cultivating metacognitive strategies for learning.

Beyond academic guidance, GenAI agents are also capable of motivational coaching. Through sentiment analysis of learner input, an agent can detect frustration or confusion and respond with encouragement, praise for effort, or strategy suggestions. This appearance of emotional attunement, made possible by LLMs' language capabilities, allows the agent to assume the role of an affective *coach*, which may boost the learner's confidence and perseverance, as suggested in pertinent research (Córdova-Esparza, 2025[5]). In short, a well-designed GenAI agent can be simultaneously a cognitive mentor and an emotional coach, blending intellectual support with empathy in a manner that outperforms static, pre-programmed AI tutors.

Crucially, shifts between different mentoring or coaching roles could occur fluidly. The same AI tutor may transition

from giving hints, to asking reflective questions, to offering encouragement, to even letting the student lead the explanation, all in one session. This versatility was nearly impossible with rigid AI tutors, but LLMs make it feasible to improvise contextually. The ability to *track discourse* progression, coupled with the capacity to identify patterns and make predictions, when rigorously designed and implemented (e.g. via the additional instructions to examine students' expressed affections during conversations), allows the agent to infer when to switch roles – for instance, adopting a supportive stance during moments of frustration as an affective coach and transitioning to a more directive role once confidence is restored. By leveraging memory of preceding turns (oftentimes passed on to the GenAI agent as the context of conversations), the agent can follow pre-determined instructions to assess whether the learner is ready for increased autonomy or, alternatively, in need of scaffolded tasks and motivational feedback to bolster confidence. This capacity contributes to the humanisation of the interaction, aligning it more closely with authentic pedagogical exchanges observed in tutor–learner contexts.

Peers and companions

Generative pedagogical agents can also play the role of a peer-like collaborator, fostering collaborative engagement and co-construction of knowledge. GenAI agents can engage in less formal, more dialogic interactions that resemble peer learning or collaborative problem-solving. For example, an agent might take on the persona of a learning partner who works through a problem alongside the student, occasionally saying "Here's how I think about it, what do you think?" rather than always instructing. This peer role leverages the conversational nature of LLMs to create a two-way exchange where the learner feels more agency. Studies have shown that multi-agent or multi-role interactions can expose learners to diverse viewpoints and promote critical thinking (Park and Seo, 2025[12]; Wang et al., 2025[13]). In one notable demonstration, GenAI agents were used to simulate different participants (a student, a teacher, a parent) engaging in a classroom-style discussion, thus providing a learner with multiple perspectives in dialogue (Li, Xie and Lee, 2024[14]). While that example involved separate AI agents for each role, a single GenAI agent could also approximate a peer by sometimes prompting the student to teach it or by playing devil's advocate. In fact, SPL incorporates a feature akin to a teachable agent mode (inspired by learning-by-teaching paradigms): the agent prompts the learner to articulate a concept or to teach it back to the system. By briefly acting as the novice who needs an explanation, the agent encourages the learner to articulate and thereby solidify their understanding – a strategy supported by educational research on learning by teaching (Bargh and Schul, 1980[15]; Debbané et al., 2023[16]). Generative AI enables this role-play by producing plausible queries and misunderstandings for the student to address, simulating a peer who learns from another peer's explanations.

Researchers are also experimenting with entirely new archetypes of AI learning companions made possible by generative models (Han et al., 2025[17]). For illustration, one can make the GenAI tutor behave like:

- a "*reflection partner*" agent which prompts learners at the end of a lesson to reflect on what they learned, what they found difficult, and how they overcame challenges. By asking metacognitive questions and perhaps sharing its own "thoughts" (generated from pedagogical prompts), the agent may foster the learner's self-reflection and self-regulation habits (Guan et al., 2024[18]);

- a "*cross-domain companion*" that accompanies a learner across different subjects and contexts, helping to connect insights from one domain to another. As LLMs are trained on broad knowledge, a single agent can potentially discuss history in history class and switch to physics in science class, all while remembering the student's general learning profile. This could enable continuity in mentorship that spans multiple disciplines and learning periods, essentially acting as a personalised learning companion over the long term. While still theoretical, early work on long-running GenAI agents with continuous memory points toward the feasibility of AI companions that persist and evolve alongside the learner (Park et al., 2023[19]). Moreover, with ones' learning profiles forming a portion of their personal world models, such a cross-domain companion may be further extended into individuals lifelong learning companions (Krinkin, 2026 (forthcoming)[20]);

- a "*motivational interlocutor*" oriented toward sustaining learner engagement and motivation. In this role, the AI agent may periodically revisit the learner's goals, highlight progress achieved, or contextualise the material in relation to the learner's personal interests – a task that LLMs can attempt by drawing on broad knowledge across domains such as sports, music, or popular culture. Through such personalisation and the maintenance of a positive tone, the agent seeks to reinforce and sustain the learner's intrinsic motivation.

In all these expanded roles, the key enabler is the GenAI agent's capacity for real-time adaptation and rich interactive communication. Whereas traditional static AI tutors relied on scripted praise or generic feedback, a GenAI agent can adapt its motivational messages and adjust task difficulty in response to individual learner behaviours (e.g. offering more gentle encouragement to a student who has made several consecutive errors). This adaptivity yields a richer and more socially attuned educational experience, one that more closely approximates the nuances of human tutoring and peer collaboration. Students interacting with these agents are not just passively receiving information

but are actively engaged in a relational experience – conversing with a mentor/peer figure that responds to them, remembers prior exchanges, and adapts accordingly.

Early user studies and anecdotal evidence from the SPL demonstration suggest that learners frequently perceive the GenAI agent as "listening" or "understanding" them to a greater extent than prior e-learning tools. This suggests that the psychological presence of the agent is enhanced; it feels less like a programme and more like a conversational partner, which can increase student willingness to persist in learning tasks.

Of course, these new capabilities also bring new challenges. Ensuring the agent's responses remain pedagogically sound while it improvises as a peer or coach is an ongoing area of research. Nonetheless, the expanded roles and adaptivity afforded by generative AI clearly have the potential to make AI pedagogical agents far more than animated digital tutors – turning them into mentors, coaches, and collaborators that enrich the social, cognitive and metacognitive dimensions of learning.

Pedagogical design and interaction frameworks

Designing effective GenAI agents requires marrying AI capabilities with established pedagogical principles and interaction design frameworks. In this section, we outline key design principles for LLM-powered educational agents, including transparency of AI decisions, scaffolded questioning techniques, multimodal engagement, and maintenance of learner agency. We then examine how these principles are implemented in practice, referencing frameworks like *ARCHED* and the structured prompt templates (often JSON-based) used in the Socratic Playground system (see Box 3.1). By emphasising features such as conversational pacing, synchrony between verbal and non-verbal cues, and interactive learner controls, we showcase how GenAI agents can balance open-ended dialogue with instructional rigor, thereby fostering learner trust and autonomy even while the interaction is AI-driven.

Box 3.1. The ARCHED framework and JSON prompts

The ARCHED and other similar frameworks

ARCHED (AI for Responsible, Collaborative, Human-centred Education Instructional Design) is a framework to counter the "black box" opacity of automated educational tools by establishing a transparent, human-in-the-loop workflow grounded in Bloom's taxonomy. Rather than fully automating course creation, the framework employs a dual-agent architecture – comprising a Learning Objective Generation System for drafting content and an Object Analysis Engine for pedagogical auditing – to ensure that educators preserve the autonomy to curate and refine AI-generated objectives. Likewise, other notable frameworks in the field include the efficiency-focused GAIDE (Generative AI for Instructional Development and Education) model and the quality-focused Academic-GPT collaboration framework (adapting the design-build-test-learn approach).

JSON prompt

Text prompts are flexible, conversational, and human-friendly but they can leave room for ambiguity. In contract, JSON (JavaScript Object Notation) prompts use explicit fields and machine-readable syntax, which reduces guesswork for the GenAI agent and makes outputs more structured and reliable.

Source: Li et al. (2025[21]) for ARCHED; Dickey and Bejarano (2024[22]) for GAIDE; Chan et al. (2024[23]) for Academic-GPT.

Transparency and explainability

As AI tutors become more complex, transparency in their operation is crucial for building trust with both learners and educators. The *ARCHED framework* proposes a human-centred approach that embeds transparency and human oversight into AI-assisted instructional design (Li et al., 2025[21]) (see Box 3.1). Within the framework, multiple specialised AI components recommend pedagogical actions and evaluate them, while human educators remain the ultimate decision-makers, ensuring the reasoning behind AI-generated content is visible and can be vetted.

Translating this to a digital tutoring agent scenario, a GenAI agent should ideally be able to explain why a particular question is asked or why certain feedback is given – or at least do so if the learner inquires. For example, an agent might preface a hint by explaining that it is intended to support clarification of the learner's understanding of a specific concept. Such meta-dialogue provides insight into the agent's pedagogical intent. Another aspect of transparency is indicating uncertainty. If the AI is not fully confident in a response (which can be estimated from model probabilities or a validation step), it can disclose that uncertainty – e.g. "Let's double-check this answer, as I'm not entirely sure". This honesty can help set the right expectations and invite joint problem-solving, rather than the student taking every AI statement as gospel.

Several modern systems have introduced mechanisms to promote their transparency and explainability (See Box 3.2).

Box 3.2. The mechanisms adopted in modern systems

- Khanmigo (from the United States): Aiming to provide safe, institutional-grade tutoring, Khanmigo enhances transparency by providing inline citations to external sources and using a "double-check" mechanism where the AI visibly verifies its own answers before displaying them;

- Squirrel AI (from China): Intending to optimise learning efficiency through granular diagnosis; it utilises "Explainable AI" to show teachers the reasoning behind specific content recommendations, though the underlying "Large Adaptive Model" generally operates as an opaque system for the student;

- Riiid (from South Korea): Focused on efficiently maximising standardised test scores, Riiid employs "Explainable AI" techniques to visualise for users exactly which knowledge components or study behaviours contributed to their predicted scores;

- Century Tech (from the United Kingdom): Intended to augment teacher capabilities and personalise learning, their platform ensures transparency by explicitly distinguishing between teacher-assigned tasks and AI-recommended "nuggets" on the student dashboard, identifying the specific skill gaps from "micro-level" learning behaviours that triggered the recommendation.

Source: https://www.khanmigo.ai/ for Khanmigo; https://squirrelai.com/ for Suirrel AI; https://support.riiid.co/hc/en-us for Riiid; https://www.century.tech/ for Century Tech.

Alternative to the mechanisms used in existing tools, the implementation of generative pedagogical agents can incorporate mechanisms for post-hoc validation of interactions. In designing SPL, the developers introduced a logging and visualisation tool for researchers and instructors that showed the agent's decision path (e.g. which prompt pattern was triggered; what the agent "thought" the student's misconception was). While this backend transparency is not directly available to learners, it allows continuous human oversight of the AI's pedagogical actions. Overall, adopting a transparent design entails making both the system's internal reasoning and its external interactions as interpretable as possible, aligning with calls for trustworthy AI in education (Khosravi et al., 2022[24]; Memarian and Doleck, 2023[25]; Khosravi et al., 2022[24]; OECD-Education International, 2023[26]).

Scaffolded questioning techniques

A cornerstone of AI-powered pedagogical agent design is the use of scaffolded dialogue, often drawing from *Socratic questioning* and related strategies. Rather than delivering answers outright, a well-designed AI-powered pedagogical agent guides the learner to construct knowledge through carefully sequenced questions. This approach is rooted in Vygotskian scaffolding and the Zone of Proximal Development, where support is provided just beyond the learner's current ability and gradually withdrawn as competence grows (Vygotsky, 1978[11]). LLM-based agents are particularly well-suited to implementing Socratic questioning, as they can generate an extensive range of probing questions and follow-ups dynamically. They can also flexibly rephrase or adjust the difficulty of questions based on learner responses.

Frameworks for intelligent tutoring often include taxonomies of questions (e.g. conceptual probes, evidence requests, counterfactual prompts) that can be encoded into the AI's prompt or decision logic. In practice, SPL uses a *JSON-based* prompt template to enforce a structured tutoring script while still leveraging generative flexibility (Hu, Xu and Graesser, 2025[4]). The prompt is divided into sections, such as "Initial_Interaction", "Following_Up", "Providing_ Feedback", etc., each containing guidance for the type of Socratic moves the agent should make. For example, in a "Following_Up" turn, the agent might be instructed (via the prompt) to ask a "why" question related to the last student statement, or to request clarification if the student's answer was incomplete. By structuring the interaction in this way, the agent's generative outputs remain pedagogically purposeful. More importantly, the JSON structure also allows the system to *track expectations and misconceptions* explicitly, that is the agent keeps lists of the key points ("expectations") the student should mention in an ideal answer as well as known common errors ("misconceptions"). Each student response is compared (via the LLM or supplementary classifiers) against these lists, and the subsequent prompt is generated accordingly – e.g. if a misconception is detected, the following question might target that misunderstanding. This method, inspired by AutoTutor's expectation-misconception tailoring (Graesser et al., 2005[11]) but modernised with LLM capabilities, ensures the question scaffolding is adaptive to the learner's input. Empirical studies have long shown the effectiveness of such scaffolding as it keeps the learner in an active constructive mode rather than a passive one, which is known to enhance learning outcomes (Chi, 2009[27]).

Adopting the scaffolding approach in agent design aligns with a broader body of research that aims to leverage LLM-driven agents to foster deeper understanding and self-directed learning (Córdova-Esparza, 2025[5]). In designing a GenAI agent, educational technology developers should thus curate a bank of pedagogically sound questioning strategies and incorporate them either through prompt patterns, few-shot exemplars, or rule-based overlays on the LLM's output.

Multimodal and embodied interaction

To truly advance beyond single-modality interactions enabled by traditional AI tutors, GenAI agents can leverage multimodal engagement – combining text or speech with other modalities like visuals, gestures, or interactive simulations. Research in multimedia learning has shown that well-coordinated verbal and visual information can enhance understanding, as long as they are synchronous and not overwhelming (Mayer, 2002[28]). Modern AI platforms allow a tutoring agent to display images, diagrams, or even manipulate virtual objects in a simulated environment alongside the dialogue. For example, if a student is learning geometry, the agent might dynamically generate a diagram of a triangle and mark angles as it guides the student through a proof. Generative models can produce descriptions of visuals or request relevant images (via integration with image search or generation models), effectively acting as a bridge between text and visuals. Furthermore, if the agent is instantiated as a virtual tutor – whether through AR/VR or screen-based interfaces – the alignment of facial expressions and gestures with dialogue constitutes an important factor in achieving natural interaction. A nod or encouraging smile rendered on the AI tutor's avatar can reinforce the tone of the agent's message (e.g. affirming the learner's progress). Yet, it is worth emphasising that the timing of these cues should align with the conversational content to avoid cognitive dissonance.

The Socratic Playground's current implementation is primarily text-based with a simple animated avatar representing the AI pedagogical agent, but the design guidelines call for gesture-text synchrony in future versions – for instance, having the avatar produce a "thinking" expression when posing a difficult question, or a cheerful expression when giving positive feedback. The literature on embodied conversational agents (e.g. Krämer et al., 2007) suggests that such non-verbal behaviours, when congruent with the dialogue, can increase learner engagement and trust in the agent. Nonverbal behaviours are used in several innovative platforms: the *DALverse* project establishes an inclusive metaverse environment for distance education, where students can engage as digital avatars in multimodal learning tasks, leading to increased engagement and retention in distance learning settings (Damasceno et al., 2024[29]).

The design implications are clear: GenAI agents should, whenever possible, be integrated into interfaces that leverage multiple modalities (e.g. text, voice, graphics) to enable richer learning interactions. However, designers must adhere to established principles of multimedia learning to ensure that these modalities complement rather than compete with one another – for example, by avoiding extraneous animations or redundant narration that merely reads on-screen text aloud, both of which can contribute to cognitive overload.

Maintaining learner agency

A frequent criticism of AI tutors is the risk of learner passivity – if the agent does too much, students might become disengaged or over-reliant on the AI. Therefore, a central design principle is the preservation and promotion of learner agency. GenAI agents can support this in several ways. One approach is through developing learners' metacognitive awareness. This may be achieved by, for example, posing open-ended questions that allow learners to guide the direction of the interaction, thereby fostering their awareness to steer their own learning journey. Even simple prompts, such as "Would you like another hint or should we try a different problem?", position learners in an active decision-making role. Interfaces can further enhance agency through interactive controls. For instance, the SPL's interface offers learners with options to request a simpler explanation, pose a question to the agent, or indicate that they wish to attempt the solution independently. These controls act as safety valves so the student can modulate the help level. Under the hood, the agent monitors these inputs and adjusts its strategy – if a student repeatedly asks for simpler explanations, the agent will reduce the complexity of its language or break problems into smaller steps; if the student wants to proceed independently, the agent will step back and take on a more observational role, ready to jump in only if asked.

Another technique to maintain agency is via implementing *turn-taking policies* that ensure the AI does not dominate the dialogue. For instance, after the agent poses a question, it should give the learner ample time to think and respond, rather than immediately filling silence with more talk. If a student seems stuck, the agent may offer a hint, but ideally after encouraging the student to articulate any partial thinking first. This aligns with the AI tutoring technique of offering minimal help to keep the student doing as much cognitive work as possible: the goal of such systems is to reach an "interactive" level of engagement, where the student and tutor co-construct knowledge (Chi, 2009[27]). From a design perspective, such an implementation can be useful to measure the proportion of conversation led by the student vs. the agent; some prototype evaluations of SPL looked at what percentage of words or turns were student-generated and aimed to maximise that over time through interface tweaks. Furthermore, the agent can foster agency by being explicitly reflective: encouraging students to set goals, ask their own questions, or evaluate the agent's suggestions. For example, the agent might say, "Do you agree with the approach I just suggested, or do you think there's a better way?" – prompting the learner to critically assess the AI's previous responses, thereby treating the student as an active participant with agency, not just a recipient of knowledge.

Summary

The design of generative pedagogical agents must carefully blend AI innovation with human-centric educational principles. Frameworks like *ARCHED* provide macro-level guidance on maintaining transparency and human control in AI-powered education systems (Li et al., 2025[21]). The micro-level design in agentic systems such as SPL (Hu, Xu and Graesser, 2025[4]) illustrates concrete features that enact those principles (for example, structured prompts, adaptive questioning, interface controls). By prioritising explainability, scaffolded interaction, multimodal engagement, and learner agency, designers can create AI tutors that are not only powerful and adaptive, but also pedagogically sound and user-friendly. As subsequent sections will show, these design considerations play a critical role in addressing the challenges and ethical implications of generative AI tutors, ensuring that technology serves as a complement to effective teaching rather than a detour from it.

Working in practice: the SPL demonstration system

To ground the discussion in a concrete example, this section introduces the operational *Socratic Playground (SPL)* prototype and examines how it functions in real educational scenarios. The SPL system[1] serves as a demonstration of generative Socratic tutoring in action (Figure 3.1), focusing initially on the domain of essay writing and critical thinking. This section will describe a typical user experience with SPL, summarise preliminary evaluation data on its effectiveness, and discuss practical challenges encountered during deployment. The lessons learned from SPL's pilot use – including user feedback and technical issues like latency and hallucinations – highlight the gap that can exist between research promise and practical deployment, offering valuable insights for future improvements.

Figure 3.1. An overview of the SPL functionalities
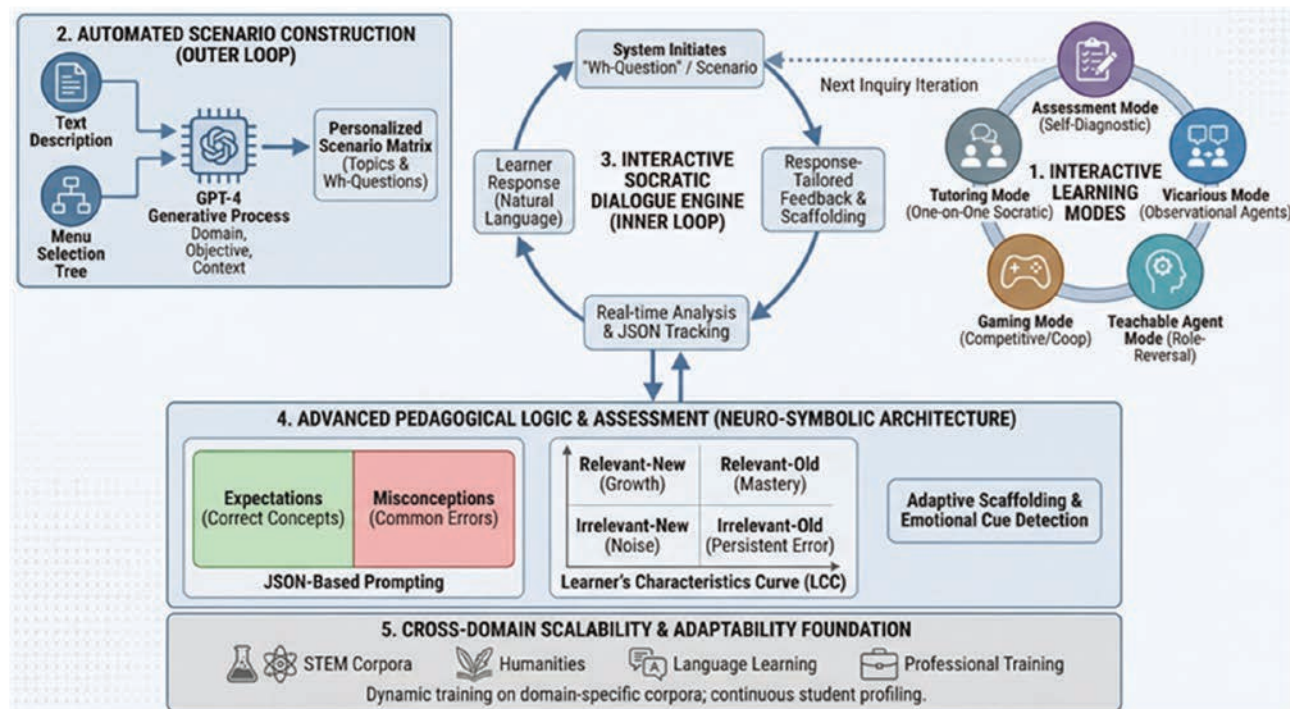
Source: Authors' own work.

Use case: Socratic tutoring for essay writing

In the piloting scenario with SPL, learners were asked to compose a short argumentative essay. For example, one prompt might be "Should renewable energy be subsidised by the government?" Instead of grading the essay outright or providing a static set of comments, the SPL agent engages the learner in a multi-turn Socratic dialogue about their essay (Figure 3.2). The session typically begins with the agent greeting the user and asking to see their draft or initial ideas. Suppose the student writes a few sentences stating their position. The agent will analyse this input (via GPT-4 and the underlying prompt structure) and then respond with a thoughtful question – often a "why" or "how" question – aimed at deepening the student's argument. For instance, if the student asserted "Yes, renewable energy should be subsidised because it's better for the environment", the agent might ask, "Why do you think government subsidies are necessary for environmental benefits, as opposed to letting the market handle it?" This kind of why-question scaffold pushes the student to elaborate on their reasoning. The student then responds, perhaps adding that "without subsidies, renewable projects might not attract investment". The agent continues this process, maybe following up with another prompt like, "Can you think of a specific example or evidence that supports that point?" Through such iterative questioning, the student is led to flesh out their argument with reasoning and evidence, essentially engaging in critical thinking about their own writing.

One notable observation from the SPL demonstration is that students often improve the quality of their reflections and explanations during these dialogues. Preliminary data collected from pilot users (university students in a writing workshop) suggest that after interacting with the Socratic agent, the students' final essays included more justification for claims and considered counterarguments more frequently than their initial drafts. While this is not a controlled study result, it aligns with the expectation that prompting learners to explain and justify would yield deeper engagement with the material (Nye, Graesser and Hu, 2014[30]). The GenAI agent essentially acts as a catalyst for self-explanation, a well-known mechanism for learning gains (Chi, 1994[31]). Users reported that the agent's questions made them think more critically: one participant noted "The AI asked me things I hadn't considered, like how exactly the subsidies work. It was challenging but it made my argument better." This anecdotal feedback resonates with our goals – the agent is not providing direct answers but improving the learner's thought process and output.

Figure 3.2. An example of Socratic tutoring session in SPL

Learning Scenario

You are a new policy analyst at a government agency tasked with reviewing proposals for renewable energy subsidies. Your first project involves analysing various reports that argue for and against government subsidies for renewable energy sources. To effectively convey your findings to stakeholders, you need to write an argumentative essay that clearly outlines the pros and cons of the topic, using appropriate text structures and organisational patterns. Your essay should be structured to facilitate understanding and persuasion, highlighting key points and supporting evidence.

AI Tutor: Hi there! Great to see you. Would you like to share your draft or any initial ideas about whether renewable energy should be subsidised by the government?

You: Yes, renewable energy should be subsidised because it's better for the environment

AI Tutor: You made a great point! Why do you think government subsidies are necessary for environmental benefits, as opposed to letting the market handle it?

You: Without subsidies, renewable projects might not attract investment.

AI Tutor: Excellent reasoning! You're highlighting the economic barrier — renewable energy often has high upfront costs, which can discourage private investors. Can you think of a specific example or evidence that supports that point?

Send My Answer

Source: Authors' own work.

Personalised scaffolding in action

The SPL system also demonstrates personalisation by adapting to different users' needs within the essay task. For a learner who is struggling to generate content (Figure 3.3), the agent takes on a more supportive, even slightly leading role. It might break down the task: "Let's start by outlining two main reasons you support subsidies. What's one reason?" If the student is totally stuck, the agent can even offer a gentle nudge like, "One reason might be related to climate change – do you want to expand on that or think of another reason?" On the other hand, for a confident learner who writes a strong initial paragraph, the agent switches to a more challenging role – perhaps by introducing a counterpoint: "Some critics argue subsidies distort the market. How would you respond to that counter-argument in your essay?" This not only personalises by difficulty but also by role: with the less confident student, the agent was a coach breaking down the task, whereas with the advanced student, it became a debate partner injecting opposing views. The underlying mechanism enabling this is the continuous

profiling of student performance and the flexible prompt template described earlier. Essentially, after each student response, the system classifies how well the student is doing and decides on a strategy (simplify vs. challenge, new question vs. give hint). This can be thought of as following the student's Zone of Proximal Development – always trying to operate just above the current level. By observing features like the length and substance of the student's answers, the system adjusts its scaffolding depth. Early user sessions show this adaptive behaviour could lead to *deeper engagement*: students at earlier stage of mastery benefit from tasks broken into manageable steps, while more advanced students remain engaged through progressively complex problems. Users in a professional development workshop who tried SPL noted that the agent felt "attentive", an impression due to this adaptive behaviour.

Figure 3.3. A conversation snapshot demonstrating the adaptability of SPL beyond simply questioning
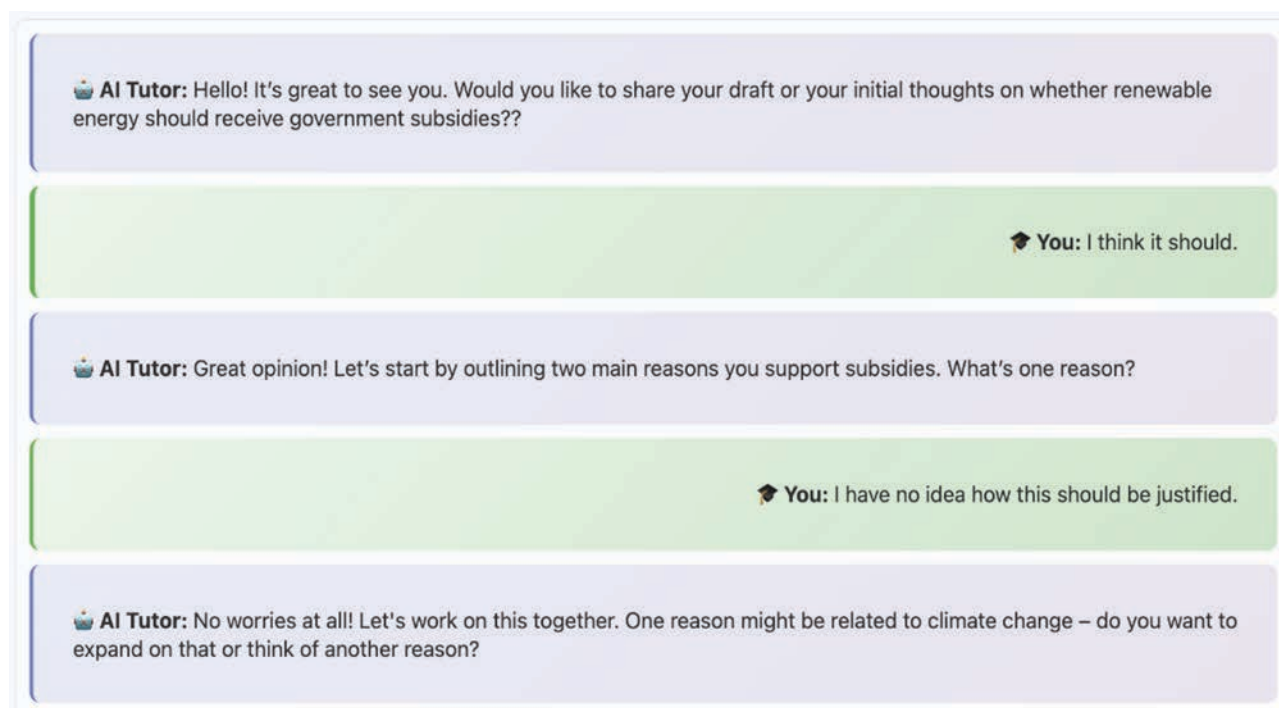
Source: Authors' own work.

Preliminary outcomes and feedback

Although comprehensive efficacy studies are still to be done (and this paper later suggests a framework for such studies), initial trials of SPL have been encouraging. In a pilot at the Hong Kong Polytechnic University involving about 20 adult learners (teachers and professionals exploring GenAI tools), participants reported high levels of satisfaction with the agent's usefulness. A post-session survey using a 5-point Likert scale found that the majority agreed (4 or 5) that the AI tutor's questions helped them think more critically about the topic. Many also indicated they would like to use such a tool regularly for brainstorming or refining their writing. From the system's perspective, logs showed that the dialogues often went through 8 to 12 turns of Q&A, with the learner contributing increasingly complex answers. Linguistic analysis of the student responses from their first turn to the final turn in each session indicated an increase in *lexical diversity and sentence complexity,* which can be seen as proxies for richer content (though more rigorous content analysis is needed). These observations align with the idea that Socratic generative tutoring fosters deeper reflection and engagement, resulting in improved outcomes like more thoughtful essays. Furthermore, previous findings suggest that a GPT-4 powered Socratic tutor (like SPL) facilitates more effective tutoring interactions than legacy ITS (Liu et al., 2024[9]), a result qualitatively and empirically reinforced by our hands-on trials.

However, the SPL demonstration also surfaced significant challenges, offering a reality check on the hype surrounding AI tutors. One practical issue was *GPT-4's latency,* as anticipated. In some cases, users had to wait 5 to 10 seconds for the AI tutor to respond, particularly as the prompt context increased (for instance, after the student had written a few hundred words of essay, sending all that plus the JSON structure to the model could prolong the response time). While many users were patient and understood this was a prototype, a few found the waiting time disruptive, especially when the conversation was flowing and then had a pause. This underlines a need for further optimisation or perhaps using a faster model for intermediate turns. Another issue observed was *occasional hallucinations* by the AI tutor. For example, in one session the student mentioned Germany's renewable energy policy, and the agent responded with a detailed "reminder" about an apparent German law that was not actually real – it had fabricated a plausible-sounding fact. The student was savvy enough to question it ("I'm not sure that law exists") and the agent then backtracked, but this could mislead less knowledgeable learners. We have since added additional checks when the agent provides factual statements, but it is a stark reminder that even a pedagogically well-intentioned AI can introduce misinformation. Ensuring factual accuracy remains an ongoing battle in generative AI tutoring, reinforcing arguments in the literature about the need for grounding and verification (Li et al., 2025[32]).

User interface and usability refinements

The SPL pilot also highlighted some interface issues. For instance, the initial version did not make it obvious that the user could ask the agent questions at any time; some users thought they could only answer the agent's questions. This one-sided interaction wasn't the intention – the system was capable of handling user-initiated questions or clarifications, but the UI cues were not clear. In response, we adjusted the interface to clarify that users can ask the AI tutor for explanations or request a hint. Another minor but interesting observation was that some users interacted with the AI tutor in a formal tone at the beginning of their conversation (for example, "Dear tutor, I have a question…"). Over time they became more conversational as they realised the agent responds like a human would. This acclimatisation suggests that building user trust and familiarity with the agent's style is part of adoption; any deployment should consider an onboarding or tutorial that lets users get comfortable with talking to the AI. From a design perspective, we found that *maintaining a user's sense of control* was vital. When a participant disagreed with the AI's suggestion, the AI persisted in pressing its point with the intention to be thorough, which left the user feeling frustrated. In subsequent tweaks, we have the agent explicitly acknowledge and respect the user's viewpoints more (for example, "That's a valid perspective. Shall we explore it further or do you want to consider alternative angles?"). This preserves the pedagogical goal of reflection but avoids the impression of the AI insisting on its way. Such fine-tuning makes the agent more like a supportive guide than an interrogator, which is important for sustained engagement.

Summary

The Socratic Playground demonstration provides a valuable case study in the real-world implementation of a generative pedagogical agent. It points to several potential benefits that are emerging in discussions of LLM-driven tutors (personalised scaffolding, improved critical thinking in student work, and positive learner reception) (Córdova-Esparza, 2025[5]; Liu et al., 2024[9]; Park and Seo, 2025[12]). At the same time, it unearths the pragmatic issues that arise when moving from controlled development to practical use: latency, occasional AI errors, interface clarity, and the delicate balance of control between student and agent. The experiences from SPL underscore a central theme of this chapter: there remains a gap between the research promise of GenAI in education and the practical deployment of these tools, which can only be closed through iterative refinement, user-centred design, and rigorous evaluation.

Framework for efficacy study

The integration of GenAI agents into education has moved beyond theoretical promise to rigorous empirical validation. In recent years, some randomised controlled trials and large-scale field studies have provided valuable insights on the current efficacy of ITS powered by GenAI. In this section, we synthesise emerging evidence across three distinct deployment models: hybrid/human-in-the-Loop (augmenting human tutors), independent tutoring (replacing or supplementing lectures), and classroom integration (supporting real-time classwork), and concludes with a streamlined framework (see Box 3.3) for evaluating the efficacy of such systems.

Emerging empirical evidence

A first scenario for the use of GenAI systems is that the AI does not teach the student directly but acts as a real-time "whisperer" for a human tutor, suggesting pedagogical moves to enhance instruction. The most prominent example is *Tutor CoPilot,* deployed in a large-scale randomised controlled trial involving 900 human tutors and 1 800 high school students (Wang et al., 2025[33]). The study found that while using the GenAI tutor improved student mastery

by 4 percentage points on average, its true power lay in "levelling up" the workforce. Students paired with lower-rated or novice tutors who used the CoPilot saw learning gains of 9 percentage points compared to the control group, effectively closing the gap between novice and expert tutoring. Analysis of chat logs revealed the mechanism: the GenAI system successfully nudged inexperienced tutors away from simply giving answers and toward using expert scaffolding strategies, such as asking guiding questions. This suggests that one of the most effective uses of GenAI is not to replace humans, but to scale expert pedagogy across a variable workforce.

Another scenario involves students interacting directly with an AI pedagogical agent to learn new concepts or accelerate their study, often outside standard classroom hours. At Harvard University, the *Harvard Physics Tutor* (a custom GPT-4 agent) was tested against a "gold standard" active learning classroom in a randomised crossover trial (Kestin et al., 2025[34]). The results were striking: students using the AI tutor achieved learning gains more than double those of the active learning group (effect size d≈0.73-1.3) and, crucially, spent significantly less time to reach that proficiency. This highlights the efficiency of "hyper-personalisation", where the GenAI addresses specific misconceptions that a classroom teacher cannot individually address for every student simultaneously.

Similarly, in the distance learning context, IU International University deployed *Syntea*, a GenAI teaching assistant, to over 10 000 students (Möller et al., 2024[35]). The primary metric of success here was "learning velocity": students using Syntea reduced the average time required to complete a course by 27% while maintaining exam performance. By acting as an always-available Socratic study partner, the GenAI agent removed the "wait time" for feedback, effectively accelerating the learning loop. In low resource settings, the text-based math tutor Rori demonstrated that high-fidelity interfaces are unnecessary for impact. Deployed via WhatsApp to over 1 000 students in Ghana, Rori produced significant math growth (effect size d=0.37) at a marginal cost of roughly $5 per student, proving that conversational AI can bridge the digital divide even on basic mobile infrastructure (Henkel et al., 2024[36]).

Thirdly, GenAI tutors can be used alongside standard instruction for practice problems. In that case, current evidence points to a high risk of cognitive offloading if "guardrails" are absent. (Cognitive offloading is the act of using external tools or resources to reduce the mental effort required to perform a task or remember information.) A study involving nearly 1 000 high school math students compared a standard "GPT Base" model against a pedagogically engineered GPT Tutor. Students given unrestricted access to the "Base" model performed 48% better during practice but 17% worse on subsequent independent exams, a phenomenon termed the "Crutch Effect". The students had learned to use the AI to bypass the cognitive struggle necessary for learning. The "GPT Tutor", intentionally engineered to withhold direct answers and prompt for self-explanation, mitigated this harm but did not yield the artificial performance boost seen in the base group (Bastani et al., 2025[37]).

Other classroom tools like *Khanmigo* (Khan Academy) have shown mixed quantitative results but strong qualitative benefits. While some trials of Khanmigo showed "no statistically significant difference" in short-term test scores compared to standard web search, students reported a significant reduction in "evaluation apprehension". They felt safer asking "stupid questions" to the AI than to a teacher (Slijepcevic and Yaylali, 2025[38]).

In summary, given this landscape of heterogenous outcomes – ranging from accelerated mastery to skill degradation – it is clear that efficacy is not inherent to the technology but dependent on implementation. More importantly, rigorous evaluation to the implemented tools should be warranted to distinguish authentic learning gains from deceptive performance boost. This calls for continuous and systematic evaluation of different uses of Gen-AI powered tools for learning. Box 3.3 provides some ideas of measures for evaluating these tools.

Challenges, ethics and practical implications

Deploying GenAI agents to facilitate technology-enhanced tutoring brings specific challenges that must be addressed to ensure these tools are responsible, equitable and educationally effective. While the potential for personalised, adaptive tutoring is vast, the implementation must navigate technical limitations and pedagogical risks.

Reliability and fairness in dialogue

A primary technical challenge in dialogue-based tutoring is the tendency of generative LLMs to "hallucinate", producing plausible but incorrect information (Zhang et al., 2025[40]). In a Socratic context, where the tutor leads the student through a chain of reasoning, a false premise introduced by the AI can derail the entire learning process. If students internalise these errors, the damage is significant. Studies have already observed students reproducing AI-introduced errors in homework tasks (Pardos and Bhandari, 2024[41]).

To mitigate this, systems increasingly employ Retrieval-Augmented Generation (RAG) to ground AI responses in trusted corpora, such as textbooks (Li et al., 2025[32]). Additionally, fairness remains a critical concern. LLMs can exhibit performance gaps across languages or dialects, potentially disadvantaging non-native speakers (Lai and Ngo, 2023[42]). Furthermore, without careful calibration, an AI tutor might inadvertently favour specific cultural perspectives or arguments, undermining the neutrality required for effective tutoring.

Box 3.3. A Framework for evaluating GenAI-powered intelligent tutoring systems (ITS)

Quantitative metrics (outcome and process)

- Standardised Learning Gains: Use pre-test/post-test designs with validated instruments such as the California Critical Thinking Skills Test (Facione, 1990[39]). Critical: Include delayed post-tests (e.g. 1 week later) without AI assistance to detect the "Crutch Effect" (performance illusion).

- Artefact Quality (Blind Grading): For writing / programming, use blind grading with rubrics (e.g. OECD's rubric for creativity and critical thinking) to assess if AI-coached drafts show superior critical thinking or creativity compared to control drafts.

- Efficiency: Measures "time-to-mastery" – whether AI allows students to reach the same proficiency standard sooner than traditional approaches.

- Interaction Granularity:

  1. Time-on-Task: Time spent engaging in the learning activities with AI (e.g. longer duration may suggest enjoyment with AI as the learning companion).

  2. Dialogue Symmetry: Ratio of student-to-AI words (e.g. low ratio may imply passive consumption while high ratio may indicate active engagement).

  3. Lexical Diversity and Syntactic Complexity: Use NLP to track if student responses become more complex over time (a proxy for deeper thinking).

Qualitative measures (phenomenology)

- Think-Aloud Protocols: Engage students to verbalise their thought process while using the AI. This identifies "mechanised convergence" (i.e. blindly accepting AI suggestions).

- Evaluation Apprehension: Survey students on their perceptions of using the ITS to support learning (e.g. comfort level asking questions to the AI vs. a human teacher).

- Teacher and Stakeholder Feedback: Interviews or focus groups to assess integration friction (e.g. whether the integrated tool disrupts class flow).

Experimental Designs

- Controlled Comparisons: Introduce Randomised Controlled Trial (or Quasi-Experimental design if RCT is not feasible) to enable comprehensive evaluation of system efficacy.

- The "Ablation" Study: For systems designed with a full suite of features, examine the systems' efficacy by introducing experimental conditions with and without specific features.

- Human-in-the-Loop RCT: Randomised access at the instructor level (not just student level) to scrutinise how the introduction of tools alters instructors' pedagogical practices.

- Longitudinal Tracking: Track usage over a period (e.g. a full semester) for the "novelty effect" wearing off; Track learning gains over a period to ensure they sustain.

Equity and safety checks

- Bias Audits: Stress-test the tools with "adversarial inputs" to check for cultural or political bias.

- Demographic Disaggregation: Analyse results by language backgrounds to ensure the tool aids learners of diverse competencies, not just the high-achievers.

Learner autonomy and cognitive dependency

Perhaps the most significant pedagogical challenge is maintaining the delicate balance between support and independence. There is a valid concern that over-reliance on AI assistance may reduce mental effort and compromise the depth of inquiry (Stadler, Bannert and Sailer, 2024[43]). If an AI tutor is too directive, or if the student passively accepts the AI's guidance, the metacognitive benefits of the method – self-evaluation and critical thinking – are lost.

Designers must thus ensure the AI empowers the learner rather than making them a passive recipient. This involves transparency in why a question is asked and explicitly prompting students to verify information, fostering the metacognitive development essential in the era of GenAI (Tankelevitch et al., 2024[44]).

The teacher-in-the-loop

Consistent with the consensus in the field, AI tutors should be viewed as tools to augment, not replace, human educators (Li et al., 2025[45]; Cukurova, 2026[46]). The ethical safeguard for these AI tutors should be a "human-in-the-loop" approach, where teachers retain oversight of the AI's guidance (Alfredo et al., 2024[47]; Colonna, 2023[48]; OECD-Education International, 2023[26]). Teachers must have the agency to determine when the AI is used – for example, assigning it for preliminary homework discussions so class time can be reserved for deeper analysis (Strauß and Rummel, 2026[49]). This requires distinct professional development to ensure teachers are literate in interpreting AI outputs and intervening when the system's logic drifts.

Implementation logistics and ethical safeguards

Beyond the specific pedagogical dynamics, the broader deployment of these agents requires strict adherence to operational and ethical standards. As Luckin and Holmes (Luckin, 2016[50]) argue, technological innovation must be paired with ethical guardrails. Key considerations include:

- **Data Privacy and Governance**: Systems must comply with regulations like European Union's GDPR or United States' FERPA. Since AI tutors collect deep behavioural data, strict anonymisation and access controls are required to protect student privacy (Colonna, 2023[48]).

- **Infrastructure and Equity**: Deploying LLMs like GPT-OSS and Qwen3 is computationally expensive. To prevent a digital divide where only wealthy institutions access high-quality GenAI-powered tutoring, strategies must include subsidised access or the use of optimised, lower-cost models.

- **Transparency and Trust**: It is an ethical imperative to transparently label the agent as an AI. Users should be informed of the system's limitations – specifically its potential to hallucinate – to encourage critical evaluation rather than blind trust.

Future directions and research roadmap

The advent of generative AI in pedagogical agents is only the beginning of a broader transformation in educational technology. This section outlines future directions and a research roadmap for advancing this field. It highlights several promising avenues:

- Authoring tools and platforms for educators to easily create and customise AI-driven tutoring content without deep technical knowledge;

- Multimodal GenAI agents that incorporate vision, speech, and possibly other sensory inputs to create more holistic learning experiences;

- Multi-agent and collaborative AI systems, where multiple AI tutors or AI-student peers interact with each other and learners to simulate group learning dynamics;

- Lifelong learning companions that accompany and support learners over extended periods (across courses or years), adapting as the learner grows;

- Cross-context adaptive deployment, ensuring these agents can transition and be effective in varied contexts (from formal classrooms to informal learning, across different subject domains and age groups). We also propose future research methodologies, including large-scale trials and longitudinal studies, to validate and refine the impact of these systems.

Finally, we emphasise that GenAI agents are an evolving technology that will require continuous evaluation of effectiveness, accessibility, and alignment with pedagogy as they develop.

Educator empowerment through authoring tools

First, for generative AI tutors to be most useful and widely adopted, educators need to be able to create and customise content easily. Relying on AI experts to build every lesson is not scalable. Therefore, a crucial area of development is *teacher-facing authoring tools* that leverage AI to help produce AI-driven lessons. For example, a teacher might input the learning objectives and key points for a lesson, and the system could generate a draft prompt template or a series of questions aligned with that objective. The teacher could then review, refine, and approve the AI-generated content. Alternatively, a teacher could demonstrate a desired dialogue flow once – either by conversing with a mock student or by outlining it explicitly, and the AI tutor could adopt that style. Additionally, AI could support the creation of simulations or narrative-based learning activities. For instance, if a teacher requests a scenario in which a student debates an AI acting a historical figure on a given topic, the system could generate an initial script for subsequent teacher revision. Such tools would dramatically lower the barrier to implementing customised AI tutoring across different subjects and languages.

A research path here includes understanding how teachers conceptualise AI behaviour and making interfaces that map to their thinking (e.g. some might prefer a rule-based interface, others might want to give examples and have the AI generalise – similar to programming by demonstration). Co-design with educators will be key; early studies should involve teachers using prototype authoring tools and measure outcomes like how quickly they can develop a new lesson, how effective that lesson is for students, and how comfortable the teachers feel about the level of control and transparency in the AI's resulting behaviour. The ARCHED framework is a step in this direction as it structures AI involvement in instructional design with human oversight at each stage (Li et al., 2025[21]). Future research can build on ARCHED to apply similar principles to real-time tutoring content creation.

Multimodal and embodied generative agents

One clear direction for future research is extending beyond text to multimodal interactions. Humans communicate and learn through a rich mix of modalities – speech, gesture, writing, drawing, etc. Future AI tutors likely will support student learning by doing the same. Already, models like GPT-4 have some multimodal capabilities (accepting image inputs, for example), and research is ongoing on integrating visual understanding with language models. A future AI-powered pedagogical agent might watch a student solve a physics problem on paper (via a camera), diagnose a misconception from their written work or diagrams, and then provide verbal guidance. Or in a virtual lab, the agent might observe how a student assembles a circuit or conducts a simulation and intervene at the right moment. Vision-enabled agents could check a student's worked solution for errors by "seeing" it, much as a teacher would glance at a notebook. Meanwhile, speech interfaces will allow more natural use in contexts where typing is inconvenient – imagine language learners practicing conversation with an AI that not only speaks but also reads facial expressions to gauge affection (e.g. confusion).

Moreover, embodied agents in AR/VR could provide immersive tutoring – for instance, a holographic science tutor that appears in an AR headset to guide a student through a chemistry experiment in a lab. Embodiment can leverage the physical environment: for instance, in mixed reality, an agent can point to parts of a model or demonstrate with virtual objects. Multimodal agents may be designed to enrich students' learning experience, better aligning with theories like Dale's Cone of Experience (Dale, 1969[51]) and Kolb's experiential learning cycle (Kolb, 1984[52]), which emphasise learning by doing and experiencing. Early trials of such approaches (like the *DALverse* platform mentioned earlier, integrating LLMs with a metaverse) have shown increased engagement and improved retention (Damasceno et al., 2024[29]). The challenge for researchers is to seamlessly integrate modalities such that the AI can interpret and generate multi-sensory data coherently. This may involve combining specialised models (for vision, for speech) with LLMs, or training unified, multi-modal models. It also raises new questions: How does one evaluate learning in these richer environments? How to ensure the added modalities truly improve learning and are not just gimmicks? These will be crucial questions to be answered as this line of research advances.

Nonetheless, a likely near-future scenario is a tutor that can speak and listen (already feasible with speech-to-text and text-to-speech integration) and perhaps use simple graphics or diagrams on the fly (e.g. drawing a chart using data provided in teacher-prepared curriculum materials). Ultimately, multi-modal generative agents aim to mimic a human tutor not just in conversation, but in full instructional presence – writing, sketching, demonstrating, and responding to non-verbal cues.

Multi-agent systems and collaborative learning

Another exciting frontier is the use of multiple agents to enrich educational interactions. Rather than one AI tutor and one student, we could have scenarios with several AI characters and one or more students. For example, a multi-agent system might include an AI tutor plus an AI peer learner; the human student can then participate in a group dialogue. This could simulate collaborative problem-solving or Socratic debates, exposing students to diverse viewpoints. (Park et al., 2023[19]) successfully developed "generative agents" that interact with each other to simulate human-like social behaviour. Researchers like Li et al. (2024[14]) have proposed multi-agent frameworks (EduMAS) for educational support. One agent might propose a solution, another critique it, and the human student could be asked to arbitrate or contribute, thereby learning through a rich discussion. Alternatively, multiple agents could take on specialised roles: one focusing on content hints, another on motivational encouragement, and a third maybe representing a historical figure or stakeholder in a debate (imagine learning civics by having AI agents role-play different political viewpoints in a discussion with the student).

These multi-agent interactions, if well-orchestrated, can model productive dialogue patterns and expose learners to argumentation and perspective-taking that is hard to achieve with a single tutor. However, designing multi-agent systems is inherently complex as the systems must be designed in ways that preserve consistency and prevent confusion (the agents must not overwhelm or contradict in a detrimental way). Research will need to explore optimal designs: What is the ideal number of agents within a multi-agent system? What role combinations are effective? Some early studies outside of education suggest multi-agent debates can improve answer accuracy by agents critiquing each other (Park and Seo, 2025[12]), but in education the goal might be more to model peer discussion or provide contrast.

Team Tutoring, where AI supports collaborative groups of students (Sottilare, Graesser and Sinatra, 2018[53]), is related; the AI could moderate or participate in a student group discussion, making sure everyone contributes (this uses multi-agent in the sense of AI + multiple humans). Overall, harnessing multiple AI agents to facilitate social learning is a promising direction, aligning with socio-constructivist theories that knowledge is often built through discourse.

Lifelong learning companions

Envisioning further ahead, generative AI agents could become continuous learning companions that support an individual over years, across many domains. Instead of separate tutors for math, science, writing, etc., one AI (or an integrated system) could accompany a student through their educational journey, maintaining a long-term model of their interests, strengths, and weaknesses. This idea resonates with the concept of personal AI assistants and also ties into educational initiatives for personalised lifelong learning (Krinkin, 2026 (forthcoming)[20]). For example, a student's "AI mentor" might help them in middle school algebra, then later adapt to help with high school physics, remembering that the student struggled with calculus concepts and proactively reinforcing those when they appear again. It could even extend beyond formal schooling: as the student goes to university or job training, the same AI knowing their learning history could tailor new learning experiences effectively.

Implementing this raises many questions – technical (how to store and update the long-term learner model securely), pedagogical (how to ensure continuity leads to cumulative benefits and not compounding of earlier biases), and ethical (ownership of that data, the right to "reset" or change one's AI companion to avoid pigeonholing based on early performance). Research by Tong and Hu (2024[54]) on self-improving adaptive instructional systems and by others on neuro-symbolic architectures (such as the NEOLAF AI service for education in the United States (Tong et al., 2023[55]) is trying to tackle the idea of AI that can improve itself and adapt over time, which is related to building a durable lifelong tutor. A roadmap for this could involve pilots where the same AI system is used across multiple courses or grade transitions, observing if retention or transfer improves because the AI can remind the student of previous knowledge or adapt to their cumulative profile.

A possibility is that a lifelong AI companion could also foster lifelong learning habits – by being present beyond the classroom, it might encourage curiosity, recommend learning opportunities, or help with personal projects (like an AI that helps a student interested in music by finding them educational resources or setting practice goals). Essentially, it would blur the lines between formal and informal learning support.

Cross-context and domain transfer

Another future direction is ensuring that generative agents can adapt to various educational contexts easily. At present, considerable effort is required to adapt systems to each new domain or use-case. In the long run, more generalist AI tutors might rapidly adapt to and learn new content. Few-shot learning capabilities of LLMs are promising here – perhaps a tutor can be given a single lesson text or some examples of questions and answers in a new subject

and then operate as a tutor in that subject. The robustness across contexts also includes adapting to different educational levels (e.g. using simpler language for a 5th grader versus a college student – something LLMs can do to an extent by role prompting). It also includes cultural context adaptation: a global AI tutor might need to know local curricula or examples relevant to the learner's environment. Research might explore how to embed contextual knowledge without a complete retraining – maybe by plugging in local knowledge bases or letting the system be easily fine-tuned by local educators.

Additionally, cross-context might involve the tutor being used in different settings: formal class vs. after- school vs. workplace training. We should explore how it needs to adjust its style (formal vs. casual, directive vs. self-directed learning mode) depending on context. The ultimate vision is an AI tutor framework that is as flexible as a human teacher who can teach different subjects, age groups, and adapt teaching style – a very remote goal, but research in transfer learning and multi-domain training of AI models is making progress.

Conclusion

The evolution of pre-scripted, AI-powered pedagogical agents towards sophisticated generative AI tutors represents a profound shift in educational technology. This chapter has traced this transformation through the lens of the Socratic Playground (SPL), demonstrating that the move from scripted tutor avatars to generative Socratic companions is not merely a technological upgrade, but a reimagining of educational possibilities. GenAI can function as a dynamic conversational partner capable of adaptive guidance and deep dialogue, provided that pedagogy remains the core driver. The "pedagogy-first" principle and integration in the system design is both an ethical imperative and the practical key to success. In essence, AI tutors must be as much products of educational craftsmanship as of computational prowess. *Digital innovations in education hinge on the synergy of advanced AI with sound teaching approaches* (Hu, Xu and Graesser, 2025[4]).

However, the rapid advancement of these systems necessitates a shift from proof-of-concept to rigorous, continuous evaluation. As GenAI agents become commonplace, the research community must move beyond novelty to conduct large-scale randomised trials that examine holistic outcomes. It is crucial to probe not only subject knowledge gains but also metacognitive shifts: do students develop better learning strategies, greater self-regulation, and sustained interest, or does the motivation fade once the novelty of the AI wears off? Furthermore, as models inevitably evolve (e.g. from GPT-4 to future iterations), the educational quality cannot be assumed to remain static. A robust process is needed to re-validate agents with each major upgrade to ensure they remain aligned with learning goals rather than becoming distractions.

This ongoing validation must prioritise inclusivity and ethics alongside effectiveness. Ensuring factual accuracy, fairness, and privacy are not optional add-ons but fundamental to the integrity of GenAI-powered tutors. To achieve this, various safeguard mechanisms in system implementation could be adopted – ranging from bias audits and human-in-the-loop frameworks to alignment with international guidelines like those from the OECD, UNESCO, and the EU. By implementing such safeguards, the international community can strive to make GenAI pedagogical agents not only effective but also trustworthy and inclusive. As features become more multimodal and autonomous, these efforts must specifically target accessibility, ensuring accommodations for diverse learning needs. When implemented carefully, these measures can foster trust. Students may welcome the personalised support and teachers appreciate the augmented capabilities provided, dispelling fears of AI as an unwanted intruder. Ultimately, navigating these challenges requires deep interdisciplinary collaboration among AI researchers, learning scientists, educators, policymakers, ethicists and other stakeholders, as no single group holds all the expertise needed to perfect these systems.

In Art Graesser's early work with AutoTutor (Graesser et al., 2005[1]), the dream was to simulate skilled tutoring dialogue. Today, GenAI and systems like the Socratic Playground bring us much closer to fulfilling those aspirations. Yet, this technology should not be pursued for novelty's sake, but to amplify and democratise the best of teaching – enabling rich, adaptive mentorship for every student, regardless of geography. Realising this vision requires a "pedagogy-first" ethos and the development and use of AI tools that augment rather than diminish human intellect. By combining the empathy of teachers, the rigour of learning and education scientists, and the computational consistency of AI, the education community can author a future where GenAI agents are harnessed responsibly to close the learning divide and become a success story for learners everywhere.

Note

1. The system is available publicly at https://uals.skoonline.org/

2. For example, a structured call like {"action": "USE_CALCULATOR", "input": "2*5*3.14"} means invoking a calculator if a multiplication task has been detected from the input "2*5*3.14".

References

Alfredo, R., V. Echeverria, Y. Jin, L. Yan, Z. Swiecki, D. Gašević and R. Martinez-Maldonado (2024), "Human-centred learning analytics and AI in education: A systematic literature review", *Computers and Education: Artificial Intelligence*, Vol. 6, https://doi.org/10.1016/j.caeai.2024.100215. [47]

Alrobai, A. and A. Cristea (2025), "Personality-Aware Conversational Intelligent Tutoring System with GenAI: Studying the Effect on Learners in Introductory Programming", in *Communications in Computer and Information Science, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-99261-2_31. [2]

Bargh, J. and Y. Schul (1980), "On the cognitive benefits of teaching", *Journal of Educational Psychology*, Vol. 72(5), p. 593. [15]

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakci and R. Mariman (2025), "Generative AI without guardrails can harm learning: Evidence from high school mathematics", *Proceedings of the National Academy of Sciences,* Vol. 122/26, https://doi.org/10.1073/pnas.2422633122. [37]

Bouamor, H. and J. Bali (eds.) (2023), *ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning.* [42]

Carvalho, P. (ed.) (2024), "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills", *PLOS ONE*, Vol. 19/5, https://doi.org/10.1371/journal.pone.0304013. [41]

Chan, K., S. Lokuge, M. Fahmideh and M. Lane (2024), AI-Assisted Educational Design: Academic-GPT Collaboration for Assessment Creation, Elsevier BV, https://doi.org/10.2139/ssrn.4996532. [23]

Chi, M. (2009), "Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities", *Topics in Cognitive Science,* Vol. 1/1, pp. 73-105, https://doi.org/10.1111/j.1756-8765.2008.01005.x. [27]

Chi, M. (1994), "Eliciting self-explanations improves understanding,", *Cognitive Science*, Vol. 18/3, pp. 439-477, https://doi.org/10.1016/0364-0213(94)90016-7. [31]

Colonna, L. (2023), "Teachers in the loop? An analysis of automatic assessment systems under Article 22 GDPR", *International Data Privacy Law*, Vol. 14/1, pp. 3-18, https://doi.org/10.1093/idpl/ipad024. [48]

Córdova-Esparza, D. (2025), "AI-Powered Educational Agents: Opportunities, Innovations, and Ethical Challenges", *Information*, pp. 16(6), 469. [5]

Cukurova, M. (2026), "A Conceptual Framework for Teacher-AI Teaming in Education: Harnessing GenAI to Enhance Teacher Agency", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [46]

Dale, E. (1969), *Audiovisual Methods in Teaching*, New York, Dryden Press. [51]

Damasceno, A., L. Silva, E. Barros and F. Oliveira (2024), "DA: Assistive Technology for Inclusion of People with Disabilities in Distance Education through a Metaverse-Based Environment", *IEEE International Conference on Advanced Learning Technologies* (ICALT). [29]

Debbané, A., K. Lee, J. Tse, and E. Law (2023), "Learning by Teaching: Key Challenges and Design Implications", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7/CSCW1, pp. 1-34, https://doi.org/10.1145/3579501. [16]

Dickey, E. and A. Bejarano (2024), "GAIDE: A Framework for Using Generative AI to Assist in Course Content Development", *2024 IEEE Frontiers in Education Conference (FIE),* pp. 1-9, https://doi.org/10.1109/fie61694.2024.10893132. [22]

Echeverria, V., K. Yang, L. Lawrence, N. Rummel and V. Aleven (2023), "Designing Hybrid Human–AI Orchestration Tools for Individual and Collaborative Activities: A Technology Probe Study", *IEEE Transactions on Learning Technologies*, Vol. 16/2, pp. 191-205, https://doi.org/10.1109/tlt.2023.3248155. [58]

Facione, P. (1990), *The California Critical Thinking Skills Test--College Level. Technical Report# 1, Experimental Validation and Content Validity.*   [39]

Giannakos, M., R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, S. Järvelä, M. Mavrikis and B. Rienties (2024), "The promise and challenges of generative AI in education", *Behaviour and Information Technology*, Vol. 44/11, pp. 2518-2544, https://doi.org/10.1080/0144929x.2024.2394886.   [7]

Gong, Y., J. Beck and N. Heffernan (2011), "How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis", *International Journal of Artificial Intelligence in Education*, Vol. 21/1-2, pp. 27-46, https://doi.org/10.3233/jai-2011-016.   [56]

Graesser, A., P. Chipman, B. Haynes and A. Olney (2005), "AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue", *IEEE Transactions on Education*, Vol. 48/4, pp. 612-618, https://doi.org/10.1109/te.2005.856149.   [1]

Guan, R., M. Raković, G. Chen and D. Gašević (2024), "How educational chatbots support self-regulated learning? A systematic review of the literature", *Education and Information Technologies*, Vol. 30/4, pp. 4493-4518, https://doi.org/10.1007/s10639-024-12881-y.   [18]

Han, Y., S. Hong, Z. Li and C. Lim (2025) (2025), "Defining and Classifying the Roles of Intelligent Learning Companion Systems: A Scoping Review of the Literature", *TechTrends*, Vol. 69/3, pp. 567-581, https://doi.org/10.1007/s11528-025-01058-0.   [17]

Henkel, O., H. Horne-Robinson, N. Kozhakhmetova and A. Lee (2024), "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana", in *Communications in Computer and Information Science, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64315-6_34.   [36]

Hu, X., S. Xu and R. Graesser (2025), *Generative AI in Education: From Foundational Insights to the Socratic Playground for Learning*, https://arxiv.org/abs/2501.06682.   [4]

Kestin, G., K. Miller, A. Klales, T. Milbourne and G. Ponti (2025), "AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting", *Scientific Reports*, Vol. 15/1, https://doi.org/10.1038/s41598-025-97652-6.   [34]

Khosravi, H., S. Buckingham Shum, G. Chen, C. Conati, D. Gašević, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq and Y. Tsai (2022), "Explainable Artificial Intelligence in education", *Computers and Education: Artificial Intelligence*, Vol. 3, https://doi.org/10.1016/j.caeai.2022.100074.   [24]

Krinkin, K. and T. Berlenko (2026), "Flipped" University: LLM-Assisted Lifelong Learning Environment. in Mahmud, M., M. Doborjeh, K. Wong, A.C.S. Leung, Z. Doborjeh and M. Tanveer, (eds), *Neural Information Processing. ICONIP 2024. Lecture Notes in Computer Science*, vol 15295. Springer, Singapore. https://doi.org/10.1007/978-981-96-6603-4_16   [20]

Kolb, D. (1984), *Experiential Learning: Experience as the Source of Learning and Development*, Prentice-Hall.   [52]

Li, H., Y. Fang, S. Zhang, S. Lee, Y. Wang and M. Botelho (2025), *ARCHED: A Human-Centered Framework for Transparent, Responsible, and Collaborative AI Assisted Instructional Design*, https://proceedings.mlr.press/v273/li25a.html.   [21]

Li, Q., Y. Xie and S. Lee (2024), E*duMAS: A Novel LLM-Powered Multi-Agent Framework for Educational Support.*   [14]

Liu, S., X. Guo, X. Hu and X. Zhao (2024), "Advancing Generative Intelligent Tutoring Systems with GPT-4: Design, Evaluation, and a Modular Framework for Future Learning Platforms", *Electronics*, Vol. 13/24, https://doi.org/10.3390/electronics13244876.   [9]

Li, Y., Z. Shan, M. Raković, Q. Guan, D. Gašević and G. Chen (2025), "When AI explains in natural language: Unveiling the impact of generative AI explanations on educators' grading and feedback practices", *Education and Information Technologies*, https://doi.org/10.1007/s10639-025-13741-z.   [45]

Li, Z., Z. Wang, W. Wang, K. Hung, H. Xie and F. Wang (2025), "Retrieval-augmented generation for educational application: A systematic survey", *Computers and Education: Artificial Intelligence*, Vol. 8, p. 100417, https://doi.org/10.1016/j.caeai.2025.100417.   [32]

Luckin, R. (2016), *Intelligence Unleashed: An Argument for AI in Education.*   [50]

Mayer, R. (2002), "Multimedia learning", in *Psychology of Learning and Motivation*, Elsevier, https://doi.org/10.1016/s0079-7421(02)80005-6.   [28]

Memarian, B. and T. Doleck (2023), "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review", *Computers and Education: Artificial Intelligence*, Vol. 5, https://doi.org/10.1016/j.caeai.2023.100152.   [25]

Modran, H., I. Bogdan, D. Ursuțiu, C. Samoila and P. Modran (2024), *LLM Intelligent Agent Tutoring in Higher Education Courses using a RAG Approach*, MDPI AG, https://doi.org/10.20944/preprints202407.0519.v1.   [10]

Möller, M., G. Nirmal, D. Fabietti, Q. Stierstorfer, M Zakhvatkin and H. Schütt (2024), *Revolutionising Distance Learning: A Comparative Study of Learning Progress with AI-Driven Tutoring*, https://arxiv.org/abs/2403.14642 (accessed on 4 December 2012).   [35]

Nye, B., A. Graesser and X. Hu (2014), "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring", *International Journal of Artificial Intelligence in Education*, Vol. 24/4, pp. 427-469, https://doi.org/10.1007/s40593-014-0029-5.   [30]

OECD-Education International (2023), *Opportunities, Guidelines and Guardrails for Effective and Equitable Use of AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/c74f03de-en.   [26]

Paladines, J. and J. Ramirez (2020), "A Systematic Literature Review of Intelligent Tutoring Systems With Dialogue in Natural Language", *IEEE Access*, Vol. 8, https://doi.org/10.1109/access.2020.3021383.   [6]

Park, B. and K. Seo (2025), "Assessing Critical Thinking through a Multi-Agent LLM-Based Debate Chatbot", *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1-13, https://doi.org/10.1145/3706599.3721207.   [12]

Park, J., J. O'Brien, C. Cai, M. Morris, P. Liang and M. Bernstein (2023), "Generative Agents: Interactive Simulacra of Human Behavior", *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1-22, https://doi.org/10.1145/3586183.3606763.   [19]

Slijepcevic, N. and A. Yaylali (2025), "Leveraging "Khanmigo" Generative AI-Powered Tool for Personalized Tutoring to Learn Scientific Concepts", J*ournal of Teaching and Learning*, Vol. 19/4, https://doi.org/10.22329/jtl.v19i4.10052.   [38]

Sottilare, R., A. Graesser and X. Sinatra (2018), *Introduction to Team Tutoring and GIFT. Design Recommendations for Intelligent Tutoring Systems.*   [53]

Stadler, M., M. Bannert and M. Sailer (2024), "Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry", *Computers in Human Behavior*, Vol. 160, https://doi.org/10.1016/j.chb.2024.108386.   [43]

Strauß, S. and N. Rummel (2026), "Fostering collaborative learning and promoting collaboration skills: what generative AI could contribute", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [49]

Tankelevitch, L., V. Kewenig, A. Simkute, A. Scott, A. Sarkar, A. Sellen and S. Rintel (2024), "The Metacognitive Demands and Opportunities of Generative AI", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1-24, https://doi.org/10.1145/3613904.3642902.   [44]

Tong, R. (2024), "Future of education with neuro-symbolic AI agents in self-improving adaptive instructional systems", *Frontiers of Digital Education,* Vol. 1(2), pp. 198–212.   [54]

Tong, R. J., C. Cao, T. Lee, G. Zhao, R. Wan, F Wang, X. Hu, R. Schmucker, J. Pan, J. Quevedo and Y. Lu (2023), *NEOLAF, an LLM-Powered Neural-Symbolic Cognitive Architecture.*   [55]

Vygotsky, L. (1978), *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press.   [11]

Wang, R. E., A. T. Ribeiro, C.D. Robinson, S.Loeb and D.Demszky (2025), *Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise.*   [33]

Wang, Z., J. Pan, D. Jin, J. Zhang, J. Cao, C. Zhang, Z. Li, P. Hansen, Y. Zhao, S. Shougian and X. Qiao "CharacterCritique: Supporting Children's Development of Critical Thinking through Multi-Agent Interaction in Story Reading", *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1-21, https://doi.org/10.1145/3706598.3713602.   [13]

Yan, L., S. Greiff, Z. Teuber and D. Gašević (2024), "Promises and challenges of generative artificial intelligence for human learning", *Nature Human Behaviour*, Vol. 8/10, pp. 1839-1850, https://doi.org/10.1038/s41562-024-02004-5.   [8]

Zhang, L., J. Lin, Z. Kuang, S. Xu and M. Hu (2024), *SPL: A Socratic Playground for Learning Powered by Large Language Model*, http://[CEUR-WS.org/Vol-3840/L3MNGET24_paper13.pdf].   [3]

Zhang, Y., Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, C. Xu, Y. Chen, L. Wang, A. Tuan, W. Bi, F. Shi and S. Shi (2025), "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models", *Computational Linguistics*, pp. 1-46, https://doi.org/10.1162/coli.a.16.   [40]

Zhu, Y., S. Wei, X. Wang, K. Xue, S. Zhang and X. Zhang (2025), "IMeNTi: Bridging Medical Calculator and LLM Agent with Nested Tool Calling", in L. Chiruzzo, A. Ritter and L. Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1   [57]

Annex 3.A. Technical aspects of educational GenAI agents

Technical infrastructure & integration

Implementing a system like the Socratic Playground requires a robust technical infrastructure that integrates the generative AI core with supporting components for memory, logic, and user interaction. This section details the architecture of SPL and similar LLM-driven tutoring systems, highlighting how various modules work together to deliver a seamless educational experience. We also discuss practical considerations such as latency, error handling, and integration with external platforms (e.g. Learning Management Systems and AR/VR environments). Key technical elements include the GPT-4 based dialogue engine with dynamic prompt injection, JSON-configured lesson scripts, state-tracking for personalisation (like modelling the learner's Zone of Proximal Development), and auxiliary tools (for example, domain-specific rubrics or calculators). Front-end design aspects, such as real-time dialogue rendering and interactive scaffolding controls, are also considered, as they affect the perceived responsiveness and reliability of the agent. Throughout this section, we emphasise technical infrastructure and integration strategies for scalability, multimodal support, and system monitoring that help maintain a smooth user experience.

Figure 3.4. The architecture overview of the SPL system
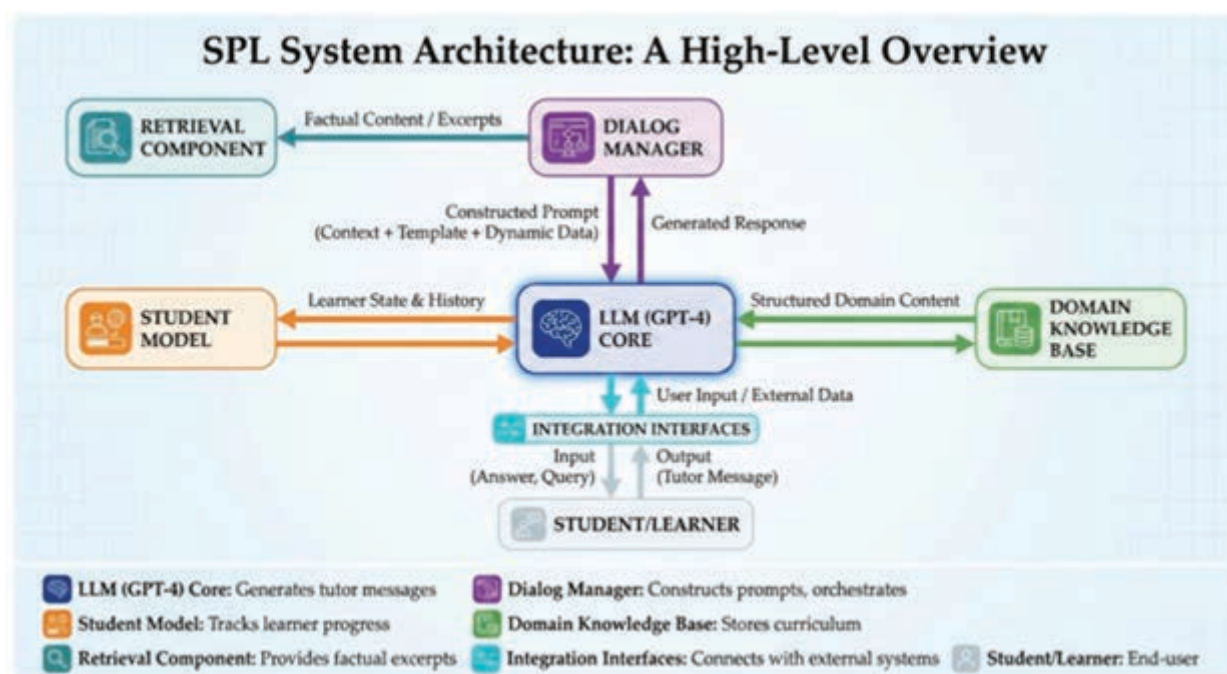
Architecture overview (GPT-4 core and modules)

At the heart of SPL is an LLM (GPT-4) that generates the tutor's messages. This model is accessed via an API, and given a prompt that encapsulates the tutoring conversation and the pedagogical goals for the next interaction. Surrounding the LLM core are several modules: a Dialog Manager, a Student Model, a Domain Knowledge Base, and various integration interfaces (See Figure 3.4). The Dialog Manager is responsible for constructing the LLM prompt each turn by combining the relevant context (e.g. recent conversation history, relevant facts or passages from the curriculum) with the instructional template (as described in the "pedagogical design and interaction frameworks" section). This often involves *dynamic prompt injection* – inserting up-to-date information such as the student's answer, identified misconceptions, or external knowledge into a prompt template before sending the prompt to GPT-4. For instance, if the student is writing an essay about climate change, the system might inject a brief excerpt from a scientific article (via a retrieval component) to ground the agent's feedback in factual content. This retrieval-augmented generation approach is a common way to

combat the LLM's tendency to hallucinate by grounding it in vetted knowledge (Modran et al., 2024[10]). The Student Model maintains a profile of the learner's progress and state. In SPL, this includes tracking which major expectations of the assignment have been met by the learner, and which misconceptions have been exhibited (as mentioned earlier), as well as simpler metrics like the student's overall accuracy rate, response times, and affective indicators (if any). One could think of this module as an evolving memory of the student's Zone of Proximal Development – it estimates the current level of challenge the student can handle with scaffolding. For example, if the student consistently answers a certain type of question correctly, the system might ramp up the difficulty or move on to a new topic; if errors occur, it might stay in the same subtopic but try a different teaching strategy. Research on learner modelling and knowledge tracing (e.g.Gong, Beck and Heffernan, 2011[56]) feeds into this component, though traditional knowledge tracing must be adapted for the unstructured dialogue context.

The Domain Knowledge Base (or Lesson Script repository) is where the JSON-configured lesson scripts come into play. Each lesson or subject area can be defined by a JSON file that lists the key concepts, common misconceptions, example problems, etc., which the agent should be aware of. In SPL's design, these JSON files have entries for "expectations" and "misconceptions" as previously described, and possibly other pedagogical metadata like suggested hints for each concept. They serve as a lightweight expert model that the AI can consult. When GPT-4 is prompted to produce a response, some of this structured information may be embedded or appended (for instance, a summary of which expectation the student hasn't covered yet can be added to the prompt, subtly nudging the AI to steer the student there). This hybrid approach (i.e. combining the generative flexibility of GPT-4 with structured domain guidance in JSON) aims to ensure content accuracy and curriculum alignment. It addresses a key technical challenge: pure end-to-end LLM tutoring might go off-topic or miss curriculum goals, but by integrating a scripted backbone (i.e. the JSON lesson plan), the AI is kept "on track" pedagogically. Notably, this approach was informed by earlier ITS frameworks like Generalised Intelligent Framework for Tutoring, i.e. GIFT (Sottilare, Graesser and Sinatra, 2018[53]) and legacy systems like AutoTutor (Graesser et al., 2005[1]), which emphasised explicit modelling of correct and incorrect knowledge. The innovation here, moving beyond traditional pre-scripted ITS, is that the heavy lifting of dialogue generation and language understanding is done by the LLM, while the structured script provides checkpoints and boundaries.

Memory and zone of proximal development tracking

One of the advantages of modern AI infrastructure is the ability to maintain long conversation histories through extended context windows or external memory stores. GPT-4's expanded context window (up to 8 000 tokens or more in some versions) means that the agent can "remember" everything said so far in a tutoring session without forgetting earlier points – something older chatbots could not. This enables more coherent and contextually relevant interactions over extended sessions. However, long-term memory across sessions (e.g. what the student did last week) requires additional solutions, such as saving a summary of each session to a student profile that can be reloaded later. In SPL, after each session, the system generates a concise summary of the dialogue and learning outcomes (compiled by GPT-4 itself) and stores it in a database. When the student returns, that summary is prepended to the conversation to give context to the agent. This approach to continuity is a practical implementation of treating the agent as a lifelong companion that accumulates knowledge about the learner (Krinkin, 2026 (forthcoming)[20]). Memory also plays into tracking the Zone of Proximal Development: by deriving attributes from interaction patterns (e.g. number of mistakes made, time spent on activities), the system infers what the student is ready to learn next. For example, if the student can answer direct questions but struggles with synthesis questions, the agent will focus support on the higher-order thinking steps – always trying to operate in the sweet spot where the student is challenged but not overwhelmed. Technically, this could be a rule like: "if the student makes two errors in a row on a concept, revert to an easier question or a sub-concept of that topic; if the student answers correctly with high confidence, progress to a harder question or next concept." Implementing such rules can be outside the LLM (in the Dialog Manager) to ensure reliability, rather than hoping the LLM deduces it every time.

Extended functionalities through tool integrations

Another important aspect of infrastructure is integrating external tools that extend the agent's functionality. For instance, in an essay writing support scenario, one might integrate a *writing evaluation* rubric tool. This could be an NLP service that scores an essay on dimensions like coherence, grammar, argument strength, etc. When a student submits a draft or a paragraph, the system can invoke this rubric tool and feed the results into the LLM prompt – enabling the agent to give targeted feedback (e.g. "Your argument is strong, but the organization could be improved. Maybe start this paragraph with a clear topic sentence."). In SPL's current version, we integrated a simple grammar checker and a fact checker. The grammar checker (an off-the-shelf API) identifies any glaring grammatical mistakes in the student's response; the agent then decides whether to mention it (often it will only do so after addressing content understanding, to not derail the student's thinking process). The fact checker (using a search engine or a knowledge base) is used when the student or

agent makes a factual claim; the system can quickly verify that claim and, if it's likely false, the agent can prompt the student to reconsider (e.g. "Are you sure about that fact? Maybe we should verify it."). These integrations act as guardrails to improve the accuracy of the tutoring dialogue and to enrich the feedback. A well-designed generative agent platform should have a modular way to plug in such tools or services. Recent research prototypes, for example, have LLM "planner" agents that can decide to call a diagram-drawing tool for geometry or a calculator tool for math problems (Zhu et al., 2025[57]). The architecture may employ function calling (e.g. via Model Context Protocol) or a JSON-based input-output (I/O) mechanism. In such cases, the language model produces an output in the form of a structured call[2], which the surrounding system executes. The computed result is then passed back to the model for subsequent processing. This can be achieved via OpenAI's function calling API or custom middleware. In summary, the technical stack is not just the LLM; it's an ensemble of AI and non-AI components working in conjunction to deliver a coherent tutoring experience.

Latency and real-time performance

One practical challenge with using large models like GPT-4 is latency. Students and teachers expect responsive systems, and long pauses can disrupt the conversational flow. GPT-4, given its size, can sometimes take a few seconds (or more, depending on server load and prompt length) to generate a response. In a live tutoring setting, even a 5-second delay might feel awkward. SPL addresses this in a few ways. First, prompts are kept as concise as possible – through prompt engineering and the use of system-level instructions that don't need to be repeated verbosely every turn. Second, the front-end provides visual feedback (like a typing indicator or an animation of the avatar "thinking") to reassure the user that the system is working, not stalled. Third, for longer responses, the system streams the output to the interface as it is generated (a capability supported by many LLM APIs). This means the student can start reading the beginning of the agent's answer while the rest is still coming, which mimics a natural dialogue more closely. In terms of infrastructure, we also consider deploying the model on powerful servers or using distillation techniques to have a smaller version for faster real-time interaction when full GPT-4 speed is not needed. There is often a trade-off between accuracy and speed; one idea is to use a faster but slightly less capable model for quick interactions and reserve the full model for more complex tasks or when the session can tolerate a delay. As generative models continue to improve, we anticipate latency will decrease, but it remains a design consideration for now.

Error handling and fallbacks

No AI system is perfect, so the infrastructure must handle errors gracefully. Hallucination, where the LLM produces a plausible-sounding but incorrect statement, is a known issue (Zhang et al., 2025[40]). To mitigate hallucination during tutoring sessions, a multi-layered approach involving prevention (via grounded prompts and post-hoc checks) and mitigation (via user interface and pedagogical strategy) may be adopted. As mentioned, integrating a retrieval mechanism to ground answers can prevent many factual hallucinations (Modran et al., 2024[10]). Additionally, after the LLM produces an answer, a lightweight verifier can assess its correctness. For example, if the question was a math problem, a separate programme can verify the solution; if the question was asking for a definition, a keyword check against a trusted source can be done. If the verifier flags a potential error, the system might either correct it before showing to student or have the agent acknowledge uncertainty and provide a fallback response (e.g. "I cannot provide a confident answer at the moment."). In the SPL implementation, in cases where the agent is not confident about its own responses, the agent is prompted to respond with a question rather than an assertion (turning a potential hallucination into a joint exploration: "That's a complex question – what do you think might be the reason? Let's work through it together."). This way, even if the AI isn't sure, it keeps the student engaged in finding the answer rather than delivering a false answer confidently. Another potentially problematic scenario is if the model output is malformed or content-inappropriate (e.g. it somehow trips a content filter or produces something irrelevant). A well-designed system should actively monitor for such occurrences and employs predefined fallback strategies, including a generic apology, a reformulated response, or a standardised prompt (e.g. "I am sorry, I didn't quite get that. Could you try asking in a different way?"), which designates to elicit a revised input from the user. Logging every interaction along with any error flags is crucial for developers to later review and refine the system. Over time, such reviews help improve prompt strategies or add rules to cover edge cases.

Front-end integration and user experience

The technical infrastructure extends to the front-end application where learners and educators interact with the agent. In SPL's web interface, the conversation with the AI tutor is displayed much like a chat, with each turn labelled by speaker (i.e. Tutor or Student). The design uses simple visual cues: the tutor's messages appear in a speech bubble next to an avatar icon, the student's entries appear on the opposite side. Important phrases in the tutor's text can be highlighted (for example, when the tutor introduces a key term, it appears in bold or a different colour). There is also the capability for the agent to display tabular data or images in-line if needed, e.g. showing a quick table of the student's quiz results

or a diagram. The front-end is built to be modular so that it can plug into Learning Management Systems (LMS) used by schools. For integration with LMS, compliance with standards like LTI (Learning Tools Interoperability) is considered – essentially allowing the AI tutor to launch within a platform like Moodle or Canvas as an external tool. This requires secure authentication (so the agent knows which student and class it is dealing with) and data reporting back to the LMS (such as scores or completion status). While an internal prototype like SPL might not fully implement all LMS integration features, designing the system with APIs for retrieving and posting grades or session summaries makes later integration feasible. Additionally, a teacher dashboard is often part of the envisioned infrastructure: a view where a teacher can see what questions the AI is asking the student, intervene if necessary, or review a transcript after the fact. This aligns with the co-orchestration model where teachers oversee AI interventions (Echeverria et al., 2023[58]). From a technical standpoint, enabling real-time observation means the system should broadcast events (e.g. via WebSockets) so that if a teacher is connected, they receive the stream of dialogue as it happens.

Scalability and monitoring

Finally, for such an infrastructure to be viable in real educational deployment, it must be scalable and maintainable. Scalability refers not only to handling many simultaneous users (which requires load balancing and possibly model instancing for heavy use times) but also scaling to new content areas. Thanks to the generative nature of the AI, the system can be content-agnostic to a degree – the same GPT-4 can tutor math or history – but it needs domain-specific scripts or knowledge bases plugged in for each subject. Thus, adding a new course or topic involves authoring the JSON script for that topic and assembling any domain resources (like a glossary or a set of source texts). A long-term technical goal is to develop authoring tools that let educators create these domain scripts through a user-friendly interface, rather than writing JSON manually. For now, that process might be semi-automated, e.g. an educator fills out a form with key concepts and common misconceptions and the system generates the JSON structure.

System monitoring is also essential for maintenance and improvement. This includes analytics on usage (which questions most commonly cause students to ask for more help, where the AI often generates suboptimal responses, etc.) and automated alerts for problematic behaviour (for instance, if the AI ever produces inappropriate content, it should be flagged and developers should be notified). In the SPL research deployment, all sessions are logged with consent, and a team periodically reviews them for quality assurance and to identify patterns that need attention, such as a certain concept that confuses the AI. By monitoring these logs, developers can update prompts or add new examples to the training/fine-tuning data to gradually improve the system. Reliability monitoring is another aspect – ensuring uptime, quick recovery from any crashes, and measuring any failures in the external tool calls.

Summary

In conclusion, the technical backbone of generative pedagogical agents like Socratic Playground involves a sophisticated orchestration of AI and software components. The GPT-4 core is leveraged for its powerful language generation, but around it we build structures (scripts, memory, tools) to ensure that the result is pedagogically coherent, factually accurate, and contextually appropriate for the learner. Integration with existing educational technology ecosystems (e.g. LMS, classroom devices, VR platforms) further enhances the practicality of the system. Through careful management of latency, resilient error handling, and thoughtful user-centric interface design, the infrastructure is designed to provide a smooth and trustworthy experience. As these systems move from prototype to real-world classrooms, the considerations discussed here – from dynamic prompting to teacher oversight dashboards – will determine how effectively generative AI can be embedded into daily teaching and learning. The In this chapter, the "Working in Practice: The SPL Demonstration System" section shows how such a system operates in practice, showcasing the Socratic Playground demonstration and lessons learned from initial deployments.

4 Fostering collaborative learning and promoting collaboration skills:
What generative AI could contribute

Sebastian Strauß and Nikol Rummel
Ruhr-Universität Bochum, Germany

This chapter illustrates the potentials of generative AI (GenAI) to support collaborative learning and reviews the emerging research. After presenting how technology can support collaborative learning, we illustrate the roles that GenAI can play during collaboration (for example serving as a repository of information or as a teacher/tutor), which aspects of the collaboration it can support (for example providing knowledge or fostering social interaction), and whether this support affects learning in terms of domain-specific knowledge and effective collaboration skills. We conclude by discussing potential ways to combine GenAI with established types of support for groups.

Introduction

Collaborating with others in small groups can be effective for learning, fostering not only domain-specific knowledge but also the knowledge and skills necessary for learning and working in teams (e.g. Hattie, 2009[1]; Pai, Sears and Maeda, 2014[2]; Tenenbaum et al., 2020[3]). At the same time, groups encounter challenges during collaboration (e.g. Barron, 2003[4]; Rummel, Spada and Hauser, 2009[5]; Strauß and Rummel, 2021[6]) and efforts have therefore been undertaken to design support that helps groups navigate these challenges. Work within the learning sciences, especially the field of computer-supported collaborative learning (CSCL), has a long-standing tradition of developing and testing computer-based support for collaborative learning. Such work has now been expanded by research exploring the use of machine learning models, which fall under the term 'generative artificial intelligence' (GenAI) in the context of collaborative learning. However, a clear picture of the landscape of GenAI in CSCL is currently lacking. This chapter therefore illustrates how generative AI systems are currently used to support groups during collaborative learning, and analyses their effectiveness in supporting collaborative learning. To this aim, we review a broad sample of current research studies in which GenAI is utilised to support collaborative learning.

To understand how GenAI can be designed to support collaborative learning, and to determine what is required to provide groups with support that is sensitive to their current needs, it is useful to elucidate how learning during collaboration occurs. Therefore, we first provide a brief overview of the core mechanisms underlying collaborative learning and illustrate why groups benefit from support. Drawing on previous work that focused on utilising computer technology to support groups, we then present the most prominent CSCL approaches to supporting groups. This allows us to place GenAI in the broader context of instructional support, to compare it to previous efforts in this field, and to identify the potentials that this technology brings to the table in terms of providing groups with personalised (i.e. adaptive) support.

The main body of this chapter describes how GenAI systems are currently used to provide groups with support. Here, we illustrate the roles that GenAI can play during collaboration (for example serving as a repository of information or as a teacher/tutor), which aspects of the collaboration can be supported by GenAI (for example providing knowledge or fostering social interaction), whether this support affects learning in terms of domain-specific knowledge and knowledge about effective collaboration, and on what grounds GenAI tools determine when and how groups can be supported.

Our review illustrates the potentials of GenAI for supporting collaborative learning and underlines the need for more systematic efforts to determine how GenAI should be designed to support collaboration, especially in terms of how it affects collaboration processes and subsequent learning. Although more research is needed, future efforts to design effective support using GenAI can leverage established insights from learning science research. We conclude the chapter with an outlook, in which we discuss potential ways forward regarding the design of GenAI support, how this technology can be combined with established types of support for groups, and how we can investigate the benefits of GenAI support and deepen our understanding of its impact on collaboration and learning processes. Importantly, we point to several issues that need to be navigated during this process.

Collaborative learning: Collaborating to learn and learning to collaborate

What is collaborative learning?

Collaborative learning can be understood as "a situation in which two or more people learn or attempt to learn something together" (Dillenbourg, 1999[7]) and a "coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (Roschelle and Teasley, 1995[8]). These definitions demonstrate that collaborative learning can encompass a variety of learning settings in which learners are not merely presented with information but rather actively co-construct knowledge through interaction. Importantly, the interaction between the group members serves the co-construction of knowledge instead of simply creating artifacts such as texts, presentations, computer code, or other artifacts.

Sometimes, the term "collaborative learning" is contrasted with "cooperative learning": While learning together is understood as "collaborative" when all group members are required to interact with each other in order to achieve their goals, it is understood as "cooperative" when the joint task can be divided into subgoals that can be achieved independently, that is, without the need for knowledge co-construction (i.e. a divide-and-conquer approach). Kreijns et al. (2024[9]) provide a more in-depth discussion of this distinction. In this chapter, we use the "collaborative" learning to refer to different arrangements of learning in small groups, without making this explicit distinction.

The beneficial effects of collaborative learning have been repeatedly demonstrated, both for the acquisition of domain-specific knowledge and its transfer (collaborating to learn: for overviews, see Hattie, 2009[1]; Pai, Sears and Maeda, 2015[10]; Tenenbaum et al., 2020[3]) and for the acquisition of collaboration skills (learning to collaborate: for overviews see Radkowitsch, Vogel and Fischer, 2020[11]; Vogel et al., 2017[12]). An important finding from learning science research is that assigning learners to a group, and assigning them a task to solve as a group, does not guarantee learning (Kreijns, Weidlich and Kirschner, 2024[9]). Collaborative learning encompasses the active shared construction of knowledge by the group members (Stahl, Koschmann and Suthers, 2006[13]; Webb, 2013[14]) which occurs through activities such as providing explanations to group members, elaborating on ideas, providing feedback on errors, posing thought-provoking questions, resolving conceptual conflicts, or cognitive modelling (King, 2007[15]). Thus, interaction patterns are key to learning through collaboration (see Dillenbourg et al., 1995[16], for an "interaction paradigm").

An important prerequisite for fruitful interaction between the group members is social interdependence, that is, each group member can only achieve their goals if they assist the other group members to achieve their own goals (Johnson and Johnson, 2009[17]). Besides interacting to co-construct knowledge, groups have to perform several other tasks, which lay the groundwork for continuing co-construction, for instance maintaining a shared understanding (common ground, (Baker et al., 1999[18]; Clark and Brennan, 1991[19]; Hadwin, Bakhtiar and Miller, 2018[20]), actively participating in discussions (Hrastinski, 2008[21]; Matschke et al., 2014[22]), or regulating the collaboration (Hadwin and Oshige, 2011[23]; Järvelä and Hadwin, 2024[24]). Besides these beneficial processes, research has also identified pitfalls that may hamper collaborative learning (for overviews see Kreijns, Weidlich and Kirschner, 2024[9]; Nokes-Malach et al., 2019[25]; Strauß and Rummel, 2021[26]).

Evidently, collaborative learning is complex and challenging for learners. Therefore, learning science research has explored how to help groups engage in effective interaction that eventually affords learning. In the next section, we

provide an overview of the rationales that guide the design of such collaboration support and illustrate the most relevant types of adaptive support, starting with support that does not rely on GenAI. This overview serves as the basis for analysing current implementations of GenAI and discussing future pathways for the development of GenAI support for collaborative learning.

Computer support for collaborative learning: Adaptive scaffolding

Providing groups with adaptive (personalised) support is not a new concept. The earliest efforts in CSCL stem from the 1990s and investigated support that is not adaptive (Jeong, Hmelo-Silver and Jo, 2019[27]; Rummel, 2018[28]). While even such non-adaptive support has been shown to enhance the beneficial effects of collaborative learning (Chen et al., 2018[29]; Jeong, Hmelo-Silver and Jo, 2019[27]; Lou, Abrami and d'Apollonia, 2001[30]; Talan, 2021[31]), adaptive support has been a constant theme in the discourse around collaboration support.

The conceptual basis for providing groups with support that is tailored to their needs is the concept of scaffolding. Scaffolding evokes the metaphor of scaffolds during the construction of buildings, denoting that learners are given the means with which to accomplish hitherto unattainable learning tasks (Kollar et al., 2025[32]; Reiser and Tabak, 2014[33]). Essential during scaffolding is that the learner performs all parts of the task on their own, despite receiving support ("independent activity", see Reiser and Tabak, 2014[33]). Scaffolds come from a more knowledgeable other, such as a teacher, learning materials, or a digital learning environment. For instance, scaffolds can model the problem-solving process, direct attention to important aspects of the task, or elicit explanations (Reiser and Tabak, 2014[33]). Given that every learner enters a learning situation with different prior knowledge, a tenet of scaffolding is adaptivity. The same is true for groups of learners. Such adaptive (or personalised) support is adjusted to the relevant characteristics of the learner, the group or their interaction, with characteristics being "relevant" if they are expected to affect groups' ability to achieve their goal (e.g. acquiring new domain-specific knowledge; (Tetzlaff, Schmiedek and Brod, 2021[34]). This resonates with works in CSCL, especially research on external collaboration scripts (Fischer et al., 2013[35]; Kollar, Wecker and Fischer, 2018[36]; Fischer et al., 2007[37]) as research on expertise and collaborative learning scripts has shown that not all groups benefit from the same amount of support. Rather, support that is too fine-grained can be expected to lead to lower performance in learners who already have high competence (expertise reversal effect, Kalyuga, Rikers and Paas, 2012[38]; over-scripting, Dillenbourg, 2002[39]) whereas learners with little prior experience may struggle to collaborate effectively when they receive too little support ("underscripting": Dillenbourg, 2002[39]). Adapting the support to a group's needs is thus seen as crucial.

Designing for adaptive support

When designing support for collaborative learning, researchers and developers have to make a number of decisions, for which the framework of CSCL design dimensions presented by Rummel (2018[28]) can serve as a conceptual guide (Figure 4.1).

First, developers must decide what the support is ultimately expected to help groups achieve, that is, the goal of the support. This goal may be acquiring new domain-specific knowledge, acquiring collaboration skills, having satisfaction with the collaboration, or other relevant outcomes. Next, it has to be decided how the instructional support can achieve this goal. Rummel (2018[28]) refers to this as the target of the support. For instance, in order to co-construct knowledge that each group member did not hold before, the group may receive support that targets their interaction, for instance prompting groups to engage in a discussion (see Weinberger and Fischer, 2006[40]). The same goal may be addressed by helping groups monitor their understanding and support each other repair misconceptions. Besides other dimensions such as timing, addressee, or availability of the support, it also has to be decided whether the support is implemented in a fixed (every groups receives the same support at the same time), adaptable (each group can decide which support they receive) or adaptive way (a system decides under which circumstances a group receives which degree of support).

The foundation for an adaptive system is a model of relevant prerequisites for collaborative learning and desired states or processes that can occur during collaboration (Deiglmayr and Spada, 2010[41]). This model underlies the processes performed by the automated support system. The early works on adaptive support for collaborative learning comes from the 2010s and leveraged techniques that represent "good old-fashioned", or symbolic, artificial intelligence – they were rule-based.

In order to provide groups with adaptive support during the collaboration, a system has to collect information about the collaboration, process this information to determine whether the group requires support, and then deploy adequate interventions that help the group achieve their goals (Deiglmayr and Spada, 2011[42]; Molenaar, 2022[43]; Walker, Rummel and Koedinger, 2009[44]).

Figure 4.1. CSCL design dimensions

CSCL design dimensions

Why?	**1) Goal**
	☐ Knowledge acquisition (e.g., domain-specific, domain-general)
	☐ Outcomes of the collaboration (e.g., problem solution, artifacts)
	☐ Affective outcomes
	☐ …

What?	**2) Target**
	☐ Cognitive (e.g., domain-specific help)
	☐ metacognitive (e.g., reflection)
	☐ Social (e.g., managing the interaction)
	☐ …

Who and for whom?	**3) Delivery agent**
	☐ Teacher
	☐ Digital system
	☐ Peer learners
	☐ …
	4) Addressee
	☐ Individual
	☐ Group
	☐ …
	5) Mediation
	☐ Direct (provided by individual/group)
	☐ Indirect (provided to the teacher, the parents)

How?	**6) Implementation**
	☐ Fix/static (one size fits all)
	☐ Adaptive
	☐ Adaptable

How?	**7) Granularity**
	☐ Task level
	☐ Step level
	☐ Turn level (e.g., speaking turns)
	8) Directivity
	☐ Implicit (enabling conditions, resources)
	☐ Explicit: general advice
	☐ Explicit: specific advice
	9) Foundation
	☐ No information on state (i.e., just alert)
	☐ Show state (i.e., raw data)
	☐ Show aggregated data (e.g., means)
	☐ Show interpretation of state (e.g., assessment "good" vs. "bad")
	10) Timing
	☐ Before the collaboration
	☐ During the collaboration
	☐ After the collaboration
	11) Availability
	☐ Always visible
	☐ Upon diagnosis (e.g., pop-up)
	☐ On demand
	☐ Upon diagnosis (e.g., pop-up)
	☐ On demand
	12) Coercion
	☐ No specific action required
	☐ Some action required (e.g., clicking "ok")
	☐ Specific action required (e.g., interaction pattern)

Source: Rummel, N. (2018[28]), One framework to rule them all? Carrying forward the conversation started by Wise and Schwarz, *International Journal of Computer-Supported Collaborative Learning,* 13(1), pp. 123 - 129

Following Molenaar (2022[43]) in the first step, a system has to collect information about the learners, the groups and the collaboration process (detect), and process this information into indicators that represent relevant insights into characteristics and processes during collaboration (e.g. the distribution of relevant domain-specific knowledge among the group members). One of the early approaches to automatically collecting and processing relevant information about the collaboration process was leveraging text classification approaches, a group of machine learning techniques from computational linguistics (Rosé et al., 2008[45]). Based on their experiences from an interdisciplinary project, Rosé et al. (2008[45]) discussed how text classification can be utilised to analyse transcripts of collaborative dialogues in order to automatically analyse the process of knowledge co-construction using a multidimensional coding scheme (instead of merely analysing the content of the dialogue). This ability to automatically assess collaboration processes in terms of relevant events (e.g. knowledge co-construction) then provided the basis for deciding upon groups' need for support. The authors urged great care when developing the indicators to be used by an automated system to assess the collaboration process. For instance, it is essential that the indicators are externally valid, and actually represent what they aim to represent (Rosé et al., 2008[45]).

These indicators inform the system about the current state of the collaboration and are then compared to a desired goal-state (diagnose). Depending on the outcome of this comparison, support may be offered (act) (see Molenaar, 2021[46]; Molenaar, 2022[43]). Deiglmayr and Spada (2010[41]) underlined the importance of developing rules that govern the behaviour of the support system. Specifically, they illustrate production rules (IF-THEN rules) that specify which support is offered under which circumstances (also see Bernacki, Greene and Lobczowski, 2021[47]; Radkowitsch et al., 2021[48]). These rules can target critical moments (for example, "IF state is detrimental, THEN offer support A") or opportune moments ("IF potential for beneficial interaction, THEN offer support B": see Deiglmayr and Spada, 2010[41]).

Following the concept of scaffolding, systems that support collaborative learning are designed to only provide support as long as the group requires it. As a consequence, the degree of support is gradually reduced (or if necessary increased) as the competence of the learner or the group changes (Reiser and Tabak, 2014[33]). Such fading can be realised, for instance, by removing certain action prompts after they have been shown for a set number of times (Wecker and Fischer, 2011[49]). Thus, support in (and for) collaborative learning (i.e. scaffolding) is not available indefinitely, but only so long as learners require it.

After illustrating the foundations of the mechanisms behind learning during collaboration and how support for these processes can be conceptualised, we turn to the most central approaches to providing groups with adaptive support. Here, we begin with providing an overview of well-established approaches.

Different approaches to providing adaptive support

Methods for supporting groups include:

1. adaptive collaboration scripts (Edwards et al., 2025[50]; Karakostas and Demetriadis, 2011[51]; Rummel, Mullins and Spada, 2012[52]; Rummel et al., 2008[53]);

2. adaptive tutoring systems that have been modified to accommodate collaborative learning and tutoring (Diziol et al., 2010[54]; Walker, Rummel and Koedinger, 2009[44]; 2011[55]; 2014[56]; Sottilare et al., 2017[57]); or

3. virtual agents in the form of collaborative conversational agents (for an overview see Earle-Randell et al., 2025[58]) and for examples: Araujo et al., 2025[59]; Burkhard et al., 2022[60]; Vizcaíno, 2005[61]) and chatbots (for general overviews see Kuhail et al., 2023[62]; Wollny et al., 2021[63]).

These approaches aim to facilitate interactions that are conducive to knowledge co-construction or that inhibit undesired interaction patterns (Earle-Randell et al., 2025[58]; Fischer et al., 2013[35]; Vizcaíno, 2005[61]; Vogel et al., 2017[12]) while providing learners with opportunities to internalise beneficial interaction patterns (Fischer et al., 2013[35]; Kollar, Fischer and Hesse, 2006[64]). Besides approaches that aim at directly fostering the collaboration in groups, some approaches instead are designed to help the teacher to monitor and orchestrate the students' collaboration (Chounta and Avouris, 2016[65]; Han et al., 2021[66]; van Leeuwen, Teasley and Wise, 2022[67]; Yang et al., 2022[68]).

These systems often use production rules (IF-THEN rules) to generate diagnoses of the collaboration or decisions about the need for support based on production rules. Despite the effectiveness of adaptive collaboration support, over the years it has not been implemented on a broad scale. One reason for this may be the complex boundary conditions for effective collaborative learning which have to be understood well in order to formulate production rules. Moreover, assessing the state of the collaboration in a valid and automated manner is far from trivial (see work on collaboration analytics, for example Wise, Knight and Shum, 2021[69]).

The recent introduction of chatbots that leverage large language models (LLMs) has sparked a new conversation about providing groups with adaptive support, with often high hopes regarding their ability to provide groups with adaptive support.

Supporting collaborative learning with generative AI

As illustrated in the previous section, there is a long-standing tradition of employing computer technology to provide groups with adaptive scaffolding. Research and development in this field was reinvigorated when the company OpenAI provided public access to their LLM GPT 3.5 in the form of 'ChatGPT'. The introduction of LLM platforms such as OpenAI's ChatGPT, Google's Gemini, Microsoft's Co-Pilot, Anthropic's Claude, Mistral's Le Chat, or DeepSeek's Deepseek have been touted as "a new type of pedagogical scaffolding" (Gong et al., 2024[70]) with the potential to reshape (Boudia, C. and Bengueddach, 2024[71]) and revolutionise education (Lai et al., 2024[72]; Yan et al., 2024[73]). Despite not being dedicated educational technology, expectations that this technology can achieve a more thorough personalisation of support than previous attempts are high. The hope is that past efforts can be continued with adaptive support that is more flexible than what has hitherto been possible.

For the context of collaborative learning, authors highlight the capabilities of GenAI (usually LLMs) to process inputs from different modalities such as text, speech, or images, as well as to retain the context of the conversation and generate adequate responses (Jong et al., 2024[74]), such as (almost) human-like, coherent texts (Boudia, C. and Bengueddach, 2024[71]; Cress and Kimmerle, 2023[75]) or other content that is similar to human-created material (Lai et al., 2024[72]). Furthermore, these tools are expected to perform automated assessments of the collaboration and provide immediate, tailored, adaptive scaffolding (Gong et al., 2024[70]) in the form of feedback (Lai et al., 2024[72]; Suciati et al., 2024[76]) guidance, and correction of errors (Güner and Er, 2025[77]; Chu, Xu and Zhai, 2024[78]). Eventually, this adaptive support is expected to lead to improved learning through collaboration.

As described above, to provide adaptive support, a system has to achieve three goals: 1) detect relevant characteristics of the group members, the group as a whole, and the collaboration process; 2) formulate a diagnosis about the current state of the collaboration by contrasting it with a desired goal state; and 3) select and deploy scaffolds (see Molenaar, 2022[43]).

Generative AI may detect and diagnose tasks. Like approaches from computational linguistics (see Rosé et al., 2008[45]), generative AI LLMs have natural language processing capabilities that can make a valuable contribution to the analysis pipeline within collaboration support. For example, Zheng et al. (2021[79]) used an LLM to label data from the interaction in groups. After labelling (i.e. categorising (or annotating) different collaborative actions according to a coding schema), the support used production rules to generate support for the groups. This use case for LLMs hinges on the accuracy with which the model assigns labels to collaboration data (e.g. audio, video, computer log files). Thus, an essential question is to what extent LLMs are capable of labelling data reliably based on a coding schema. Some studies found acceptable labelling performance across multiple labels when compared to human coding (as indicated by interrater reliability scores greater than 0.7, e.g. Amarasinghe et al., 2023[80]), while others reported lower overall performance (Huang et al., 2025[81]; Yin et al., 2025[82]) or high performance only for certain labels (Wang et al., 2023[83]). Thus, there is potential to integrate LLMs into the analysis that lays the foundation upon which to provide groups with adaptive support. But there are still obstacles to overcome, especially in terms of consistently accurate diagnoses so that groups do not receive inadequate support due to misclassifications. In this context, Wong et al. (2025[84]) recently demonstrated that the accuracy of automatic coding can be further improved by using models that are able to leverage inputs from multiple modalities, such as speech (audio) and written text.

What roles can generative AI assume in order to support collaborative learning, and what is being supported?

In this chapter, we focus on how researchers and developers have leveraged GenAI models to provide groups with scaffolds and to what effect. Therefore, we collected a broad sample of articles from scientific journals and conference proceedings that reported on studies in which generative AI systems were implemented in collaborative learning scenarios. Notably, we observed that studies in this field tended to use LLMs as the foundation of support, while rarely leveraging them to expand the capabilities of already established types of collaboration support, such as collaboration scripts or pedagogical agents. To gain a clearer picture of the landscape of GenAI support for collaborative learning, we analysed these studies and determined 1) which role a support system may assume during collaboration to scaffold the collaboration, 2) which aspects of the collaboration are scaffolded (i.e. the dimension 'target' in the framework proposed by Rummel (2018[28]), 3) whether the tools are effective in supporting knowledge acquisition during collaborative learning, and 4) on what grounds the adaptive behaviour of the tools is designed.

Inspecting current studies on collaborative learning with generative AI, we derived four different ways in which generative AI was integrated into collaborative learning settings in order to afford or support the collaboration. These are characterised below, with examples provided (see Figure 4.2 for an illustrative overview).

Figure 4.2. Roles GenAI can assume to support collaborative learning
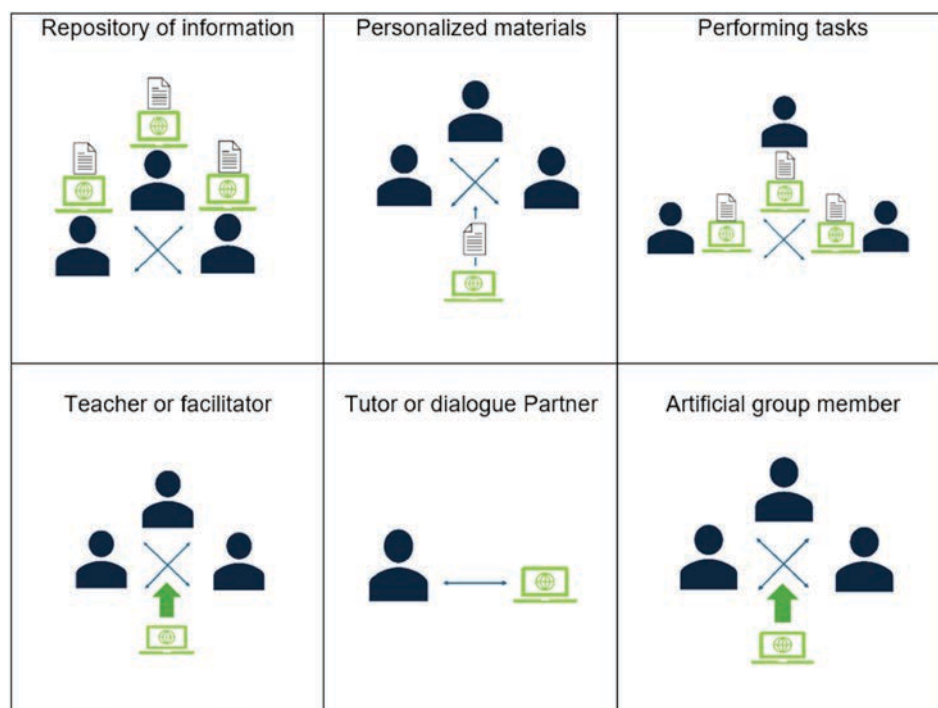
Repository of information

First, there are studies in which GenAI (i.e. LLMs) served as a repository of information, which groups can query to obtain information that can be used to solve their tasks (e.g. Chu, Xu and Zhai, 2024[78]; Darmawansah et al., 2025[85]; Feng, 2025[86]), thus serving the function of a web-search. For example, Darmawansah et al. (2025[85]) implemented ChatGPT to support groups during argumentative knowledge co-construction. In the first phase of the collaborative activity, groups were tasked to research information, which they subsequently used to develop arguments. During this phase, groups could search for information about different topics using ChatGPT or other sources. In the subsequent phases of the activity, students discussed using their arguments.

Personalised learning material and performing tasks for the group

Generative AI can further be a source for personalised learning material (Naik et al., 2024[87]), that is, a tool that collects information about the group, its collaboration, or its results, and generates additional learning material that the group can use. For instance, Naik et al. (2024[87]) implemented GenAI in a collaborative scenario where groups first developed a solution for a problem and then contrasted their solution to an alternative solution, subsequently discussing trade-offs between the two solutions. In this case, the group's solution was processed by ChatGPT using prompts from the researchers, who tasked the model to create an alternative solution, a so-called contrasting case. Contrasting cases present learners with alternative solutions to the problem (i.e. cases). By contrasting their own solution to another case, learners can gain a deeper understanding of the underlying principles of the correct solution to the problem (Alfieri, Nokes-Malach and Schunn, 2013[88]). The group then had to discuss the merits and drawbacks of its own and the generated solution. In other studies, generative AI systems were used as tools that performed tasks for the groups, such as generating narrations or images based on inputs from the group (Wei et al., 2025[89]).

Teacher or facilitator

Other studies used GenAI to intervene and scaffold the collaboration (Cai et al., 2024[90]; Feng, 2025[86]; Liu et al., 2024[91]), thereby assuming the role of a teacher or facilitator. In this role, GenAI is positioned outside of the group, monitors the collaboration, and provides support. For example, Cai et al. (2024[90]) implemented a chatbot based on ChatGPT, which followed pre-specified rules to facilitate active participation during collaboration. To this end, the system monitored the participation and alerted the group if the participation was uneven, prompting the group to ensure that all group members contributed. Furthermore, the bot scripted the discussion based on the stages suggested by Tuckman (1965[92]) to promote knowledge construction, for instance by asking follow-up questions and steering the conversation back on-topic if necessary.

Feng (2025[86]) described a system that only provided support upon request. Specifically, groups could ask the system to provide summaries of or feedback on the group's discussion. An analysis of the interaction between group members and the chatbot revealed that most of the groups' requests were cognitive interactions, such as asking task-related questions or requesting the chatbot to perform tasks like formatting outputs. In other examples where GenAI assumed this role, the system was designed to promote reflection about the collaboration process, for instance by monitoring a group's interaction in terms of group norms and promoting reflection about the interaction in the group (Ko and Foltz, 2025[93]), by helping supported groups to interpret and reflect upon information from a learning analytics dashboard during a debriefing session after a collaboration phase (Echeverria et al., 2025[94]), by providing feedback on texts that were created during a knowledge-building activity (Cao et al., 2025[95]), or by facilitating the process of providing peer feedback (Greisel et al., 2025[96]).

Tutor or dialogue partner

A GenAI system may also assume the role of a tutor or dialogue partner. This role relates to that of a teacher, but instead of providing guidance on the interaction in the group, the GenAI system serves as a partner in a dialogue aimed at developing domain-specific knowledge. For example, (Ahlström et al., 2025[97]) described a collaborative learning scenario where groups interacted with digital characters in a virtual reality (VR) environment (e.g. a hurricane evacuee or a climate scientist) in order to collect information that was needed to solve a task. These characters represented a storytelling element and were not designed to act as pedagogical agents (Ahlström et al., 2025[97]).

A more widespread implementation of pedagogical dialogue partners can be found in arrangements where an individual learner interacts with an artificial agent acting as a tutor or partner for Socratic dialogue (e.g. Dang et al., 2025[98]; Goda et al., 2024[99]; Song et al., 2025[100]). As one such example, Goda et al. (2024[99]) described a chatbot-based system that was structured around a set of principles for Socratic dialogue, such as providing learners with structured questions that are expected to elicit critical thinking and analyse their own reasoning instead of providing direct answers to students' questions. In a similar vein, Dang et al. (2025[98]) developed an LLM chatbot for interaction during a mixed-reality learning activity. The chatbot was designed to make use of the conversation history between learner and chatbot in order to provide personalised responses that sought to promote critical thinking, such as asking open-ended questions and encouraging elaboration and reflection (Dang et al., 2025[98]). We wish to note that this application of generative AI arguably represents a fringe case within the broader landscape of collaborative learning, which traditionally emphasises interaction between at least two human learners. We include it here since the interaction with an artificial interaction partner can elicit cognitive processes conducive to learning that can also be found in collaboration between human learners (e.g. eliciting argumentation and self-explanations). However, a separate discussion about whether these contexts truly represent collaborative learning is required.

Artificial group member

Finally, generative AI may also be implemented as an artificial group member. In this role, the system is positioned as part of the group and participates in the learning activity, for example by contributing domain-specific knowledge to dialogues (e.g. contribution by Hernandez-Leo et al., 2025[101]; Liu et al., 2024[91]; Zhu et al., 2023[102]) or by facilitating interaction (e.g. An et al., 2024[103]; Liu et al. (2024[91]) implemented support for learning about interdisciplinary collaboration: students were assigned the role of different experts and then collaborated to solve a problem in a team. In one group, a conversational agent based on ChatGPT functioned as a peer group member. This agent played an engineer and provided respective expert knowledge to the discussion in the group. Similarly, Hernandez-Leo et al., (2025[101]) implemented an artificial peer in a knowledge-building activity. This agent participated in the activity and was designed to elicit discussions and promote critical thinking skills by submitting its own answers to the knowledge-building environment, rating other students' answers, and participating in the knowledge-building discussion.

Targets of support

While assuming these different roles, the support can scaffold different aspects of the collaboration to achieve its goal, that is different targets (Rummel, 2018[28]). In the following, we present examples of different targets that we identified in the literature (Figure 4.3). Notably, in some studies, groups were free to choose how they used the GenAI tool (e.g. An et al., 2024[103]; Feng, 2025[86]; Wei, Li and Lan, 2024[104]). Not all of the studies explicitly reported the target of the support, in which case the target(s) of the support can only be assumed based on the reported results or illustrative screenshots provided by the authors.

Figure 4.3. Targets of GenAI support and examples
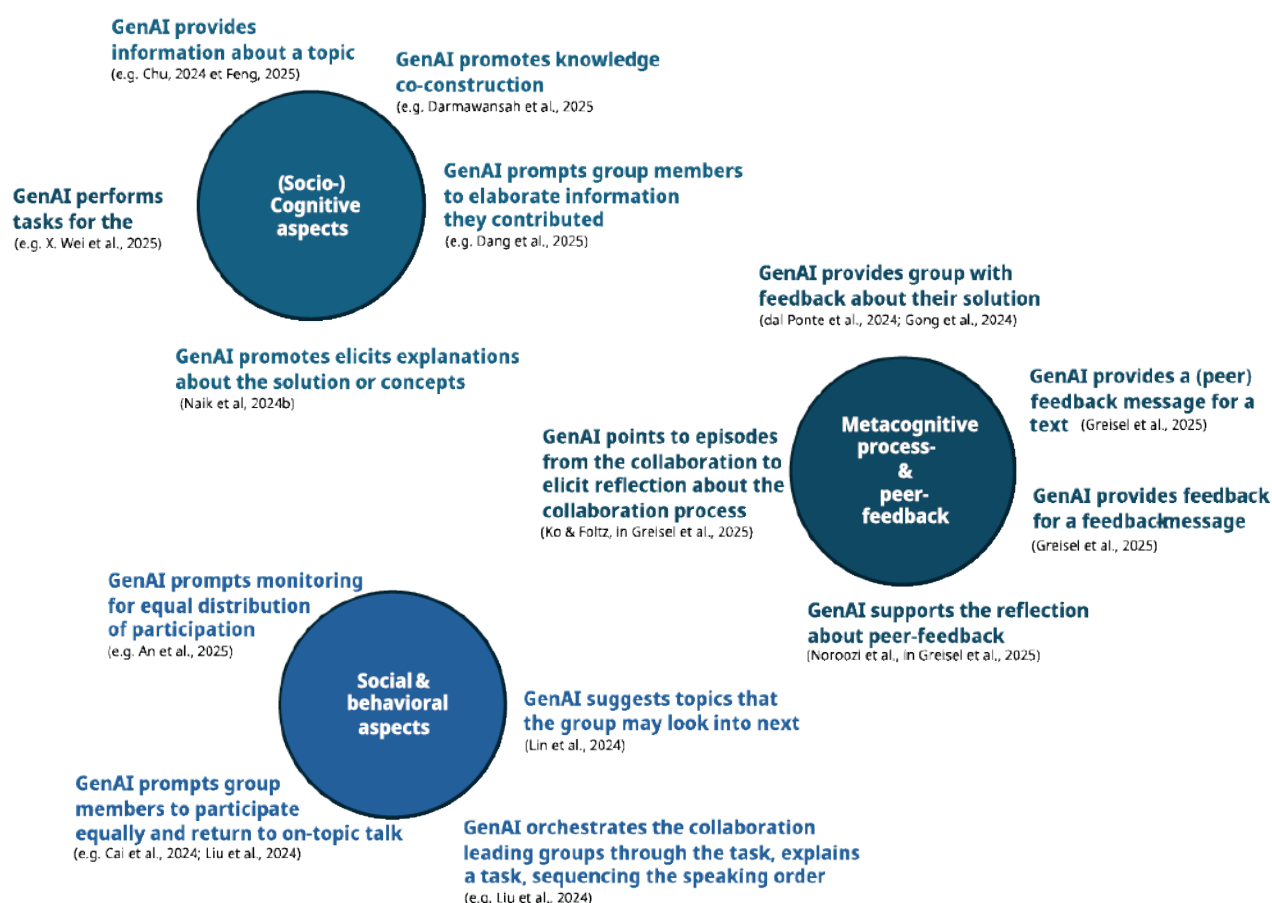
(Socio-)Cognitive aspects

Support can target cognitive aspects of the collaboration, for instance by providing information or resources that the group needs to solve its task and learn (e.g. Chu, Xu and Zhai, 2024[78]; Darmawansah et al., 2025[85]; Feng, 2025[86]; Lin et al., 2024[105]; Zhu et al., 2023[102]) implemented a chatbot based on ChatGPT that acted as a peer during the collaboration. One version of the chatbot in this study provided disciplinary knowledge that was necessary to solve the collaborative task. In other studies, the LLM-based system aimed to elicit cognitive processes such as elaborating on new information (e.g. Dang et al., 2025[98]).

There are also approaches, where GenAI tools are provided so that groups offload tasks to them. For instance, Wei et al. (2025[89]) provided groups with a combination of different GenAI tools (ChatGPT, Midjourney, Runway) which performed different tasks for the group, such as creating storyboards, images, and videos. These approaches exist cannot necessarily be conceptualised as scaffolding.

Other authors explored how generative AI can scaffold socio-cognitive interaction patterns that afford the co-construction of knowledge in the group (e.g. An et al., 2024[103]; Darmawansah et al., 2025[85]) or promote reflection processes (e.g. Naik et al., 2024[87]). For instance, the systems by Lin et al. (2024[105]) and Naik et al. (2024[87])

highlighted certain parts of the groups' solutions and asked the group members to reflect on their solution and explain concepts included in the solution.

Metacognitive processes and (peer) feedback

Metacognitive aspects of the collaboration have often been scaffolded by providing groups with feedback about their work (e.g. dal Ponte, Dushyanthen and Lyons, 2023[106]; Feng, 2025[86]; Gong et al., 2024[70]). Dal Ponte et al. (2023[106]) presented a system that evaluated a group's solution and then provided the group with feedback about the quality of the solution. Groups could then react to the feedback and revise their solution. Moreover, when groups were freely able to choose how to use an LLM chatbot, some explicitly requested feedback on their written solution (Gong et al., 2024[70]).

Another popular method of providing support in the context of feedback is peer feedback (Hornstein et al., 2025[107]; Huisman et al., 2019[108]). Greisel et al. (2025[96]) presented two approaches on how generative AI may facilitate this process: The first approach (contribution by Greisel et al., 2025[96]) consists of a system that processes a student-written text and generates a message with feedback for this text, while another student (the peer reviewer) also reads the text and creates their own message with feedback. Subsequently, the peer reviewer reviews both feedback messages and creates a feedback message that combines the machine-generated message and the human-generated message, before sending the feedback message to the student who created the text. In the second approach, an LLM provided feedback on a peer reviewer's feedback message. With the help of this machine-generated feedback, the peer reviewer then revised their own feedback before sending it to the original student. Noroozi et al. (2025[96]) further noted that an LLM may also help learners to reflect on the peer feedback they have received or to revise their text using peer feedback.

Another way in which generative AI systems can provide feedback relates to promoting collaborative reflection about the collaboration process. The system described by Ko and Foltz (contribution in Greisel et al. (2025[96]) monitored the collaboration in terms of previously negotiated group norms and provides examples from the collaboration process so that the learners can reflect upon their collaboration after collaborating.

Social and behavioural aspects

Some systems target social aspects of the collaboration, such as the participation of the group members (Cai et al., 2024[90]; Gong et al., 2024[70]; Liu et al., 2024[91]) or off-topic discussions (e.g. Cai et al., 2024[90]). As described above, the systems implemented by Cai et al. (2024[90]) or Liu et al. (2024[91]) monitored the amount of participation in the groups and prompted groups to regulate the distribution of participation if it was uneven. Similarly, An et al. (2025[109]) implemented an artificial group member that monitored the interaction to detect moments of uneven participation, prompting inactive group members to contribute to the discussion.

The system by Cai et al. (2024[90]) additionally monitored the content of the conversation to identify off-topic talk. If off-topic conversations were detected, the system would prompt the group to return to the main task.

Other systems focused on guiding and orchestrating the collaboration on a behavioural level, for example by leading the group through the task (Cai et al., 2024[90]; Lin et al., 2024[105]; Liu et al., 2024[91]). For instance, Liu et al. (2024[91]) designed a system that moderated the collaboration by guiding group members through the different phases of the collaborative task, explained tasks and orchestrated turn-taking during discussions. The system presented by Lin et al. (2024[105]) served as a summarisation aid to create feedback for the group and propose potential further topics with which the group may familiarize itself. Specifically, based on the contributions of the group members, the system created feedback messages that summarised (the model's "understanding" of) what the group members already understood and how they could proceed with their task.

Socio-emotional aspects

Finally, some systems were designed to support socio-emotional aspects during collaboration, which includes motivating students, regulating emotions during learning, or giving compliments. While studies such as those reported by Dang et al. (2025[98]) and Feng (2025[86]) investigated the interaction between learners and generative AI systems in terms of socio-emotional processes, studies in which generative AI is deployed to specifically promote beneficial socio-emotional processes or states are rare.

What are the effects of generative AI support on knowledge acquisition?

As we have illustrated, GenAI based systems (usually LLMs) can assume several different roles and target different aspects of the collaboration. The ultimate aim of supporting groups is to afford interaction patterns that benefit groups in terms of achieving their goals, for instance learning. Learning may encompass the acquisition of domain-specific knowledge as well as knowledge that enables group members to collaborate more effectively. It is important

to note that beneficial effects of a support tool depend on whether and to what extent the tool can elicit conducive interaction patterns. Against this background, asking about "the" effect of generative AI on collaborative learning is not precise enough. Instead, we should wonder how learners and groups interact with a tool while working on a task, which individual and collaborative learning processes are sparked in this way, and how this in turn affects outcomes such as the co-construction and acquisition of knowledge. This insight comes from the so-called media/methods debate which is essential for contemporary research on technology-enhanced leaning (see Clark, 1994[110]; Kozma, 1994[111]; Tennyson, 1994[112]; Weidlich et al., 2025[113]; see Sailer et al., 2024[114] for a meta-analysis). In the following, we briefly summarise the evidence on the effectiveness of generative AI support for learning.

Acquiring domain-specific knowledge

Of the studies described above, several employed a quasi-experimental or experimental design, usually in the field (such as university courses), which is the gold-standard to investigate effects of support on outcomes such as domain-specific knowledge (Lin et al., 2024[105]; Naik et al., 2024[115]; Naik et al., 2024[87]). Notably, the studies we identified all focused on knowledge about computer programming. For example, Lin et al. (2024[105]) conducted a field experiment with university students, comparing the effect of support from a GenAI system on knowledge about Arduino programming with a control condition that did not receive any support during collaboration. The GenAI system served as a chatbot that provided information in the first phase of the collaboration and later provided the group with a summary of the group's current understanding as well as with topics the group should subsequently explore. The authors reported a medium-sized positive effect ($\eta^2 = 0.035$), suggesting that the groups benefited from the interaction with the chatbot of support in terms of knowledge about programming.

Naik et al. (2024[116]) examined the effect of personalised reflection episodes in which groups of approximately three students were prompted to discuss specific aspects of their current tasks compared to a control condition without support. Using ChatGPT, the reflection prompts were tailored to content of the groups' discussion. While the authors reported that all students increased their knowledge about programming, they did not find a statistically significant effect of the personalised reflection prompts on students' knowledge.

In another study, Naik et al. (2024[117]) investigated the effect of different sources of contrasting cases (self-generated, personalised by generative AI) on students' domain-specific knowledge in computer programming. The authors compared a control condition that did not perform a contrasting cases comparison (but did perform a reflection activity) with a condition in which groups had to generate a contrasting case for their solution on their own, and with a condition in which ChatGPT was prompted to construct a contrasting case for the group based on the group's solution. While the analyses did not yield an overall effect of the machine-generated contrasting cases on students' knowledge, the results suggested that students with lower prior knowledge benefited more from machine-generated contrasting cases than did students with higher prior knowledge. In turn, students with higher prior knowledge showed a higher post-test score when their group had to generate a contrasting case on their own .

In summary, several studies investigated the effectiveness of different approaches regarding how systems that leverage generative AI support the acquisition of domain-specific knowledge. These studies tended to find positive effects of support on students' knowledge. Nevertheless, it should be noted that the studies employed different approaches to fostering collaborative learning, and that due to methodological limitations, caution is warranted when interpreting the results. We will discuss these points in more detail at the end of this chapter.

Acquiring collaboration skills

To initiate and maintain effective collaboration that eventually leads to learning, groups need to perform a broad variety of actions (see e.g. King, 2007[15]; Schürmann et al., 2023[118]; Strauß et al., 2025[119]). Despite the crucial role of collaboration skills, only a few of the studies we found investigated them as an outcome (or dependent) variable. One example is the study by Darmawansah et al. (2025[85]), who compared argumentative speaking performance, as well as the complexity of the arguments, before and after collaborating with support from a generative AI system when learning English as a foreign language. In one phase, the groups could query ChatGPT for information about the topic, and in a subsequent phase, groups used predefined prompts for ChatGPT that processed the input from the group, for instance "Rewrite these arguments using the argumentation model". The results revealed a significant large effect ($\eta^2 = 0.33$) of the scaffolding on students' argumentative speaking performance and on the complexity of the arguments that students were able to provide after the collaboration.

Another study, by X. Wei et al. (2025[89]), examined the effect of a combination of different GenAI tools (ChatGPT, Midjourney, Runway) during a digital storytelling activity on students' collaborative problem- solving skills. Over several weeks, students created digital videos that told short stories. Groups in the experimental condition used the different GenAI tools to create storyboards, images, and videos. As the dependent variable, the authors used

a questionnaire on which participants rated statements regarding their past collaborative behaviour and potential reactions to hypothetical collaborative situations. Learners who collaborated using these tools achieved a higher score on this measure than did their peers who collaborated without GenAI tools (large effect, $\eta^2 = 0.16$).

Taken together, there is still very limited research on the effects of different approaches to using generative AI in order to scaffold collaborative learning with the aim of promoting collaboration skills. Further studies in the field have investigated other outcomes or characteristics of the collaboration process, such as group performance (An et al., 2024[103]; Naik et al., 2024[116]), interaction patterns and participant roles that emerged during collaboration when participants had access to a generative AI system (An et al., 2024[103]; Gong et al., 2024[70]; Zhu et al., 2023[102]), reflective thinking (Lin et al., 2024[105]), the perceived influence of ChatGPT on group dynamics (dal Ponte, Dushyanthen and Lyons, 2023[106]) or the overall perception of a generative AI chatbot (Cai et al., 2024[90]).

Rationales guiding the adaptivity

Research emphasises that scaffolds are effective if they are designed in accordance with didactic principles (Kollar et al., 2025[32]). Against this background, we explored how generative AI systems were designed so as to provide groups with adaptive support.

Relying on the model/ outsourcing adaptivity to the large language model

One approach utilised by authors to afford personalised scaffolding was to rely on the generative model (i.e. LLMs) to perform monitoring, diagnosis, and decisions about specific instructional actions (e.g. Chu, Xu and Zhai, 2024[78]; Greisel et al., 2025[96]; Jong et al., 2024[74]). In some cases, groups could use the generative AI as they wished; consequently, the specific adaptive actions from the system were dependent on groups' prompts or the behaviour that was transmitted to the LLM (e.g. Chu, Xu and Zha, 2024[78]; contributions in Greisel et al.,2025[96]; contributions in Hernandez-Leo et al., 2025[101]; Wei et al., 2024[104]). In these cases, if the model received a request aimed at receiving feedback or guidance (e.g. Gong et al., 2024[70]; Suciati et al., 2024[76]), the guidance would be dependent on the qualities of the model (i.e. LLMs such as ChatGPT 3.5).

Prompting instructions/ providing the large language model with general instructions

Other researchers used prompts to instruct the GenAI system how it should react. One such approach is to specify a role that the system should assume and specific evaluation criteria that the system should use to assess inputs from a group (e.g. An et al., 2024[103]; Dang et al., 2025[98]; dal Ponte, Dushyanthen and Lyons, 2023[106]; Feng, 2025[86]). For instance, groups participating in the study by dal Ponte et al. (2023[106]) were tasked to develop evaluation plans, which were subsequently assessed by ChatGPT. The authors provided the model with an instructional prompt delineating the role that the model should assume (i.e. an expert) as well as criteria against which the model should evaluate the group's solution: "Act as an expert in the learning health systems framework with a focus on socio-technical evaluation plans. [...] Meticulously analyse its content based on the specific categories" (p. 2). These categories included the method of evaluation: "Assess the appropriateness, robustness, and feasibility of the chosen evaluation method" (p. 2), or the data source: "Critically examine the listed data sources for their relevance, reliability, and potential to address the evaluation's objectives" (p. 2).

An example of a prompt that was used to design an artificial peer can be found in An et al. (2024[103]), who tested a chatbot for collaborative learning. The authors gave GenAI models (Ernie-Speed-128k and Qwen LLM) prompts such as "You are a student agent named Alice who will participate in a group discussion with several students on [topic]" (p. 3), or "Once the discussion starts, you should guide the conversation and help students delve deeper into the questions being discussed by sharing various viewpoints." (p. 3). The artificial agent was implemented in a chatroom where students could collaborate and participated in the discussion.

A more complex system, which makes use of multiple LLM-based agents, was described by Wu et al. (2025[120]), who examined a multi-agent system that was used to support groups of learners during a programming task. The authors specified the system's behaviour by giving it directions about the behaviour the system should exhibit (e.g. style and persona), the format the system's responses should have, and example responses (Wu et al., 2025[120]).

An alternative approach to researchers or teachers providing the evaluation criteria that should be used by a generative AI system is to allow the groups to determine the criteria themselves. One such example was presented by Ko and Foltz (contribution in Greisel et al. (2025[96]), who described a generative AI system that monitors groups' discussions against the background of group norms that were designed by the groups prior to the collaboration. Afterwards, the system facilitates the reflection on the collaboration by providing examples from the collaboration process in which these norms were visible.

Production rules

As an alternative to using qualitative instructions to steer adaptive interventions by the generative AI system, interventions may be based on production rules (IF-THEN statements; some authors also use the term 'triggers', e.g. Naik et al., 2024[116]). For instance, the system described by Cai et al. (2024[90]) counted the number of contributions made by each group member. If the relative number of contributions of a group member fell below 10%, then the system would issue a prompt to encourage this group member to participate more. Similarly, Naik et al. (2024[117]) implemented several production rules that used expression-based matching to identify specific types of code fragments that groups used while solving their programming tasks. Each production rule was aimed at encouraging reflection about one of five learning goals. For example, if the code entered by the students included a specific expression that aimed at altering the data type of a column, then the system would prompt the group to discuss their selection.

Fine-tuning and providing context information

Two further approaches to guide a generative AI system in providing instructions to groups are to fine-tune the system and to provide it with additional information about the learning context. Fine-tuning consists of adapting a pre-trained model for specific tasks or use cases using a smaller, specific dataset. One example of fine-tuning can be found in An et al. (2024[103]). While the authors stated that they used "extensive tutor dialogue data [...] to train the LLM, ensuring the generation of more professional responses for instructional and pedagogical guidance" (p. 2), they did not provide further details regarding the corpus of tutoring dialogue or the training procedure. Lin et al. (2024[105]) also mentioned that they fine-tuned ChatGPT before implementing it into their GPT-Assisted Summarization Aid (GASA), but did not provide further information on this.

Instead of fine-tuning an LLM, Feng (2025[86]) provided their LLM chatbot with presentation slides and a description of the collaborative problem-solving task that were used in the course that students attended, in addition to specifying the chatbot's role and desired behaviour (see above). Feng (2025[86]) argued that this approach is associated with lower costs than fine-tuning and allows for adjustments as soon as the system is in use.

Conclusions: leveraging the potentials of GenAI for collaborative learning

In response to generative AI, educational researchers and developers are exploring novel ways to adaptively support collaborative learning using this new type of machine learning model (usually LLMs). To determine how best to design effective support, and to establish for which learning outcomes such support has proven beneficial, we analysed a broad sample of studies. Overall, we identified and illustrated the potential roles of GenAI support in collaborative learning settings – often higher education settings. Upon closer scrutiny of the studies, we saw that GenAI systems were often designed to assume the role of a tutor or facilitator, or to function as a repository of information (akin to a search engine), while the role of artificial group members was explored less frequently. The support targeted different aspects of the collaboration in order to promote collaborative learning, for example by providing domain-specific information, facilitating discussions, eliciting reflective thinking, or nudging equal participation of all group members. Moreover, a significant proportion of the studies ultimately relied on human intelligence to implement evaluation criteria and to specify rules to govern the desired pedagogical behaviour of the generative AI (rather than relying on the LLM to make suggestions).

Given the small number of studies that have systematically investigated the benefits of GenAI support for knowledge acquisition, it is currently somewhat difficult to determine how these potentials of GenAI to support collaborative learning can best be leveraged. In terms of promoting the acquisition of domain-specific knowledge, results are mixed, with two of the three studies that focused on domain-specific knowledge as an outcome finding small to medium effects, although these studies targeted a very specific domain, namely computer programming (Lin et al., 2024[105]; Naik et al., 2024[87]; Naik et al., 2024[115]). Studies that focused on learning to collaborate – instead of collaborating to learn – were especially rare. However, Darmawansah et al. (2025[85]) reported promising findings with respect to supporting students' argumentation skills.

Our overview shows that the potentials of GenAI are indeed being explored, and conceptual and empirical work continues. Previous research on collaborative learning has already accumulated a wealth of insights into boundary conditions for effective collaboration. These insights represent a fruitful basis for continuing research on adaptive support that leverages GenAI, especially LLMs. Most importantly, collaborative learning scenarios should create social interdependence (Johnson and Johnson, 2009[17]), while support is most beneficial when it can elicit interaction processes cognitive, metacognitive, motivation or affective processes that benefit the knowledge construction in the group.

Conceptually, it is important to discuss which roles GenAI can realistically play in the process of detecting and diagnosing collaboration processes, how it can generate adequate instructional support, and how it may be integrated into existing well-researched approaches to supporting collaboration. Empirically, it will be beneficial to explore whether GenAI models actually provide the support that designers are expecting, whether the supporting actions are in line with beneficial actions identified by previous research, and whether supporting actions by GenAI elicit beneficial collaboration processes. Ideally, the conceptual and empirical perspectives should be combined in order to design and test collaboration support. In the final section, we share ideas on designing collaboration support as well as essentials for the design of future studies in this field. In doing so, we hope to encourage researchers and developers to collaborate and to continue charting out the potentials of GenAI for collaborative learning.

Outlook: Impulses for the future of GenAI in CSCL

After illustrating the current landscape of the use of generative AI in collaborative learning contexts, the question arises of what the future may hold. As researchers, educators, developers, and policymakers, we must play an active role in shaping this future by utilising educational theories and empirical evidence to design instructional support that benefits learning in small groups. This section presents our thoughts on the current discourse and hopes regarding the development and implementation of generative AI for collaborative learning, but also regarding the necessary research to inform the development of support for collaborative learning.

GenAI to support collaborative learning: stand-alone solution and auxiliary function

While there is already ample research on collaborative learning and adaptive support for collaboration, first hopes on leveraging GenAI to support collaborative learning tend to rely on the use of LLMs as a stand-alone solution for instructional support. Given the rich tradition in adaptive collaboration support, this trend raises an important question: which aspects of the design of LLMs warrant the assumption that LLMs are indeed capable of performing all actions necessary for effective support in collaborative learning?

As illustrated above, to generate adaptive support for groups, generative AI models (and humans) must be capable of performing activities such as those laid out in Molenaar's detect-diagnose-act framework (Molenaar, 2021[46]). Research in the areas of CSCL, learning analytics, and AI in education have been exploring how to derive indicators from multimodal data streams that represent objective, reliable, and valid operationalisations reflecting relevant characteristics of learners, groups, and collaborative interaction. These efforts have shown that this is no trivial task. Besides conceptual work (Drachsler and Goldhammer, 2020[121]; Wise, Rosé and Järvelä, 2023[122]), there are also empirical approaches that aim at exploring multimodal data streams (e.g. video, voice, computer log files) to detect aspects of the collaboration process (Sottilare et al., 2018[123]; Zhou, Suraworachet and Cukurova, 2024[124]; Järvelä et al., 2021[125]). Notably, the results described in the overview by Schneider et al. (2021[126]) point to persisting challenges to derive valid indicators for different aspects of collaborative learning.

Supposing that the type and goal of the support have already been decided, a GenAI model then needs to determine whether support is required and then generate supporting actions that elicit productive interaction (e.g. by providing feedback, contrasting cases, or prompts). In other words, instead of merely predicting a 'next token' (i.e. parts of words or sentences) generative AI systems are assumed to be capable of predicting 'the next instructionally adequate actions'. Given criticisms that LLMs represent 'stochastic parrots' (Bender et al., 2021[127]) and potentially generate advice that reproduces neuromyths as Richter et al. (2025[128]) illustrate, one may wonder whether it is realistic to expect LLMs to be capable of providing instructionally beneficial support, in the sense of a 'next token pedagogy'. In this regard, a thorough discussion among researchers and developers is needed to clarify our expectations of LLMs in terms of their ability to detect relevant characteristics of collaborative learning and groups, leverage this information for diagnoses, and generate adequate supporting actions that address the diagnosed states. This discussion should be accompanied by empirical investigations into the functioning of LLM-based support that compares machine outputs with best practices derived from research on learning and instruction. As we have seen, many researchers specify pedagogical guidelines that the LLMs are expected to follow. However, whether the systems indeed produce the desired outputs in response to events during collaboration or groups' requests has rarely been reported (c.f. An et al., 2024[103]).

Exploring contributions of GenAI to established types of support

A second approach (and question to discuss) is whether generative AI applications such as LLMs can complement or expand established means of collaboration support. Given the wealth of research on collaboration support, it is surely beneficial to explore how to leverage the capabilities of generative AI in combination with already existing types of collaboration support. The most prominent features of LLMs are their natural language processing capabilities and their potential dialogic nature (if implemented alongside a chat interface). A core value of collaborative learning is the

interaction between humans, who interact to support each other and co-construct knowledge. Provided that it does not undermine human-human contact and interactions, generative AI may have its place in collaborative learning. For instance, it may facilitate various aspects of the collaboration (i.e. targets), such as social interaction, socio-cognitive processes, or group metacognition (see above). To this end, the support must elicit beneficial interaction patterns, including sharing information, giving explanations, and monitoring the collaboration (King, 2007[15]; Kreijns, Kirschner and Jochems, 2003[129]; Nokes-Malach et al., 2019[130]).

Here are a few potentially fruitful contributions of generative AI to already established types of collaboration support. First, an LLM may adapt the phrasing of an external collaboration script (scriptlets: Kollar, Wecker and Fischer, 2018[36]) to the competence of the individual group members, for example by using easier terminology or providing suggestions about how the prompt from the script may be performed. If the script is more fine-grained, for instance on the level of individual utterances or sentence-starters (scriptlets: Kollar, Wecker and Fischer, 2018[36]), an LLM may help formulate messages that are tailored to the context of the task and the group's conversation history. At the same time, there is the potential for groups to interact with and adapt the support during collaboration: if support is delivered by a chatbot, groups may ask clarifying questions, such as how to implement a particular collaboration strategy. Groups may also modify the degree of support, for instance by increasing or decreasing the amount or granularity of support, or by selecting specific aspects of their collaboration for which they wish (or no longer wish) to receive support. In this case, GenAI provides not only adaptive support, but also adaptable support, that is, the group can modify the support based on their needs during collaboration (see Rummel, 2018[28]; Vogel et al., 2022[131]; Wang, Kollar and Stegmann, 2017[132]).

When integrated into conversational agents (Burkhard et al., 2022[60]) or intelligent tutoring systems (Diziol et al., 2010[54]; Walker, Rummel and Koedinger, 2011[55]; Walker, Rummel and Koedinger, 2014[56]), an LLM may render the artificial agent's text output more human-like, allow for natural language input from the learners, provide prompts that are more context-aware, and take into account the history (i.e. context) of the conversation between the learner and the artificial agent, such as the goals that the group is currently pursuing. Accordingly, learners may engage with the support messages more actively. The same may apply to systems that do not take the role of a facilitator, but of a peer that has to be taught by the group and thus elicits beneficial interaction patterns (regarding teachable agents, see for example Brophy et al. (1999[133]) and Hayashi et al. (2025[134]).

The effects of such applications of GenAI should be subject to empirical inquiry. It is essential to note that outputs from LLMs are likely to suffer from encoded biases and stereotypes (Bender et al., 2021[127]); for examples in learning contexts, see Kotek et al., (2023[135]) or incorrect information that may even include neuromyths about learning and teaching (Richter et al., 2025[128]). Researchers, developers, and teachers must be sensitive to such issues and test a system's performance before implementing it on a broader scale.

Finally, generative AI, and especially LLMs, may be a valuable addition to the analytics pipeline of adaptive support systems due to their natural language processing capabilities, as illustrated by Wong et al. (2025[84]) and Zheng et al. (2021[79]). In this context, GenAI is not directly responsible for diagnosing the collaboration or deploying pedagogical actions but is part of the analysis of the interaction (e.g. dialogue in the group).

Designing support: evidence-based and ethical

So far, we have discussed various specific ways to leverage GenAI to support collaboration. Next, we highlight some more general aspects to consider when designing support that includes GenAI. Importantly, the design of adaptive collaboration support needs to centre on challenges that groups experience, as opposed to employing GenAI for its own sake. Thus, researchers and developers should design support that targets specific challenges that may arise during collaborative learning. Here, the discussion returns to our assessment of the current state of research on the effects of GenAI to facilitate collaborative learning. While the evidence regarding tangible beneficial effects is currently limited though positive, we can conceptualise collaboration support that might be effective. The mechanisms underlying effective collaboration and the challenges for groups are well documented (overviews of the central theories and empirical evidence are collected in Cress et al., (2021[136]) and Fischer et al. (2018[137]). Kollar et al. (2025[32]) and Trentepohl et al. (2025[138]) illustrated the importance of designing support carefully based on established theories and empirical evidence. Luckin and Cukurova (2019[139]) illustrate one way in which insights and methods from the learning sciences can guide the design of effective instructional support.

We have proposed conceptualising adaptive support as scaffolding. From this perspective, support does not remain available to the groups, but is rather faded out as learners become increasingly competent to perform all aspects of the task on their own (Reiser and Tabak, 2014[33]). The ultimate goal of scaffolding is to achieve learning that develops over time and leads to independent performance, as opposed to increasing the mere performance in the moment (and perhaps only as long as support is available).

More specific guidance in designing support can be found in the framework of CSCL design dimensions presented by Rummel (2018[28]) (see Figure 4.1), to which developers can refer in order to familiarise themselves with design decisions and potential options and subsequently utilise theoretical and empirical work to inform the specific design of the support. One perspective that may be fruitful when reflecting on these more fine-grained design decisions is the concept of cognitive offloading (Risko and Gilbert, 2016[140]). This perspective highlights the question of which (cognitive) activities are delegated (e.g. an object in the environment or an artificial agent) and thus not (yet) performed by human learners. While offloading or delegating certain tasks may lead to an increase in performance (Vaccaro, Almaatouq and Malone, 2024[141]), it should be kept in mind that specific activities represent opportunities for learners, and delegating them may thus have undesired effects (Fiorella and Mayer, 2016[142]) or desired difficulties (Bjork and Bjork, 2020[143]). From this perspective, there are good arguments for on-loading (and supporting) activities such as generating and collaboratively exploring different arguments and explanations or supporting other group members in monitoring and regulating their engagement during collaboration.

Finally, the design, effects, and consequences of supporting learning, especially when leveraging machine learning models, should be considered in the light of ethical, legal, and social implications (ELSI) or aspects (ELSA) (Nakazono, 2023[144]; Zwart, Landeweerd and van Rooij, 2014[145]). Technology is never neutral, and we therefore need to reflect not only upon our use of technology but also its design (ethics by design). Guidance for such reflection might be found in the field of applied ethics (Simis, 2024[146]), and especially from specific ethics such as the ethics of technology (Hansson, 2017[147]), ethics of artificial intelligence (Gunkel, 2024[148]), and ethics of artificial intelligence in education (e.g. Holmes and Porayska-Pomsta, 2023). A more practical approach for ethical reflection is offered by Simis (2024[146]; 2025[149]), who provides suggestions and leading questions. Given the current trends of exploring the use of LLMs in educational settings, we summarise some insights that might be included during ethical reflection: while there are undoubtedly potential benefits of utilising GenAI as part of educational technologies, we must be aware of potential costs (legal, social, or economic) associated with them. Bender (2024[150]) and Bender et al. (2021[127]) describe some of the costs linked to specific LLMs, including the question of which materials companies can legally use to train their models (Brittain, 2025[151]; Creamer, 2025[152]; O'Brien and Ortutaty, 2025[153]) or precarious working conditions of the data workers involved in the training and fine-tuning of models (e.g. Ali et al., 2024[154]; Hao, 2024[155]). Other costs are biases and stereotypes that are encoded in machine learning models, and the energy necessary to train models and process user queries (see Cornell University, November 3, 2022[156]; Crawford, 2024[157]), or other "hidden costs" (Greenbaum and Gerstein, 2025[158]). With this in mind, we have to discuss which costs we are willing to tolerate in exchange for benefits such as positive effects (of a particular magnitude) on learning outcomes and goals such as the UN Sustainable Development Goals (OECD, 2022[159]; United Nations, 2015[160]), for instance quality education, gender equality, or decent work and economic growth. Here, we have to keep in mind that the costs associated with LLMs can be expected to differ depending on the specifics of the development process (training) as well as our way of implementing them in educational contexts.

Conducting research on the effects of generative AI-based support for collaborative learning

As we have illustrated, more research on the effects of GenAI support is necessary to better inform the design of collaborative learning settings. Therefore, we sketch out relevant design aspects of empirical studies that are essential for creating robust and sufficiently reliable evidence to inform stakeholders' decisions about the design of effective collaboration environments.

Research topics

Our review revealed that greater research attention needs to be placed on roles of support, goals of support, and aspects of the collaboration that instructional support may address to achieve these goals. Thus, one avenue for future research is to explore these more thoroughly. Such research could be guided by the framework of CSCL design dimensions proposed by Rummel (2018[28]). Given the crucial role of collaboration processes for learning, research should also investigate processes that occur in the group, such as which roles the learners in the group take when interacting with each other and with an artificial agent (e.g. Gong et al., 2024[70]; Strijbos and Weinberger, 2010[161]), the quality of the interaction (Strauß et al., 2025[119]) the regulation processes (Hadwin and Oshige, 2011[23]; Järvelä and Hadwin, 2013[162]), or the processes of knowledge co-construction (Popov, van Leeuwen and Buis, 2017[163]). As noted by Cukurova (2025[164]), research should not be limited to cognitive outcomes. Collaborative learning offers opportunities to learn how to build relationships with peers, practice emotional regulation, build self-esteem, or acquire metacognitive capabilities.

However, future research might not only expand on potential targets and goals of collaboration support, or systematically investigate other CSCL design dimensions (Rummel, 2018[28]), but might also investigate the

consequences of introducing artificial agents into groups, such as altering the interaction in the group and thus how information is processed by human learners. For example, the review by Vaccaro et al. (2024[141]) highlights that the success of human-machine teams is not guaranteed. Instead, performance benefits of human-AI teams are only more likely if the human alone would outperform an AI alone.

From the perspective of cognitive offloading (Risko and Gilbert, 2016[140]), the introduction of artificial agents may lead human learners to delegate tasks to the artificial agent instead of performing these tasks themselves. This may undermine activities that are essential for learning, especially generative learning activities (for overviews see Brod, 2021[165]; Fiorella, 2023[166]; Fiorella and Mayer, 2016[142]) that pose desired difficulties for the learners (Bjork and Bjork, 2020[143]; Bruin et al., 2023[167]). The same applies to monitoring and controlling the collaboration. For the context of individual learning, Fan et al. (2025[168]) found that some learners delegated metacognitive processes to an artificial agent, a finding they termed "metacognitive laziness". At the same time, it is important to acknowledge that cognitive offloading is not necessarily always detrimental, and can be part of the fading-out process like Reiser and Tabak (2014[33]) (facilitative) co-regulation of collaboration (Rogat and Adams-Wiggins, 2015[169]).

Designing studies to gather empirical evidence

In pursuing avenues like those described above, we seek to gain insights that will allow us to determine how to design effective collaboration support for different educational contexts.

The basis for designing empirical studies is to test hypotheses on the expected mechanisms underlying how the support affects collaboration processes. Ideally, we can then develop experiments that isolate the effect of this support on outcomes such as knowledge acquisition. Conducting such studies in the context of collaborative learning comes with additional challenges, such as small sample sizes, randomisation, and statistical analyses (for a detailed discussion see Janssen and Kollar (2021[170]). Experiments on the effects of GenAI-based support are especially informative for educational practice if an experiment compares this newly designed support with a 'strong' control condition, such as other types of support (e.g. collaboration scripts, conversational agents). A comparison of outcomes from groups receiving adaptive support that utilises GenAI with the outcomes from groups receiving no support or receiving a 'business-as-usual' instruction may confound the effects of the support with other factors, such as the novelty factor, and may not be very informative for educational practice (Weidlich et al., 2025[113]). When comparing conditions, it is further vital to check whether the adaptive system indeed provides the intended support (i.e. implementation check). Otherwise, the results provide little insight into how to design and implement adaptive support. This is especially relevant for situations requiring insights about the conditions under which different types of support are effective, enabling us to select between alternatives.

Ideally, there will come a time when ample evidence from rigorous research is available that can be synthesised into overviews such as systematic literature reviews or meta-analyses. This appears to be a challenge pertinent for the current generation of research on the effects of generative AI on learning, as illustrated by Weidlich et al. (2025[113]). Thus, we need more research that yields robust findings.

We can tackle the complexity of developing and evaluating GenAI support for collaborative learning by bringing together expertise from fields such as the learning sciences, artificial intelligence, computer science, and educational practice. Such interdisciplinary collaboration can lead to more comprehensive research designs, more nuanced and robust data analysis, and consequently a deeper understanding of how a specific type of support affords collaborative learning. While such collaborations require more time and effort, we believe that the costs required to generate robust and reliable findings are justifiable given the consequences of premature conclusions (i.e. slow science (Frith, 2020[171]; Weidlich et al., 2025[113]) such as investing time to develop and implement ineffective support, or hampering learning. One framework for conducting interdisciplinary research that aims to exert an impact on practice in authentic learning contexts is design-based research (The Design-Based Research Collective, 2003[172]; Puntambekar, 2018[173]; Topali, Ortega-Arranz and Molenaar, 2026[174]).

Instead of 'moving fast and breaking things', we advocate that the costs for conducting research that produces reliable insights are well spent if harms are mitigated and we are able to have confidence that the support indeed benefits our learners. As researchers, educators, developers, and policymakers, it is our responsibility to shape a future where educational technology is used to afford meaningful collaboration and support effective interaction that serves learning. To this end, technologies need to be designed thoughtfully. As Schleicher (2018[175]) notes, "digital technologies are also creating opportunities that will amplify great teaching, even if great digital technology can never replace poor teaching" (p. 68). Therefore, considering insights from research on collaborative learning is the backbone of the design of support that benefits groups and not only fosters the acquisition of domain-specific knowledge, but also helps them become capable team members.

Acknowledgements

References

Ahlström, E., F. Sanctis, T. Nousiainen, and P. Näykki (2025), "Human-AI-human collaboration: Learning about climate change in virtual reality", in J. Oshima, B. Chen, F. Vogel and S. Järvelä, *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning - CSCL 2025*, International Society of the Learning Sciences. [97]

Alfieri, L., T. Nokes-Malach and C. Schunn (2013), "Learning Through Case Comparisons: A Meta-Analytic Review", *Educational Psychologist*, Vol. 48(2) pp. 87–113. [88]

Ali, A. M., A.H. Adan, A.O., Adan, A. A. Guyo and ... A. M. Ahmed (2024), "Open letter to President Biden from tech workers in Kenya", *Foxglove*, https://www.foxglove.org.uk/open-letter-to-president-biden-from-tech-workers-in-kenya/ (accessed on 25 November 2025). [154]

Amarasinghe, I., F. Marques, A. Ortiz-Beltrán and D. Hernández-Leo (2023), "Generative Pre-trained Transformers for Coding Text Data? An Analysis with Classroom Orchestration Data", in Viberg, O., I. Jivet, P. J. Muñoz-Merino, M. Perifanou and T. Papathoma, *Responsive and Sustainable Educational Futures*, Vol. 14200, pp. 32–43, Springer Nature Switzerland. Cham. [80]

Anderson, E., G. C. Lin, A. Farid, M. Fenech, B. Hanks, E. Klopfer, ... E. Doherty (2025), "Exploring GenAI Technologies within Collaborative Learning", *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning - CSCL 2025*, pp. 499-506, https://doi.org/10.22318/cscl2025.991659. [95]

An, S., Y. Li, Y., Y. Ma, G. Cheng and G. Xu (2024), *Developing an LLM-Empowered Agent to Enhance Student Collaborative Learning Through Group Discussion*, International Conference on Computers in Education. [103]

An, S., S. Zhang, T. Guo, S. Lu, W. Zhang and Z. Cai (2025), "Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment", *Computers* & Education, Vol. 227. [109]

Araujo, A., P.M. Papadopoulos, S. McKenney and T. Jong (2025), "Investigating the Impact of a Collaborative Conversational Agent on Dialogue Productivity and Knowledge Acquisition", *International Journal of Artificial Intelligence in Education*. [59]

Baker, M. T. Hansen, R. Joiner and D. Traum (1999), "The role of grounding in collaborative learning tasks", in Dillenbourg, P., *Advances in learning and instruction series. Collaborative learning. Cognitive and computational approaches*, pp. 31–63, Pergamon. [18]

Barron, B. (2003), "When Smart Groups Fail", *Journal of the Learning Sciences*, Vol. 12(3), pp. 307 -359. [4]

Bender, E. (2024), "Resisting Dehumanization in the Age of 'AI'", *Current Directions in Psychological Science,* Vol. 33(2), pp. 114–120. [150]

Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell (2021), *On the Dangers of Stochastic Parrots*, ACM. [127]

Bernacki, M., M. Greene and N. Lobczowski (2021), "A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)?", *Educational Psychology Review*, Vol. 33(4), pp. 1675–1715. [47]

Bjork, R. and E. Bjork (2020), "Desirable difficulties in theory and practice", *Journal of Applied Research in Memory and Cognition*, Vol. 9(4), pp. 475–479. [143]

Boudia, C. and A. Bengueddach (2024), "Innovating Higher Education: The Rise of ChatGPT in Collaborative Learning-A Short Review", *Proceedings of ICRES*, pp. 1347–1370. [71]

Brittain, B. (2025), "Meta fends off authors' US copyright lawsuit over AI", *Reuters*. [151]

Brod, G. (2021), "Generative Learning: Which Strategies for What Age?", *Educational Psychology Review,* Vol. 33(4), pp. 1295–1318. [165]

Brophy, S., G. Biswas and T. Katzlberger, J. Bransford and D. Schwartz (1999), "Teachable agents: Combining insights from learning theory and computer science", *Artificial Intelligence in Education*, Vol. 50, pp. 21–28.   [133]

Bruin, A. B., F. Biwer, L. Hui, E. Onan, L. David and W. Wiradhany (2023), "Worth the Effort: the Start and Stick to Desirable Difficulties (S2D2) Framework", *Educational Psychology Review*, Vol. 35(2).   [167]

Burkhard, M., S. Seufert, W. Cetto and S. Handschuh (2022), *Educational Chatbots for Collaborative Learning: Results of a Design Experiment in a Middle School*, International Association for Development of the Information Society.   [60]

Cai, Z., S. Park, N. Nixon and S. Doroudi (2024), *Advancing Knowledge Together: Integrating Large Language Model-based Conversational AI in Small Group Collaborative Learning*, ACM.   [90]

Chen, J., M. Wang, P.A. Kirschner and C-C.Tsai (2018), "The Role of Collaboration, Computer Use, Learning Environments, and Supporting Strategies in CSCL: A Meta-Analysis", *Review of Educational Research*, Vol. 88(6), pp. 799–843.   [29]

Chounta, I. and N. Avouris (2016), "Towards the real-time evaluation of collaborative activities: Integration of an automatic rater of collaboration quality in the classroom from the teacher's perspective", *Education and Information Technologies,* Vol. 21(4), pp. 815–835.   [65]

Chu, X., J. Xu and X. Zhai (2024), *Investigating the Knowledge Building Process of Collaborative Learning between Student and Virtual Tutor Supported by ChatGPT: A Discourse Analysis*, International Society of the Learning Sciences.   [78]

Clark, H. and S. Brennan (1991), "Grounding in communication", in L. B. Resnick, J. M. Levine and S. D. Teasley, *Perspectives on Socially Shared Cognition*, pp. 127–149, American Psychological Association, Washington.   [19]

Clark, R. (1994), "Media and method", *Educational Technology Research and Development*, Vol. 42(3), pp. 7–10.   [110]

Crawford, K. (2024), "Generative AI's Environmental Costs Are Soaring–and Mostly Secret", *Nature*, https://www.nature.com/articles/d41586-024-00478-x.   [157]

Creamer, E. (2025), "US authors' copyright lawsuits against OpenAI and Microsoft combined in New York with newspaper actions", *The Guardian*.   [152]

Cress, U. and J. Kimmerle (2023), "Co-constructing knowledge with generative AI tools: Reflections from a CSCL perspective", *International Journal of Computer-Supported Collaborative Learning*, Vol. 18(4), pp. 607–614.   [75]

Cress, U., C. Rosé, A. Wise and J. Oshima (eds.) (2021), *International Handbook of Computer-Supported Collaborative Learning*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-65291-3.   [136]

Cukurova, M. (2025), "The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence", *British Journal of Educational Technology*, Vol. 56(2), pp. 469–488.   [164]

dal Ponte, C., S. Dushyanthen and K. Lyons (2023), "Close…but not as good as an educator." - Using ChatGPT to provide formative feedback in large-class collaborative learning", *Proceedings of the 14th International Conference on Learning Analytics & Knowledge*, ACM   [106]

Dang, B., L. Huynh, F. Gul, C. Rosé, S. Järvelä and A. Nguyen (2025), "Human– AI collaborative learning in mixed reality: Examining the cognitive and socio-emotional interactions", *British Journal of Educational Technology*, pp. 1–24.   [98]

Darmawansah, D., D. Rachman, F. Febiyani and G.-J. Hwang (2025), "ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills", *Education and Information Technologies*, Vol. 30(3), pp. 3803–3827.   [85]

Deiglmayr, A. and H. Spada (2011), "Training for fostering knowledge co-construction from collaborative inference-drawing", *Learning and Instruction*, Vol. 21(3), pp. 441–451.   [42]

Deiglmayr, A. and H. Spada (2010), "Developing Adaptive Collaboration Support: The Example of an Effective Training for Collaborative Inferences", *Educational Psychology Review*, Vol. 22, pp. 103–113.   [41]

Dillenbourg, P. (2002), "Over-scripting CSCL: The risks of blending collaborative learning with instructional design", in P. A. Kirschner, *Three worlds of CSCL. Can we support CSCL?*, Heerlen, Open Universiteit Nederland, pp.61-91.   [39]

Dillenbourg, P. (1999), "What do you mean by 'collaborative learning'?", in P. Dillenbourg, *Cognitive and Computational Approaches*, pp. 1–19, Elsevier, Oxford.   [7]

Dillenbourg, P., M. Baker, A. Blaye and C. O'Malley (1995), "The evolution of research on collaborative learning", in H. Spada, & P. Reiman, *Learning in Humans and Machine: Towards an Interdisciplinary Learning Science*, pp. 189–211, Elsevier, Oxford.   [16]

Diziol, D., E. Walker, N. Rummel and K.R. Koedinger (2010), "Using Intelligent Tutor Technology to Implement Adaptive Support for Student Collaboration", *Educational Psychology Review*, Vol. 22(1), pp. 89–102.   [54]

Drachsler, H. and F. Goldhammer (2020), "Learning Analytics and eAssessment—Towards Computational Psychometrics by Combining Psychometrics with Learning Analytics", *Radical Solutions and Learning Analytics,* pp. 67–80, Springer Singapore.   [121]

Earle-Randell, T. V., S. Zhang, N. Schroeder, K.E. Boyer and E. Dorley (2025), "How Virtual Agents Can Shape Human-Human Collaboration: A Systematic Review", in Cristea, A. I., E. Walker, Y. Lu, O. C. Santos and S. Isotani, *Artificial Intelligence in Education*, Vol. 15879, pp. 468–486). Springer Nature Switzerland.   [58]

Akoyak-Yildiz, M., R. Alfredo, O. Chernikova, M. Cole, J. Cooke, C. Damşa, ... V. Echeverría (2025), *Enhancing Team-Based Medical Simulations: Learning Through Reflection with Analytics and AI Tools*, International Society of the Learning Sciences.   [94]

Edwards, J., A. Nguyen, J. Lämsä, M. Sobocinski, R. Whitehead, B. Dang, . . . Järvelä, S. (2025), "Human- AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners", *British Journal of Educational Technology*, Vol. 56(2), pp. 712–733.   [50]

Fan, Y., L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, . . . D. Gašević (2025), "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance", *British Journal of Educational Technology*, Vol. 56(2), pp. 489–530.   [168]

Feng, S. (2025), "Group interaction patterns in generative AI -supported collaborative problem solving: Network analysis of the interactions among students and a GAI chatbot", *British Journal of Educational Technology*.   [86]

Fiorella, L. (2023), "Making Sense of Generative Learning", *Educational Psychology Review*, Vol. 35(2).   [166]

Fiorella, L. and R. Mayer (2016), "Eight ways to promote generative learning", *Educational Psychology Review*, Vol. 28(4), pp. 717–741.   [142]

Fischer, F., C. Hmelo-Silver, S. Goldman and P. Reimann (2018), *International Handbook of the Learning Sciences*, Routledge, New York, NY, https://doi.org/10.4324/9781315617572.   [137]

Fischer, F., I. Kollar, K. Stegmann and C. Wecker (2013), "Toward a Script Theory of Guidance in Computer-Supported Collaborative Learning", *Educational Psychologist*, Vol. 48(1), pp. 56–66.   [35]

Fischer, F., I. Kollar, H. Mandl and J. Haake (eds.) (2007), *Scripting Computer-Supported Collaborative Learning*, Springer US, Boston, MA, https://doi.org/10.1007/978-0-387-36949-5.   [37]

Ko and Fultz (2025), *Exploring GenAI Technologies within Collaborative Learning*, in Anderson, E., G. C. Lin, A. Farid, M. Fenech, B. Hanks, E. Klopfer, . . . M. Scardamalia (Eds.). International Society of the Learning Sciences.   [93]

Frith, U. (2020), "Fast lane to slow science", *Trends in Cognitive Sciences*, Vol. 24(1), pp. 1–2.   [171]

Gašević, D., J. Jovanovic, A. Pardo and S. Dawson (2017), "Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance", *Journal of Learning Analytics*, Vol. 4(2), pp. 113–128.   [180]

Goda, Y., M. Arame, M. Toda, J. Handa and M. Yamada (2024), *Design and Development of Mondo-GPT: A Generative AI-Integrated Chatbot for Socratic Inquiry and Critical Thinking Enhancement*, Proceedings of Society for Information Technology & Teacher Education International Conference, pp. 2753–2760.   [99]

Gong, R., R. Jiang, C. Guo, W. Hu and Y. Li (2024), *Roles emerging during the knowledge construction process in collaborative learning: Does a generative AI-support chatbot matter?*, ACM.   [70]

Graesser, A. C., P.W. Foltz, Y. Rosen, D.W. Shaffer, C. Forsyth and M.-L. Germany (2018), "Challenges of Assessing Collaborative Problem Solving", in Care E., Griffin P. and M. Wilson (eds), *Assessment and Teaching of 21st Century Skills*, pp. 75–91, Springer International Publishing, Cham.   [179]

Greenbaum, D. and M. Gerstein (2025), "Hidden human costs of AI Waiting for Robots: The Hired Hands of Automation", *Science*, p. 32.   [158]

Greisel, M., J. Hornstein, I. Kollar, O. Noroozi, G. Haddadian, X. Gao, . . . N. Rummel (2025), "Enhancing Peer Feedback Practices with Generative AI", in J. Oshima, B. Chen, F. Vogel, & S. Järvelä, International Society of the Learning Sciences.   [96]

Güner, H. and E. Er (2025), "AI in the classroom: Exploring students' interaction with ChatGPT in programming learning", *Education and Information Technologies*, Vol. 30, pp. 12681–12707.   [77]

Gunkel, D. (ed.) (2024), *Handbook on the Ethics of Artificial Intelligence*, Edward Elgar Publishing.   [148]

Hadwin, A., A. Bakhtiar and M. Miller (2018), "Challenges in online collaboration: effects of scripting shared task perceptions", *International Journal of Computer-Supported Collaborative Learning*, Vol. 13(3), pp. 301–329.   [20]

Hadwin, A. and M. Oshige (2011), "Self-regulation, coregulation, and socially shared regulation: Exploring perspectives of social in self-regulated learning theory", *Teachers College Record*, Vol. 113(2), pp. 240–264.   [23]

Han, J., K.H. Kim, W. Rhee, W. and Y.H. Cho (2021), "Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation", *Computers & Education*, Vol. 163.   [66]

Hansson, S. (ed.) (2017), *The Ethics of Technology*, Rowman & Littlefield International.   [147]

Hao, K. (2024), "Cleaning Up ChatGPT Takes Heavy Toll on Human Workers", *Wall Street Journal* https://www.wsj.com/tech/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483#comments_sector (accessed on 25 November 2025).   [155]

Hattie, J. (2009), *Visible Learning*, Routledge.   [1]

Hayashi, Y., S. Shimojo and T. Kawamura (2025), "Experimental Investigation on Reciprocal Teaching Using Misconception-Based Teachable Agents in Collaborative Learning", *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning - CSCL 2025*, pp. 46-54, https://doi.org/10.22318/cscl2025.171809.   [134]

Hernández-Leo, D., E. Theophilou, J. Oshima, T. Minematsu, K. Matsueda, S. Naganuma, . . . A. Weinberger (2025), "Collaborating with Generative AI for Learning?", *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning - CSCL 2025*, International Society of the Learning Sciences.   [101]

Hornstein, J., M.V. Keller, M. Greisel, M. Dresel and I. Kollar (2025), "Enhancing the peer-feedback process through instructional Support: A meta-analysis", *Educational Psychology Review*, Vol. 37(2), pp. 1–34.   [107]

Hrastinski, S. (2008), "What is online learner participation? A literature review", *Computers & Education*, Vol. 51(4), pp. 1755–1765.   [21]

Huang, K., R. Ferreira Mello, C. Pereira Junior, L. Rodrigues, M. Baars and O. Viberg (2025), *That's What RoBERTa Said: Explainable Classification of Peer Feedback*, ACM.   [81]

Huisman, B., N. Saab, P. van den Broek and J. van Driel (2019), "The impact of formative peer feedback on higher education students' academic writing: a Meta-Analysis", *Assessment & Evaluation in Higher Education*, Vol. 44(6), pp. 863–880.   [108]

Janssen, J. and I. Kollar (2021), "Experimental and Quasi-Experimental Research in CSCL", in Cress, U., C. Rosé, A. F. Wise and J. Oshima, *International Handbook of Computer-Supported Collaborative Learning*, pp. 497–515, Springer International Publishing, Cham.   [170]

Järvelä, S. and A. Hadwin (2024), "Triggers for self-regulated learning: A conceptual framework for advancing multimodal research about SRL", *Learning and Individual Differences*, Vol. 115.   [24]

Järvelä, S. and A. Hadwin (2013), "New Frontiers: Regulating Learning in CSCL", *Educational Psychologist*, Vol. 48, pp. 25–39.   [162]

Järvelä, S. et al. (2021), "What multimodal data can tell us about the students' regulation of their learning process?", *Learning and Instruction*, Vol. 72, p. 101203, https://doi.org/10.1016/j.learninstruc.2019.04.004.   [125]

Jeong, H., C. Hmelo-Silver and K. Jo (2019), "Ten years of Computer-Supported Collaborative Learning: A meta-analysis of CSCL in STEM education during 2005–2014", *Educational Research Review*, Vol. 28.   [27]

Johnson, D. and R. Johnson (2009), "An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning", *Educational Researcher*, Vol. 38(5), pp. 365–379.   [17]

Jong, S., J. Wester, T.S. Schrills, K. Secher, C. F. Griggio and N. van Berkel (2024), *Assessing Cognitive and Social Awareness among Group Members in AI-assisted Collaboration*, ACM.   [74]

Kalyuga, S., R. Rikers and F. Paas (2012), "Educational Implications of Expertise Reversal Effects in Learning and Performance of Complex Cognitive and Sensorimotor Skills", *Educational Psychology Review*, Vol. 24(2), pp. 313–337.   [38]

Karakostas, A. and S. Demetriadis (2011), "Enhancing collaborative learning through dynamic forms of support: the impact of an adaptive domain-specific support strategy", *Journal of Computer Assisted Learning*, Vol. 27(3), pp. 243–258.   [51]

King, A. (2007), "Scripting Collaborative Learning Processes", in F. Fischer, I. Kollar, H. Mandl, & J. M. Haake, *Scripting Computer-Supported Collaborative Learning: Cognitive, Computational and Educational Perspectives*, pp. 13–37, Springer US, Boston, MA.   [15]

Klingsieck, K. (2018), Kurz und knapp – die Kurzskala des Fragebogens „Lernstrategien im Studium" (LIST) [Short and sweet—The short version of the questionnaire "Learning Strategies of University Students" (LIST)], *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology*, Vol. 32(4), pp. 249–259. https://doi.org/10.1024/1010-0652/a000230. [178]

Kollar, I., F. Fischer and F. Hesse (2006), "Collaboration scripts - a conceptual analysis", *Educational Psychology Review*, Vol. 18(2), pp. 159–185 [64]

Kollar, I., M. Greisel, T. Özbek, L. Spang and F. Vogel (2025), "Computer-Supported Collaborative Learning", in Gegenfurtner, A. and I. Kollar, *Designing Effective Digital Learning Environments*, pp. 179–192), Routledge Taylor & Francis Group, London; New York. [32]

Kollar, I., C. Wecker and F. Fischer (2018), "Scaffolding and Scripting (Computer-Supported) Collaborative just Learning", in F. Fischer, *International Handbook of the Learning Sciences*, pp. 340–350, Routledge, New York, NY. [36]

Kotek, H., R. Dockum and D. Sun (2023), *Gender Bias and Stereotypes in Large Language Models*, Association for Computing Machinery. [135]

Kozma, R. (1994), "Will media influence learning? Reframing the debate", *Educational Technology Research and Development*, Vol. 42(2) pp. 7–19. [111]

Kreijns, K., P. Kirschner and W. Jochems (2003), "Identifying the pitfalls for social interaction in computer-supported collaborative learning environments", *Computers in Human Behavior*, Vol. 19(3), pp. 335–353. [129]

Kreijns, K., J. Weidlich and P. Kirschner (2024), "Pitfalls of Social Interaction in Online Group Learning", *The Cambridge Handbook of Cyber Behavior,* pp. 588–630, Cambridge University Press. [9]

Kuhail, M. A., N. Alturki, S. Alramlawi and K. Alhejori (2023), "Interacting with educational chatbots", *Education and Information Technologies*, Vol. 28(1), pp. 973–1018. [62]

Lai, J. W., W. Qiu, M. Thway, L. Zhang, N.B. Jamil, C.L. Su, . . . F.S. Lim (2024), "Leveraging Process-Action Epistemic Network Analysis to Illuminate Student Self-Regulated Learning with a Socratic Chatbot", *Journal of Learning Analytics*, pp. 1–19. [72]

Lin, C.-J., H.-Y. Lee, W.-S. Wang, Y.-M. Huang and T.-T. Wu (2024), "Enhancing reflective thinking in STEM education through experiential learning: The role of generative AI as a learning aid", *Education and Information Technologies,* Vol. 30. [105]

Liu, J., Y. Yao, P. An and Q. Wang (2024), *PeerGPT: Probing the Roles of LLM-based Peer Agents as Team Moderators and Participants in Children's Collaborative Learning*, ACM. [91]

Lou, Y., P. Abrami and S. d'Apollonia (2001), "Small Group and Individual Learning with Technology: A Meta-Analysis", *Review of Educational Research*, Vol. 71(3), pp. 449–521. [30]

Luckin, R. and M. Cukurova (2019), "Designing educational technologies in the age of AI: A learning sciences-driven approach", *British Journal of Educational Technology*, Vol. 50(6), pp. 2824–2838. [139]

Matschke, C., J. Moskaliuk, F. Bokhorst, T. Schümmer and U. Cress (2014), "Motivational factors of information exchange in social information spaces", *Computers in Human Behavior,* Vol. 36, pp. 549–558. [22]

Molenaar, I. (2022), "Towards hybrid human-AI learning technologies", *European Journal of Education*, Vol. 57(4), pp. 632–645. [43]

Molenaar, I. (2021), "Personalisation Of Learning: Towards Hybrid Human-AI Learning Technologies", *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, OECD Publishing, Paris, https://doi.org/10.1787/589b283f-en. [46]

Naik, A., J.R. Yin, A. Kamath, Q. Ma, S.T. Wu, C. Murray, . . . C.P. Rose (2024), "Generating Situated Reflection Triggers About Alternative Solution Paths: A Case Study of Generative AI for Computer-Supported Collaborative Learning", *Artificial Intelligence in Education*, Vol. 14829, pp. 46–59, Springer Nature Switzerland, Cham. [115]

Naik, A., J.R. Yin, A. Kamath, Q. Ma, S.T. Wu, C. Murray, . . . C.P. Rose (2024), "Generating Situated Reflection Triggers About Alternative Solution Paths: A Case Study of Generative AI for Computer-Supported Collaborative Learning", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64302-6_4. [116]

Naik, A., Yin, J., Kamath, A., Ma, Q., Wu, S., Murray, R., . . . Rose, C (2024), "Providing tailored reflection instructions in collaborative learning using large language models", *British Journal of Educational Technology*, Vol. 56(2), pp. 531-550, https://doi.org/10.1111/bjet.13548. [87]

Naik, A. et al. (2024), "Providing tailored reflection instructions in collaborative learning using large language models", *British Journal of Educational Technology*, Vol. 56/2, pp. 531-550, https://doi.org/10.1111/bjet.13548.   [117]

Nakazono, N. (2023), "ELSI (Ethical, Legal, and Social Issues) Education on Digital Technologies: In the Field of Elementary and Secondary Education",in Keane, T., C. Lewin, T. Brinda and R. Bottino, *Towards a Collaborative Society through Creative Learning*, 1st ed. 2023 ed., Vol. 685, pp. 57–68, Springer Nature Switzerland;Imprint Springer, Cham.   [144]

Nokes-Malach, T. J., C.D. Zepeda, J.E. Richey and S. Gadgil (2019), "Collaborative Learning", in Dunlosky, J., *The Cambridge Handbook of Cognition and Education*, 1st ed. ed., pp. 500–527, Cambridge University Press, Cambridge.   [25]

Nokes-Malach, T. et al. (2019), "Collaborative Learning", in *The Cambridge Handbook of Cognition and Education*, Cambridge University Press, https://doi.org/10.1017/9781108235631.021.   [130]

O'Brien, M. and B. Ortutaty (2025), "Judge tosses authors' AI training copyright lawsuit against Meta", *PBS News*, https://www.pbs.org/newshour/arts/judge-tosses-authors-ai-training-copyright-lawsuit-against-meta (accessed on 31 August 2025).   [153]

OECD (2022), *The Short and Winding Road to 2030: Measuring Distance to the SDG Targets*, OECD Publishing, Paris, https://doi.org/10.1787/af4b630d-en.   [159]

Pai, H., D. Sears and Y. Maeda (2015), "Effects of Small-Group Learning on Transfer: a Meta-Analysis", *Educational Psychology Review*, Vol. 27(1), pp. 79–102.   [10]

Pai, H., D. Sears and Y. Maeda (2014), "Effects of Small-Group Learning on Transfer: a Meta-Analysis", *Educational Psychology Review,* Vol. 27/1, pp. 79-102, https://doi.org/10.1007/s10648-014-9260-8.   [2]

Popov, V., A. van Leeuwen and S. Buis (2017), "Are you with me or not? Temporal synchronicity and transactivity during CSCL", *Journal of Computer Assisted Learning*, Vol.33(5), pp. 424–442.   [163]

Puntambekar, S. (2018), *Design-Based Research (DBR),* Routledge.   [173]

Radkowitsch, A., M. Sailer, R. Schmidmaier, M.R. Fischer and F. Fischer, F. (2021), "Learning to diagnose collaboratively – Effects of adaptive collaboration scripts in agent-based medical simulations", *Learning and Instruction*, Vol. 75.   [48]

Radkowitsch, A., F. Vogel and F. Fischer (2020), "Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts", *International Journal of Computer-Supported Collaborative Learning*, Vol. 15(1), pp. 5–47.   [11]

Reiser, B. and I. Tabak (2014), "Scaffolding", in R. K. Sawyer, *The Cambridge Handbook of the Learning Sciences*, pp. 44–62, Cambridge University Press.   [33]

Richter, E., M.W. Spitzer, A. Morgan, L. Frede, J. Weidlich and K. Moeller (2025), "Large language models outperform humans in identifying neuromyths but show sycophantic behavior in applied contexts", *Trends in Neuroscience and Education,* Vol. 39.   [128]

Risko, E. and S. Gilbert (2016), "Cognitive Offloading", *Trends in Cognitive Sciences*, Vol. 20(9), pp. 676–688.   [140]

Rogat, T. and K. Adams-Wiggins (2015), "Interrelation between regulatory and socioemotional processes within collaborative groups characterized by facilitative and directive other-regulation", *Computers in Human Behavior*, Vol. 52, pp. 589–600.   [169]

Roschelle, J. and S. Teasley (1995), *The Construction of Shared Knowledge in Collaborative Problem Solving,* Springer.   [8]

Rosé, C., Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger and F. Fischer (2008), "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning", *International Journal of Computer-Supported Collaborative Learning*, Vol. 3(3), pp. 237–271.   [45]

Rummel, N. (2018), "One framework to rule them all? Carrying forward the conversation started by Wise and Schwarz", *International Journal of Computer-Supported Collaborative Learning*, Vol. 13(1), pp. 123–129.   [28]

Rummel, N., D. Mullins and H. Spada (2012), "Scripted collaborative learning with the cognitive tutor algebra", *International Journal of Computer-Supported Collaborative Learning*, Vol. 7(2), pp. 307–339.   [52]

Rummel, N., H. Spada and S. Hauser (2009), "Learning to collaborate while being scripted or by observing a model", *International Journal of Computer-Supported Collaborative Learning*, Vol. 4(1), pp. 69–92.   [5]

Rummel, N., A. Weinberger, C. Wecker, F. Fischer, A. Meier, E. Voyiatzaki, . . . M. Joshi (2008), "New challenges in CSCL: Towards adaptive script support", in Kanselaar, G., V. Jonker, P. A. Kirschner and F. J. Prins, *International Perspectives in the Learning Sciences: Creating a learning world. Proceedings of the Eighth International Conference for the Learning Sciences – ICLS 2008*, pp. 338–345, The Netherlands, Utrecht.   [53]

Sailer, M., R. Maier, S. Berger, T. Kastorff, T. and K. Stegmann (2024), "Learning activities in technology-enhanced learning: A systematic review of meta-analyses and second-order meta-analysis in higher education", *Learning and Individual Differences*, Vol. 112.   [114]

Schleicher, A. (2018), "Educating Learners for Their Future, Not Our Past", *ECNU Review of Education*, Vol. 1(1), pp. 58–75.   [175]

Schneider, B., G. Sung, E. Chng and S. Yang, S. (2021), "How Can High-Frequency Sensors Capture Collaboration? A Review of the Empirical Links between Multimodal Metrics and Collaborative Constructs", *Sensors*, Vol. 21/24, p. 8185, https://doi.org/10.3390/s21248185.   [126]

Schürmann, V., T. Spahn, N. Marquardt and D. Bodemer (2023), *Evaluating Peer Collaboration in Higher Education: Behaviorally Anchored Rating Scales*, International Society of the Learning Sciences.   [118]

Simis, C. (2025), *Digitale Technologien an Hochschulen : Ethische Orientierung für die Praxis*, Ruhr-Universität Bochum, https://doi.org/10.13154/294-13250.   [149]

Simis, C. (2024), D*igitale Technologien an Hochschulen : Ethische Orientierung für die Praxis*, Ruhr-Universität Bochum, https://doi.org/10.13154/294-12405.   [146]

Song, Y., L. Huang, L. Zheng, M. Fan and Z. Liu (2025), "Interactions with generative AI chatbots: unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving", *International Journal of Educational Technology in Higher Education*, Vol. 22(1).   [100]

Sottilare, R. A., C. Shawn Burke, E. Salas, A.M. Sinatra, J.H. Johnston and S.B. Gilbert (2018), "Designing Adaptive Instruction for Teams: a Meta-Analysis", *International Journal of Artificial Intelligence in Education*, Vol. 28(2), pp. 225–264.   [123]

Sottilare, R., C. Shawn Burke, E. Salas, A. Sinatra, Johnston, J., & Gilbert, S. (2017), "Designing Adaptive Instruction for Teams: a Meta-Analysis", *International Journal of Artificial Intelligence in Education*, Vol. 28/2, pp. 225-264, https://doi.org/10.1007/s40593-017-0146-z.   [57]

Spörer, N. and J. Brunstein (2006), "Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren", Z*eitschrift für Pädagogische Psychologie*, Vol. 20(3), pp. 147–160.   [177]

Stadler, M., K. Herborn, M. Mustafić and S. Greiff (2019), "Computer-Based Collaborative Problem Solving in PISA 2015 and the Role of Personality", *Journal of Intelligence*, Vol. 7(3), p. 15.   [176]

Stahl, G., T. Koschmann and D. Suthers (2006), "Computer-supported collaborative learning: An historical perspective, An historical perspective", in R. K. Sawyer, *Cambridge Handbook of the Learning Sciences*, pp. 409–426, Cambridge University Press.   [13]

Statista (ed.) (2022), "Power consumption when training artificial intelligence (AI) based large language models (LLMs) in 2023 (in megawatt hours)", *Statista*, https://www.statista.com/statistics/1384401/energy-use-when-training-llm-models/ (accessed on Sept. 01, 2025).   [156]

Strauß, S. and N. Rummel (2021), "Problematic interaction patterns during online-collaboration. A library and a survey", in C. E. Hmelo-Silver, B. de Wever, & J. Oshima, *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning - CSCL 2021*, Bochum.   [26]

Strauß, S. and N. Rummel (2021), "Promoting regulation of equal participation in online collaboration by combining a group awareness tool and adaptive prompts. But does it even matter?", *International Journal of Computer-Supported Collaborative Learning*, Vol. 16, pp. 67–104.   [6]

Strauß, S., I. Tunnigkeit, J. Eberle, A. Avdullahu, A. and N. Rummel (2025), "Comparing the effects of a collaboration script and collaborative reflection on promoting knowledge about good collaboration and effective interaction", *International Journal of Computer-Supported Collaborative Learning*, Vol. 20, pp. 121–159.   [119]

Strijbos, J. and A. Weinberger (2010), "Emerging and scripted roles in computer-supported collaborative learning", *Computers in Human Behavior*, Vol. 26(4), pp. 491–494.   [161]

Suciati, S., L.M. Silitonga, C.-Y. Wiyaka, Huang and A.A. Anggara (2024), *Enhancing Engagement and Motivation in English Writing Through AI: The Impact of ChatGPT-Supported Collaborative Learning*, Springer Nature Switzerland.   [76]

Talan, T. (2021), "The Effect of Computer-Supported Collaborative Learning on Academic Achievement: A Meta-Analysis Study", *International Journal of Education in Mathematics, Science and Technology*, Vol. 9(3), pp. 426–448.   [31]

Tenenbaum, H. R., N.E. Winstone, P.J. Leman and R.E. (2020), "How effective is peer interaction in facilitating learning? A meta-analysis", *Journal of Educational Psychology*, Vol. 112(7), pp. 1303–1319.   [3]

Tennyson, R. (1994), "The big wrench vs. integrated approaches: The great media debate", *Educational Technology Research and Development*, Vol. 42(3), pp. 15–28.   [112]

Tetzlaff, L., F. Schmiedek and G. Brod (2021), "Developing Personalized Education: A Dynamic Framework", *Educational Psychology Review*, Vol. 33(3), pp. 863–882.   [34]

The Design-Based Research Collective (2003), "Design-Based Research: An Emerging Paradigm for Educational Inquiry", *Educational Researcher*, Vol. 32(1), pp. 5–8.   [172]

Topali, P., A. Ortega-Arranz and I. Molenaar (2026), "Transitioning from general-purpose to educational-oriented GenAI: maintaining teacher autonomy", in *OECD, Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [174]

Trentepohl, S. (2025), *Prepare - Perform - Reflect - Repeat, An Intervention Perspective on Self-Regulated Learning in Higher Education*, Ruhr-Universität Bochum, Bochum.   [138]

Tuckman, B. (1965), "Developmental sequences in small groups", *Psychological Bulletin*, Vol. 63, pp. 384–399.   [92]

United Nations (2015), *Transforming Our World: The 2030 Agenda for Sustainable Development*, United Nations, New York.   [160]

Vaccaro, M., A. Almaatouq and T. Malone (2024), "When combinations of humans and AI are useful: A systematic review and meta-analysis", *Nature Human Behaviour*, Vol. 8(12), pp. 2293–2303.   [141]

van Leeuwen, A., S. Teasley and A. Wise (2022), "Teacher and Student Facing Learning Analytics", in Lang, C., G. Siemens, A. F. Wise, D. Gašević and A. Merceron, *The Handbook of Learning Analytics*, pp. 130–140, SOLAR.   [67]

Vizcaíno, A. (2005), "A simulated student can improve collaborative learning", *Artificial Intelligence in Education*, Vol. 15(1), pp. 3–40.   [61]

Vogel, F., I. Kollar, F. Fischer, K. Reiss, K. and S. Ufer (2022), "Adaptable scaffolding of mathematical argumentation skills: The role of self-regulation when scaffolded with CSCL scripts and heuristic worked examples", *International Journal of Computer-Supported Collaborative Learning*, Vol. 17(1), pp. 39–64.   [131]

Vogel, F., C. Wecker, I. Kollar and F. Fischer (2017), "Socio-Cognitive Scaffolding with Computer-Supported Collaboration Scripts: a Meta-Analysis", *Educational Psychology Review*, Vol. 29(3), pp. 477–511.   [12]

Walker, E., N. Rummel and K. Koedinger (2014), "Adaptive Intelligent Support to Improve Peer Tutoring in Algebra", *International Journal of Artificial Intelligence in Education*, Vol. 24(1), pp. 33–61.   [56]

Walker, E., N. Rummel and K. Koedinger (2011), "Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity", *International Journal of Computer-Supported Collaborative Learning*, Vol. 6(2), pp. 279–306.   [55]

Walker, E., N. Rummel and K. Koedinger (2009), "A research framework for providing adaptive collaborative learning support", *User Modeling and User-Adapted Interaction*, Vol. 19(5), pp. 387–431.   [44]

Wang, D., D. Shan, Y. Zheng, G. Chen and Y. Lu (2023), "Can ChatGPT Detect Student Talk Moves in Classroom Discourse? A Preliminary Comparison with Bert", *Proceedings of the 16th International Conference on Educational Data Mining,* pp. 515–519.   [83]

Wang, X., I. Kollar and K. Stegmann (2017), "Adaptable scripting to foster regulation processes and skills in computer-supported collaborative learning", *International Journal of Computer-Supported Collaborative Learning*, Vol. 12(2), pp. 153–172.   [132]

Webb, N. (2013), "Information processing approaches to collaborative learning", in C. E. Hmelo-Silver, C. A. Chinn, C. K. Chan, & A. M. O'Donnell, *The International Handbook of Collaborative Learning*, pp. 19–40, Taylor and Francis, Hoboken.   [14]

Wecker, C. and F. Fischer (2011), "From guided to self-regulated performance of domain-general skills: The role of peer monitoring during the fading of instructional scripts", *Learning and Instruction*, Vol. 21(6), pp. 746–756.   [49]

Weidlich, J., D. Gašević, H. Drachsler and P.A. Kirschner (2025), "ChatGPT in Education: An Effect in Search of a Cause", *Journal of Computer Assisted Learning*, Vol. 41(5).   [113]

Weinberger, A. and F. Fischer (2006), "A framework to analyze argumentative knowledge construction in computer-supported collaborative learning", *Computers & Education*, Vol. 46(1), pp. 71–95.   [40]

Wei, R., K. Li and J. Lan (2024), *Improving Collaborative Learning Performance Based on LLM Virtual Assistant,* IEEE.   [104]

Wei, X., L. Wang, L.-K. Lee and R. Liu (2025), "The effects of generative AI on collaborative problem-solving and team creativity performance in digital story creation: an experimental study", *International Journal of Educational Technology in Higher Education*, Vol. 22(1).   [89]

Wise, A., S. Knight and S. Shum (2021), *Collaborative Learning Analytics*, Springer International Publishing.   [69]

Wise, A., C. Rosé and S. Järvelä (2023), "Editorial: Nine elements for robust collaborative learning analytics: A constructive collaborative critique", *International Journal of Computer-Supported Collaborative Learning*, Vol. 18(1), pp. 1–9.   [122]

Wollny, S., J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger and H. Drachsler (2021), "Are We There Yet? - A Systematic Literature Review on Chatbots in Education", *Frontiers in Artificial Intelligence*, Vol. 4.   [63]

Wong, K., B. Wu, S. Bulathwela and M. Cukurova (2025), "Rethinking the Potential of Multimodality in Collaborative Problem Solving Diagnosis with Large Language Models", in Cristea, A. I., E. Walker, Y. Lu, O. C. Santos and S. Isotani, *Artificial Intelligence in Education*, Vol. 15878, pp. 18–32, Springer Nature Switzerland, Cham.   [84]

Wu, Z., S. Jiaxin, R. Charles Murray, C. Rosé, C. and M. San Andres (2025), *LLM Bazaar: A Service Design for Supporting Collaborative Learning with an LLM-Powered Multi-Party Collaboration Infrastructure*, International Society of the Learning Sciences.   [120]

Yang, K. B., Z. Lu, V. Echeverria, J. Sewall, L. Lawrence, N. Rummel and V. Aleven (2022), *Technology Ecosystem for Orchestrating Dynamic Transitions Between Individual and Collaborative AI-Tutored Problem Solving*, Springer International Publishing;Imprint Springer.   [68]

Yan, L., S. Greiff, Z. Teuber and D. Gašević (2024), "Promises and challenges of generative artificial intelligence for human learning", *Nature Human Behaviour*, Vol. 8(10), pp. 1839–1850.   [73]

Yin, S. X., Z. Liu, D. H.-L. Goh, C.L. Quek and N.F. Chen (2025), *Scaling Up Collaborative Dialogue Analysis: An AI-driven Approach to Understanding Dialogue Patterns in Computational Thinking Education*, ACM.   [82]

Zheng, L., L. Zhong, J. Niu, M. Long and J. Zhao (2021), "Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning", *Educational Technology & Society*, Vol. 24(3), pp. 174–193.   [79]

Zhou, Q., W. Suraworachet and M. Cukurova (2024), "Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions", *Education and Information Technologies*, Vol. 29(1), pp. 1071–1098.   [124]

Zhu, G., X. Fan, C. Hou, T. Zhong, P. Seow, A.C. Shen-Hsing, . . . T.L. Poh (2023), *Embrace Opportunities and Face Challenges: Using ChatGPT in Undergraduate Students' Collaborative Interdisciplinary Learning*.   [102]

Zwart, H., L. Landeweerd and A. van Rooij (2014), "Adapt or perish? Assessing the recent shift in the European research funding arena from 'ELSA' to 'RRI'", *Life Sciences, Society and Policy*, Vol. 10(1), p. 11.   [145]

5

Developing creativity with generative AI:
A conversation with Ronald Beghetto

This chapter is an interview between Ronald Beghetto, Professor at Arizona State University (United States), and the OECD Secretariat. After defining creativity, Beghetto presents his approach of building AI tools to experience creativity as well as the tools he developed. He argues for a slow use of generative AI in which teachers, students (and humans more broadly) remain in charge of their ideas and use generative AI to achieve a personal goal.

What creativity entails

OECD: You are an expert on creativity and how to foster it in education. What do you see as the main principles?

Ron Beghetto: The way I see creativity is very simple: creativity is a potential we all have, but not something we possess. We possess the capacity and potential to do something creative, but whether this is the case is usually judged after the fact. We never know in advance whether the process or outcome will be creative. The definition generally used in the field is that creativity requires something to be both new and meaningful or useful. It is not just originality, but originality constrained by criteria, objectives, and meaning. Generating a lot of wild solutions is just meaningless originality. Creativity must also address or solve a problem or task. For example, if you are a cook and you combine ingredients in a completely novel way but the dish is inedible, that is not creative. It has to be tasty, edible, and appealing. Creativity is a blend of originality and appropriateness, personally meaningful or meaningful to your audience.

In education, the great advantage is that we are very good at specifying criteria and constraints. We just have to open up the process so that people can meet those objectives in different and unexpected ways. That introduces uncertainty. Structured uncertainty is key. If everything is predetermined – what the problem is, how to solve it, and what the answer looks like – then we have engineered creativity out of education. But if you provide structure by saying, "this is what we want, but how you do it is up to you," that creates space for creativity.

On the teaching side, part of fostering creativity is helping educators become comfortable with the uncertainty of not knowing how students will reach objectives. You need to be clear about the criteria and then let students find their own paths. Core principles are: 1) be comfortable with uncertainty; 2) provide necessary structure and support without predetermining everything; 3) balance predetermined criteria and openness; and 4) recognise that domain knowledge is essential. Students who are creative in dance or music may not be in science, and vice versa. They must have knowledge and experience in a domain to produce something new and appropriate.

OECD: When OpenAI released ChatGPT, you quickly designed some GenAI tools to support different aspects of the creative process. Could you tell us about it?

For me personally, when the "ChatGPT moment" happened, I was able to get research access via an API key, so I could build my own tools powered by GPT models as early as 2022. My first thought was: this is pretty interesting… There was this little playground area, where you could test ideas and then build something. I had been working for a while with educators on protocols to support possibility thinking, usually in a human-to-human context. I wondered whether this tool could be trained to serve as a digital facilitator, especially if you do not have partners for possibility thinking. The problem was that I did not know how to code in Python. I had learned BASIC, the programming language, a long time ago, but that was about it. So, I spent a weekend working with ChatGPT itself, just asking it to teach me how to build a Python app, which it did. Remarkably, I had a functional app within a day or two, something that would have taken me years if I had been trying to learn from scratch via YouTube videos. Because I had a very specific goal and some domain knowledge, I knew exactly what I wanted for my bot: not just to provide answers, but to interact with users in a more Socratic way.

That experience was pretty amazing. I quickly started using ideas and knowledge from my work and the field to build standalone tools that could be free to use. That was a big realisation for me: I was building something very different from how I saw most people using ChatGPT at the time. The interface looks like a search engine, so it almost predisposes people to type in a question and get an answer. These models are designed to do that. This, I think, sets people on two divergent pathways. One is where the tool becomes a rich partner in possibility thinking, something that augments and can be steered in ways anchored in good principles for supporting creative thinking. This is what Vlad Glăveanu and I call a "slow AI experience," where the system always asks for more context, because context engineering is far more effective than prompt engineering alone. The second path is "fast AI", with people using it in a one-off way, typing in a question and running with the first polished response they get. Early on, I noticed (and I am increasingly convinced) that education is at a critical inflexion point between these two possible futures.

Fast versus Slow AI uses

OECD: Tell us more about those two paths. What have you observed in your research and teaching?

Ron Beghetto: Let's start with the second path, that of "fast AI". To me, this path leads to overdependence on AI, where students and teachers essentially become digital puppets. There is actually some empirical evidence starting to show this, especially with students, but I think it is happening with teachers as well. You can imagine a student who has an assignment deadline looming, they have a few ideas for an essay, but just before the deadline, they paste in the instructions and a few thoughts, and have ChatGPT or another tool produce the essay for them. Maybe they tweak it, maybe they don't. There are reports that some students use AI-generated content without any modification. For instance, Anthropic's Claude released a usage report looking at a million users with EDU emails – presumably mostly students, but probably some faculty as well. They found that nearly half were using it in this direct-response way: asking questions and receiving answers. Some were even explicitly requesting the AI to produce text that would not be detected by plagiarism tools.

But I think educators are also becoming digital puppets. For example, an educator with 160 papers to grade might think, "I'll just see what ChatGPT can do. Here are my criteria; here's the feedback I usually give." And soon you end up in this absurd, detrimental space where AI is speaking through students to another AI speaking through teachers. Just sitting with that idea is rather grim and dystopian. Yet this is happening, at least part of the time.

The other approach – "slow AI", the one I advocate for – is helping educators and students learn to work creatively and responsibly with AI to become more dynamic thinkers. It is about using AI as a partner in possibility thinking as if it were just a new perspective, like turning to a colleague. In that way, it is fine if it is not completely accurate, because you should never trust any single source uncritically. You should check different perspectives. That, I believe, can be really powerful. But it requires slowing things down. You must start with your own thinking, then, just as you would with a colleague, get some feedback, bring it back to yourself or your team, and work through it. This is the difference between having AI do the work or the creative thinking for us, and working with it to augment our thinking.

OECD: From your own experience, how would you encourage teachers and learners in exploring the slower path?

Ron Beghetto: What I have increasingly realised is that educators and students need to learn to build with generative AI, just as I did. I think that is the most effective way.

There is a lot of rhetoric about AI literacy, which is fine, but it tends to be superficial. "Use it ethically, beware of bias", and so on. All true, but you do not really understand it unless you try to build something yourself. There is a "vibe coding moment" emerging, enabling people to start building tools. But you need a clear goal, prior content knowledge, and a sense of what you want to build.

In autumn 2024, I started a course with doctoral students, who therefore had some domain expertise. We began with: "What kind of AI assistant could you build to support your professional goals?" I taught them the process of using these tools to build something for their work, or for other educators or students. I call it the "build to learn, learn to create" approach. You build first, and then you start to see the strengths and limitations of your product. It was remarkable what this group produced – most students had never built any AI tool before, maybe one had tinkered a little bit, but nothing more. But because they had clear goals and knew what they wanted to achieve, they built tools that they are now using in their dissertations or professional practice.

Then I thought, why not open this up to undergraduates and teachers? So, since autumn 2025, I have been teaching two courses: one for undergraduates of all majors and one for graduate students. I have also been running workshops for teachers, showing how this approach can be used in a more principled way – a slower AI approach where you teach the AI to respond in a Socratic way. Almost obnoxiously Socratic, in fact: always asking questions, seeking context, supporting the maintenance of human ideas and agency – never simply giving direct answers, but suggesting possibilities: "What if you tried this?" or "What if you tried that?" Keeping ownership of ideas with the human.

Principles and tools to foster creativity with generative AI

OECD: How can teachers make the most of generative AI to foster creativity – especially when they are usually averse to uncertainty? And are the principles different for students?

Ron Beghetto: I think the principles are essentially the same for teachers and students. We have primarily been working with teachers, because their role is critically important, particularly when working with younger students. Many of these tools have minimum age requirements in their terms of service. You should not simply turn students loose with them. Teachers need to be part of the process, to be in the loop.

First, teachers have to be comfortable with the uncertainty of not knowing exactly how to use these tools. Many teachers have been experimenting, but many still do not see themselves as creative. Many people in general, including teachers, tend to think that kids are more creative than adults. That belief is problematic. They think kids are freer and play more. But again, they are conflating creativity with pure originality. Yes, young people often come up with all sorts of wild ideas. As you grow older, you learn the constraints and realities of the world. But, again, creativity is *constrained* originality: it must be appropriate for the task and grounded in knowledge.

Teachers are actually well-positioned to guide that, but they need to understand creativity properly and be clear about why they are using generative AI. So teachers must have clear purpose and goals, use their own experience and domain knowledge, and be open to uncertainty and different perspectives.

Let's take practical examples. Sometimes, you have a lesson you have taught for years, and it does not work very well. You want to change it and make it more creative, but you are too close to it, too familiar. A simple heuristic is to make the familiar unfamiliar. You are playing with the tensions between structure and uncertainty, familiarity and unfamiliarity. Because generative AI tools are dialogic (they can have meaningful conversations with you), you can say: "I don't know how to do this; here is what I am thinking." But you still maintain control: "These are my goals; this is my context." If teachers are not willing to build tools themselves, they at least need to learn how to interact with AI in a way that slows the process down. That means having clear goals, pushing back, just as you would with a colleague, and providing detailed context. For example: "This is what I want to do; here are my materials; here is how I expect the interaction to happen." That is an aspect of context engineering, moving beyond prompt engineering. And you can say

to the AI chatbot: "Share possibilities, not answers. Preface them with 'what if' so I remember this is just one perspective." I think this is where it starts: teachers modelling this careful, reflective use.

Second, I think teachers need sustained experience with these tools before introducing them to students. In my courses, I demonstrate examples of the tools I have built, but I tell students: "Don't build these same tools; build something that addresses a problem or need that you identify."

For students, making the best, or most creative, use of student-facing applications relies on similar precautions. Most students are already using AI, often as a kind of companion, including for social and emotional support. It can be persuasive, sometimes too persuasive. For example, a student might think: "I like writing poetry, but this thing writes better poetry than I ever could. I'll just have it do it for me." We do not want that. Or: "This advice sounds very reasonable." But you must remain critical. This is just one voice. Get other perspectives, including from humans you trust. So, again, the following principles apply: embrace uncertainty, ground your work in knowledge and clear goals, be open to different perspectives, and constrain the process so outputs are relevant and feasible.

OECD: Tell us a bit about the different tools you have built with generative AI.

Ron Beghetto: On my website, readers can find short videos showcasing a few examples of the bots I've built with generative AI. I even had AI narrate the videos, along with my own narration.

One tool I developed is for the *AI Possibility Lab*. It is an ecosystem of tools I use in my classes and beyond, with students, teachers, and educational leaders. All my AI-solutions are built around a simple pedagogical framework: first, prioritise human-to-human dialogue, to clarify why you even want to use AI. And second, if you are stuck, then turn to generative AI tools. The *Possibility* Lab has a facilitator agent that knows and connects with all the other tools. You can say: "This is a problem I've been working on" or "I don't even know how to think about this." The facilitator will ask for context and suggest the most suitable tools to use. There are tools to help you become aware of possibilities (e.g. using analogies); explore those possibilities in depth (testing assumptions, considering scenarios); refine possibilities (thinking through unintended consequences); and plan and implement new ones (setting goals, monitoring progress, developing full projects).

Another tool is the *Lesson Unplanning Bot*. It helps teachers take over-planned, predetermined lessons – the kind you hate teaching – and breathe creative life into them. It helps you *unstructure* the plan, introduce structured uncertainty, and reimagine the lesson.

And yet another tool is the *Legacy Project Bot*. This one helps students develop creative projects that make an impact in their schools or communities, like addressing food waste or designing a safe after-school space.

These three examples are based on my work and other relevant scholarship. They are grounded in my definitions of creativity. Importantly, all three are designed to empower and maintain creative agency, rather than surrender it to the machine.

OECD: Let us talk about the emerging empirical evidence. There are studies comparing creativity outputs where people are allowed to use generative AI or not. One shows that individual outputs (judged by human raters) are typically more creative when AI is used as a help to provide a first idea, but there is less collective originality among those who used GenAI. What do you make of that?

Ron Beghetto: My hunch is that, yes, these tools can augment creativity. I know it from experience. But you cannot forget the knowledge and experience of the user. They can bring less experienced users up to a certain level. But without deeper knowledge, you do sometimes get homogenised outputs, and less diversity than if you were working with a highly skilled creative collaborator. I think if someone already has good ideas and can judge what the AI produces, rejecting what does not make sense and keeping what does, they can certainly be more creative. There is also evidence that even experts sometimes dismiss AI contributions that could be valuable. Or conversely, audiences sometimes rate AI outputs as superior to human ones. Evidence is still emerging, but the same criteria apply: do not be too dogmatic or you might overlook something creative. Build on domain knowledge, be open to uncertainty, and show flexibility.

OECD: And what about their accuracy?

Ron Beghetto: Humans hallucinate too. Humans say inaccurate things. Creativity sometimes thrives on "hallucinations", and there may be something worth pursuing there. But I would not rely entirely on generative AI tools for factual answers. I use them to support new thinking. The human must do the fact-checking and empirical testing.

Beyond text generation: multimodality and general artificial intelligence

OECD: Beyond text, what do you think about generative AI tools that produce music, video, images? Can we also use them in creative ways? Will they replace human creativity?

Ron Beghetto: Again, it depends on mindset and orientation. If you approach them with no clear question or purpose – "Just do this for me" – they can indeed replace your creativity. Or they simply become overwhelming. That is another reason why you should always start with a project or goal, not simply: "I have a deadline, please do this for me." Sometimes, of course, that will happen. But ideally, you approach them thinking: "I need some feedback or examples."

I would typically use different generative AI tools: ChatGPT, Gemini, Claude, and open-source models. Each has a slightly different "personality." I set the ground rules and provide context. Then, I treat them like a panel of colleagues. I present the same problem to each one, I share my initial thinking, and I compare perspectives. If one says something interesting, I might take that and ask another one to build on it. Or ask: "Poke holes in this idea: how might it fail?" That is, I think, the most powerful use: as a panel of different perspectives, always with you in control. And yes, sometimes you will want to add music or visuals. But you must remain the one deciding when and why.

These tools can accelerate and augment what you can already do, and take you further, just like working with any skilled collaborator. They hold a lot of "knowledge" so they can speed up learning. But you have to cross-check everything, just as you would with human sources.

We should absolutely not limit their use to higher education. Younger students are already using them anyway. They just need to learn to use them in a principled and responsible way, checking, questioning, and developing critical thinking. And remember that this is evolving rapidly. What we are discussing now will soon be out of date. This is not like any other subject or technology I have seen in my life. The acceleration is unprecedented.

OECD: What is your view on the future?

Ron Beghetto: The big threat is a crisis of meaning in education. If education is just about delivering inert content for students to reproduce, machines will do that better. And if students become digital puppets – "do this for me" – and teachers also outsource their feedback, education loses its purpose. That is why philosophers have always said education must be meaningful, experiential, purposeful. Otherwise, people will say: "Leave the inert knowledge to the machines - I'll just get the answer when I need it."

I think we are living in an important moment. I am actually quite optimistic, but we must be honest about the risks. This is a very different moment, not just another new technology. It is one thing to think about it as a productivity tool in industry. But in education, which is about learning, it is quite a different thing. And when you are a digital puppet, you are not really learning, and that is the crisis. Education has moved slowly for a long time, but perhaps this will accelerate some much-needed reflection about what it is for.

References for further reading

Beghetto, R., W. Ross, M. Karwowski, M., and V. P. Glăveanu (2025), "Partnering with AI for instrument development: Possibilities and pitfalls", *New Ideas in Psychology*, Vol. 76, p. 101121, https://doi.org/10.1016/j.newideapsych.2024.101121.   [4]

Beghetto, R. and F. Zamana (2025), "A Principled Approach to Human Creativity x AI in Education", in *The Cyber-Creativity Process*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-84535-2_3.   [5]

Doshi, A. and O. Hauser (2024), "Generative AI enhances individual creativity but reduces the collective diversity of novel content", S*cience Advances*, Vol. 10/28, https://doi.org/10.1126/sciadv.adn5290.   [3]

Katz-Buonincontro, J. and T. Kettler (eds.) (2025), *The Oxford Handbook of Creativity and Education,* Oxford University Press, https://doi.org/10.1093/oxfordhb/9780197698181.001.0001.   [2]

Vincent-Lancrin, S., C. González-Sancho, M. Bouckaert, F. de Luca, M. Fernández-Barrerra, G. Jacotin, J. Urgel, and Q. Vidal (2019), *Fostering Students' Creativity and Critical Thinking: What it Means in School*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/62212c37-en.   [1]

6

AI in Education Unplugged:
A conversation with
Seiji Isotani

This chapter is a conversation between Seiji Isotani, Professor of Computer Science and Learning Technology at the University of Sao Paulo (Brazil), and the OECD Secretariat. It focuses on remedying digital divides and focuses on AI in Education Unplugged, that is, the design of AI tools and AI interventions for places where digital infrastructure, connectivity and skills are limited. The chapter starts by presenting the idea and applications of AI Unplugged before zeroing in on the current and promising uses of generative AI systems in this context, notably with small language models. Isotani is now Faculty Director of the Learning Analytics and Artificial Intelligence Program at the University of Pennsylvania (United States).

AI Unplugged

OECD: Where does the concept of AI Unplugged come from?

Seiji Isotani: When we look at the AI space, people have been working on how to improve education with AI for decades. The AI in Education (AIED) society was created in 1997, almost 30 years ago. In the beginning, researchers tried to mimic what a teacher does, for example by trying to understand pedagogical strategies and their impact on learning. Others started working on "user" or student modelling, for example trying to predict how much students know based on their potential behaviour. Then we moved more towards what we call open learning models, which not only assess how students are doing but also present that information to students, so they are informed about what computers believe they know about them. It's more of a cognitive and metacognitive process.

The consistent problem in AIED has always been the expectation that students and schools, in general, have a minimum infrastructure to run AI. And as AI grows, the standard for this minimum requirement gets higher, right? This has become a problem, which is why people have discussed the digital divide for so many years. For

me, the question is: How can I bring the benefits of AI to people and regions where infrastructure is almost non-existent?

This completely shifts the research question and how you approach AI. Instead of creating the most innovative AI technology, you actually work with the community and ask, "okay, what does this community need, and how can I bring AI to them?" This is super interesting because by understanding users, we can create different technologies for that particular space. In 2022, about 90% of the world population, including those in low-income countries, had access to mobile phone subscriptions or mobile phones. If you go to poor places in Brazil, in countries near the Amazon, in Africa, or in parts of India, you will always see a mobile phone.

Then we looked at the statistics on Internet access, using data from 2022 and going back 30 years, so we could project trends. What we realised is that only 15% of people in low-income communities have access to stable Internet. This completely changes how we think about AI design, because we can consider using mobile phones and other low-cost equipment, but we cannot assume that the Internet will be available

everywhere. Even with the ambitious ongoing efforts to bring Internet access to everyone, the data give you the feeling that it's going to take about 100 years for low-income communities to have the same Internet access as higher income communities. And this considers the trends of things evolving and getting better.

OECD: So, what is the minimum infrastructure needed for AI Unplugged?

Seiji Isotani: We are developing a framework to implement AIED considering different degrees of infrastructure. To date, what we realised is that the minimum access required is a mobile phone (and not even a smartphone). That's the only thing we need. And some access to the Internet. It doesn't need to be all the time, every day, or every hour. At some point during the week you may have access to the Internet. If you have these two things – limited access to the Internet and a mobile phone, just a good enough mobile phone – then we can use AIED Unplugged. If you have Wi-Fi access once every week, I think it's enough, because it's sufficient for us to update information from the local equipment to a server. This allows us to do more intensive processing, analyse the data, update anything that needs updating, and then return information about the students.

AI Unplugged in action

OECD: What are the things that AI Unplugged allows you to do? Could you give us an example?

Seiji Isotani: Sure. Right after the COVID-19 pandemic, the Brazilian government asked us to help improve the writing skills of students in 5th to 9th grade. In Brazil, pupils spent almost two years out of school without writing, so when they returned they couldn't write well. The government asked us to try to do something about this at scale in Brazil, considering all the inequalities in the country. We accepted the challenge.

So, we performed a data analysis on mobile phone access, Internet access, and so on. In Brazil, most schools (over 90%) have some Internet access, but mainly for administrative tasks available at the principal's office, leaving students and teachers with little to no access. In this context, at least one location in most schools in Brazil has Internet access. This is an interesting and important feature.

We created an application enabling teachers to take pictures of students' essays (Portela et al., 2024[1]). They would ask students to write essays on a sheet of paper, and then take a photo of those essays (see Figure 6.1). Whenever the Internet was available – it could be the next day or two days later, or whenever possible, usually during lunchtime – a teacher would go to the principal's office and leave their mobile phone there. Our application would then upload all the photos to our server, perform the intensive processing, and return all the analysis of these essays. It would then provide the teacher with a dashboard for a particular student, for the whole class, or for a group of classes. The goal was to provide analysis and recommendations to empower teachers to improve how they support students in their writing. We analysed the outcomes with about half a million students in Brazil across 1 500 different municipalities over a year, and we found benefits from doing this. So, this type of AI technology that requires a minimal physical infrastructure offers significant support for teachers and helps students write essays through its automatic evaluation. We are now doing the same thing for basic mathematics and other subjects as well.

OECD: For that purpose, you didn't have to use generative AI, right?

Seiji Isotani: Yes, in that case we didn't use generative AI. We are now trying to use generative AI and compare it with the traditional methods we used. Right now, our previous AI model is still better at detecting incorrect words because what happens with generative AI is that it corrects the student's mistake when it processes a photo, thus correcting something that shouldn't be corrected in this context. For us to evaluate students, we need to know exactly what the student wrote, including their mistakes! But in one or two years, we will probably be in a position where generative AI will substitute all this previous work we've done.

OECD: How do you ensure that the recommendations given to teachers are pedagogically sound?

Seiji Isotani: That's a very good question. We have a library of the best pedagogical strategies that the field of the learning sciences provides to us. We try to match students' challenges or difficulties with these

Figure 6.1. AI Solution to Support Essay Correction
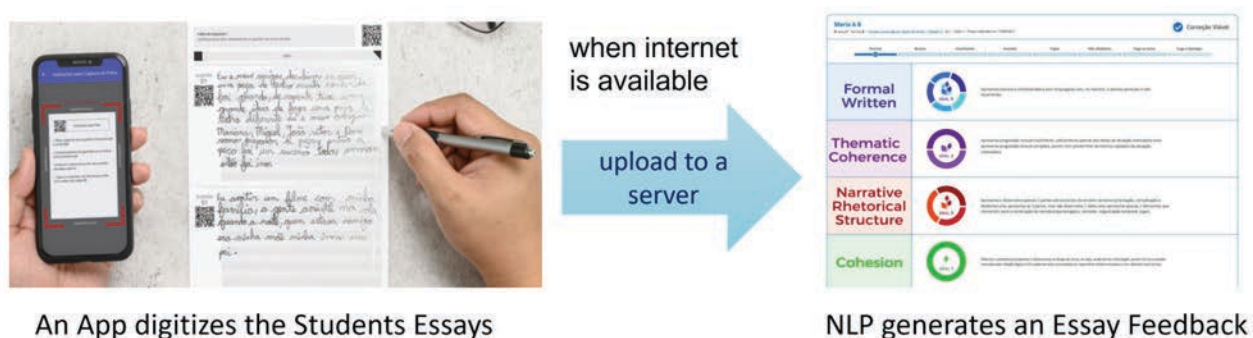
Source: Portela et al., (2024[11]), "AI in Education Unplugged Support Equity Between Rural and Urban Areas in Brazil", *Proceedings of the 13th International Conference on Information & Communication Technologies and Development*, pp. 143-154, https://doi.org/10.1145/3700794.3700810.

materials and strategies. For example, one of the materials we use is "WordGen." "WordGen" is a set of materials created by colleagues from Harvard University and other institutions. They work not only to support reading and writing but also to support reflection. Their general idea is that to support reading and writing, you need to engage students in interesting interactions. Just to give an example, imagine you are a student and you need to write and defend a position. You could talk about climate change, or about economic challenges. But students who are just starting to learn are usually more interested in local challenges. For example, in "WordGen," they have one problem related to lunch in the school cafeteria: should we have pizzas or salads for students? It's much more interesting to discuss this topic. Or, should we have mobile phones in school or not? It's interesting because you can also provide data about this: How good is a pizza? How good is a salad? What benefits do you get from each? Then you can position yourself in one of these directions, and during the debate, you bring all those ideas you collected from your reading and start to debate. Then you become much more critical when you are discussing. So, the idea of producing critical thinking or improving critical thinking becomes relevant in this case. In this context, the role of AI could be to match students' difficulties with potential pedagogical strategies and materials that can be used by teachers. The AI might assess the class and say "Okay, so students are not well-versed in semantic analysis, so we need to improve their capacity to understand more complex sentences, to retain their

meaning." Based on that, it can recommend "WordGen" together with pedagogical strategies for reflection and peer reading that we know from the learning sciences are proven to work for that specific need.

OECD: Thank you. We have a good example here of how to use unplugged AI when the AI tool is teacher-facing. Do you have examples of AI applications when it is used directly by the students and not by the teacher?

Seiji Isotani: Right now, I don't think we have a good example of a student using it directly. Because one of the foundations of AIED Unplugged is trying to reduce the amount of equipment needed. In low-income communities, most students do not have devices – so a scenario where they would all have one is not really helpful. We always think in this context that the proxy can be a teacher, a parent, or a mentor. We empower this proxy user so the students receive the greatest benefit. This is one of the challenges of using AIED Unplugged, because students who benefit from it are vulnerable. Current AI still has some biases. We can avoid them to some extent by having a proxy who tries to understand what makes sense and what doesn't to support that particular student. So, we work with an intermediary that prevents students from being affected by additional biases, that's the idea.

GenAI Unplugged

OECD: Let's move now to the small generative AI models. How could they help? They are interesting because, of course, they can be used to support teachers. They can also directly support students, even if it's through their parents or siblings, or whoever has a device. They could help students to develop their AI literacy in contexts with little resources. What are the possibilities for actually using generative AI in the unplugged model?

Seiji Isotani: Yes, that's a key question that we are discussing in our group.

I think the first point is breaking another barrier to access AI and technology in general. Even with mobile phones, we still require the user to have at least a minimum knowledge to use a mobile phone. When we use these AIED Unplugged models, they could run on a mobile phone or any other device. The interaction interface can be voice, which completely changes the game. Users are not just clicking; they are talking. And by talking, we can have a communication and interaction that students and teachers are already knowledgeable about.

Then, whenever you have a more knowledgeable partner and you want to ask and learn from them, you start asking questions and analysing their responses and trying to use them in your everyday life. So, when I think of GenAI in the AIED Unplugged model, it's like a companion that helps you solve challenging problems. It's not meant to substitute for anything, but it will enhance your capability to do things better. When you go to schools in the Amazon in Brazil, or in remote areas in many countries, you see that schools often don't have distinct grades. Five-year-old, six-year-old, ten-year-old students are all together. But they need different kinds of support, and the teacher isn't knowledgeable about everything and also need support. In this scenario, GenAIED Unplugged can actually provide specific or tailored support for students when they need it, at their level. So, that's something we are trying to produce right now. Our work is exploring whether we can ask something and receive a response in a good enough amount of time so that people with no connectivity can have a fruitful conversation. If it is possible, then the next question is, "are those interactions adequate or correct, or do they help students learn something?" These are the next steps that needs to be addressed.

OECD: How does it work? If I put a Small Language Model (SLM) on my phone, does that mean that when I'm not connected to the Internet, I can still interact and get responses from the chatbot and everything?

Seiji Isotani: Yes. You would just have a small or mini version of any LLM on your mobile phone (ChatGPT, Llama, DeepSeek, Mistral etc.) and it would work offline. It's a smaller version of an LLM, which means that hallucination is more problematic and it is not as powerful. Responses are sometimes incomprehensible. Words can be invented, so there are several problems to use it for learning right now. But on the good side, we do have a mini version of the web in our hands, so we can ask questions, get answers, and get help in different ways. So, the interaction process doesn't change much compared to an LLM: their capability to respond in different ways is limited, but they are still capable of doing things. One work that we are yet to publish, but are finalising, explores how many different pedagogical strategies can actually be used by LLMs in both an online and an offline environment (SLMs). In an online environment, LLMs are much more capable of using different strategies. If you ask them, "use the Socratic method to teach me something," they will do it, that is, ask questions and not give you the answer. On the contrary, if you are using a SLM on a mobile phone and you ask "use the Socratic method," probably - and we have observed this - in almost all cases, they will just give you the response. They won't use the strategies you want. So, these are some examples of the limitations.

OECD: Does it make a difference for a small model to be offline or online? Would the small model online be more performant than the small model offline, or would it be more or less the same?

Seiji Isotani: In our case, it doesn't change anything if it's online or offline, because the only thing we are doing is analysing interactions. If you want to create models that update over time or can search for current information on the Web according to student interactions, then an online model can make a huge difference. But if you are just thinking about the interaction itself, then offline is fine.

OECD: You have studied the trade-offs between different types of small language models. What are the lessons from your work on that?

Seiji Isotani: The lessons learned are that every community has different challenges, and we realised that AIED Unplugged can potentially address about 70% or 80% of these challenges. We are always thinking about these large language models, but perhaps for educational purposes, we should consider more seriously small language models. Small language models seem to be more effective for different purposes such as handling specialised, domain-specific tasks, like offering short feedback on a decimal misconception; they do not need to be online and they are not as high-cost as large language models. In addition, AI agents in the space of small language models also seem to be a promising path forward that few people talk about. People are discussing these huge agents that can do several different things and produce a final result. But when we think about AIED Unplugged, perhaps we can think of agents running in small language models to complete small, specialised tasks for teachers, for example, lesson planning, or creating specific materials for a particular activity for a student of a certain age. These are small agents that you can actually create to help teachers produce better quality materials and support them in their activities. They don't need to be able to do everything.

For students, agents are really interesting as well. Think about students in high-income families and what they have access to: if they struggle, their family might hire a private tutor so they can improve. Students with mental health issues will have access to a psychologist or someone working on their well-being. If uncertain about their career, a career adviser will help them, right? Three different things. Students in low-income families don't have access to this, but with small language models, using a mobile phone, you can actually run three different specialised agents: one focusing on the student's cognitive capability, another on the student's well-being, and another on the student's career. Then AI agents and students' caregivers can collaborate with students to make the best decision about what they should do next. So, I think this is a very promising path forward because, in many cases, those students lack any kind of support right now.

OECD: Large Gen AI models are increasingly getting trained for specific purposes, a bit the way you were describing. Is it possible to do something like that with a small model? Could you train a small model to be more focused on academic or on social and emotional skills, or to really have some kind of specialty? And can it be combined with other types of "good old-fashioned AI"?

Seiji Isotani: Yes, it's possible to tailor a small language model to specific tasks and activities for education. For example, we can use RAG (Retrieval-Augmented Generation) and train those models on data just from Wikipedia for example, or any other source. Their responses would then be heavily based on Wikipedia information. This is completely feasible. We can also do it with other different sets of materials. We can think about books, materials from OECD, or several other different resources. So, these students can actually have agents that will help them in different domains. We could use ontologies and knowledge representation (i.e. symbolic AI) to create hybrid approaches that can potentially have better results without relying solely on a single AI technology.

OECD: How much time does it usually take for a small language model to respond to a question? And does it make a difference if I type it or if I speak?

Seiji Isotani: Yes, typing or speaking makes a difference because the model needs to transcribe your oral input from what you're saying and then generate the response. When you type, it's faster. In January 2025, the last time we experimented with voice, it took more than a minute to get a response. So, if you ask a complex question, it can take much more time. But I believe that with new models and optimisations we will probably get better results over time.

OECD: One last question. Here we're talking about language models. Generative AI is more than just language models. It can generate pictures, music, videos, and so on. How much of that can a small model do?

Seiji Isotani: That's a big limitation right now. With the capabilities we have, small language models are not enough to generate images, videos, and so on. You need a lot of processing power and energy to do that. And even now, with the current large GenAI models, it may take several minutes to generate a good quality image. So, right now, small GenAI models don't have

the capability to generate these high-quality materials. But my bet is that this is just a matter of time. Every two years, the power required to generate an image is reducing. OpenAI showed that in the beginning, responding to any prompt would cost USD 5, and now it's a matter of a pence - one cent of a dollar. So, I think things will continue to improve, and these smaller language models will be able to do something like generating images and performing more tasks.

References for further reading

Barros, A., M. Monteiro, L. Rodrigues, D. Dermeval, S. Isotani, and I. I. Bittencourt (2025), "Evaluating Large Language Model Quality in Resource-Constrained Environments: An Educational Stakeholders' Survey on Accuracy, Completeness, and Readability in Brazil", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-98417-4_16.   [4]

Monteiro Santos, M., A. Barros, L. Rodrigues, D. Dermeval, T. Primo, I. Ibert, and S. Isotani (2024), "Near Feasibility, Distant Practicality: Empirical Analysis of Deploying and Using LLMs on Resource-Constrained Smartphones", *Proceedings of the 13th International Conference on Information & Communication Technologies and Development*, pp. 224-235, https://doi.org/10.1145/3700794.3700817.   [3]

Monteiro, M., A. Barros, L. Rodrigues, D. Dermeval, I.I. Bittencourt, S. Isotani, and V. Macario (2025), "Small Device, Big Decision:" Comparing Lightweight LLMs' Computational Performance and Output Quality for AIED Unplugged", in *Communications in Computer and Information Science, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-99267-4_20.   [2]

Portela, C. P. Palomino, G. Challco, Á. Sobrinho, T. Cordeiro, R. Mello, D. Dermeval, I. I. Bittencourt, and S. Isotani (2024), "AI in Education Unplugged Support Equity Between Rural and Urban Areas in Brazil", *Proceedings of the 13th International Conference on Information & Communication Technologies and Development*, pp. 143-154, https://doi.org/10.1145/3700794.3700810.   [1]

Rodrigues, L., G. Guerino, T. E. V. Silva, G. C. Challco, L. Oliveira, R. S. da Penha, R. F. Melo, T. Vieira, M. Marinho, V. Macario, I. I. Bittencourt, D. Dermeval and S. Isotani (2024), "MathAIde: A Qualitative Study of Teachers' Perceptions of an ITS Unplugged for Underserved Regions", *International Journal of Artificial Intelligence in Education*, Vol. 35/1, pp. 2-30, https://doi.org/10.1007/s40593-024-00397-y.   [5]

Uema, M., T.P.C. Souza, D. Dermeval, I. I. Bittencourt, S. Isotani (2025), "Designing for Meaningful Access: Towards a Framework for AI in Education Unplugged", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-98459-4_22.   [6]

Part 2

Augmenting teachers' performance with generative AI

7 A conceptual framework for teacher-AI teaming in education:
Harnessing generative AI to enhance teacher agency

Mutlu Cukurova
University College London, The United Kingdom

As generative artificial intelligence (GenAI) technologies rapidly permeate education, this chapter offers a conceptual analysis of how AI is reshaping teacher agency and argues that strengthening this agency is paramount in an AI-driven era to ensure educational practices remain human-centred, ethically grounded, and conducive to the ongoing development of teacher competence. It introduces a framework that distinguishes between: (i) *replacement*, (ii) complementarity, and (iii) *augmentation* of teacher competence. Building on the latter two, the chapter proposes a five-level teacher-AI teaming framework comprising *transactional, situational, operational, praxical,* and *synergistic* modes of interaction. It highlights the unique affordances of GenAI to help us achieve more praxical and synergistic teacher-AI interactions, as well as GenAI's potential to enhance transactional, situational and operational teaming in a diverse set of tasks. The chapter concludes that while the replacement paradigm can indeed yield productivity gains, these benefits may come with costs that must be acknowledged and discussed to support informed decision-making.

Introduction

The growing presence of generative artificial intelligence (GenAI) technologies in educational contexts presents a paradox of empowerment and concern. On one hand, AI promises to relieve teachers of burdensome administrative tasks (Roy et al., 2024[1]), provide personalised learning insights (VanLEHN, 2011[2]), and complement instructional capabilities including for lesson planning, for classroom implementation and for assessment. For example, AI could help teachers plan their lessons by supporting their review of the knowledge to be taught, by providing ideas on alternative pedagogical strategies or by defining students' needs and familiarising them with such needs (van den Berg and du Plessis, 2023[3]). AI could also support classroom instruction through immediate feedback to students (Di Mitri, Schneider and Drachsler, 2021[4]) or to teachers about their practice (Demszky et al., 2025[5]) or about their interventions (Aslan et al., 2019[6]); and AI can be used for modelling students' mastery (Minn, 2022[7]), to generate assessment items (Chan et al., 2025[8]), supported essay scoring (Seßler et al., 2025[9]) or feedback on certain aspects of their teaching (Zhang et al., 2025[10]).

On the other hand, there are significant concerns regarding potential threats, including the erosion of teacher autonomy (Miao and Cukurova, 2024[11]), de-professionalisation of teaching (Berendt, Littlejohn and Blakemore, 2020[12]), and ethical pitfalls if AI is misapplied (Holmes et al., 2021[13]). Many education experts are also concerned that over-reliance on AI to perform teachers' tasks, such as marking, feedback generation, and lesson planning could risk skill atrophy for teachers (Felix and Webb, 2024[14]). Therefore, the notion of teacher agency (i.e. teachers' ability to exercise professional judgment, innovate in practice, and maintain control over pedagogical decisions) has come to the forefront of debates on AI in education.

This chapter emphasises the need and urgency for moving beyond polarised narratives, rejecting both the dystopian fear of AI teachers replacing humans and the utopian hype that AI alone will solve all educational problems. Instead, it calls for evidence-informed strategies that harness AI's potential while safeguarding teachers' agency, rights, and professional integrity. As professionals increasingly offload tasks, that were traditionally viewed as uniquely human, to AI, we must engage in profound and forward-looking reflection about the essential roles and identities of professionals. What are the enduring and uniquely human qualities at the heart of education? How can we safeguard, elevate, and flourish them as technology transforms lives and the broader teaching and learning ecosystem? What is the role of a teacher in education where technological advancements allow the generation and sharing of information effectively and efficiently?

To frame the inquiry into these difficult questions, the chapter outlines a continuum of AI integration in teaching, distinguishing between *replacement, complementarity,* and *augmentation* of teacher competence with AI in educational practice. *Replacement* through automation refers to tasks being offloaded to AI systems and carried out by them with no further teacher intervention. *Complementarity* requires a careful consideration of what teacher competence means for a given task and a specific AI tool's design, development, and deployment features that can complement the specific aspects of the teacher's competence. *Augmentation* suggests a deeper assimilation of AI models into teachers' cognitive and pedagogical processes to be able to improve teachers' competence through iterative interactions so that AI-augmented teachers can achieve the task better than AI or teacher alone (i.e. competence augmentation increases the likelihood of teacher-AI teams to perform better than the best of teacher or AI alone). Understanding these conceptual models is crucial for identifying when AI use supports teacher agency, when it might erode it, and when teachers' agentic interactions can lead to competence augmentation for teachers, and thus to more effective teaching.

How do teachers use GenAI in education: early benefits and concerns

A review of the literature reveals that while AI applications in education are burgeoning, our understanding of how exactly teachers are integrating GenAI into their practices at a scale remains nascent and continually evolving. Additionally, the ways in which teachers engage with GenAI vary significantly across countries and jurisdictions, making it challenging to comprehensively capture the full range of practices in this chapter. However, for instance, based on the open call for evidence of the use of GenAI in education by the UK Department for Education, which received (non-representative, but elaborate) responses from 567 participants, the majority of whom were teachers, the public generally believed GenAI to offer various opportunities (Department for Education, 2023[15]). These include freeing up teachers' time, improving teaching and educational materials, providing additional support to students (particularly those with special educational needs and disabilities (SEND) and those for whom English is an additional language (EAL)), and enhancing subject-specific applications (e.g. STEM). Overall, these perceived benefits argued to outweigh concerns about GenAI in education (e.g. students' overreliance on GenAI, academic misconduct, fear of GenAI replacing face-to-face teaching, and the "digital divide"). Most use cases observed in self-declared surveys and interviews of teachers also indicate that teachers use GenAI to develop lesson materials, ensuring alignment with curricular objectives while saving time on content preparation (Department for Education, 2024[16]). In assessment, teachers tend to use it to support their marking and provide personalised formative feedback for students. Beyond the classroom, teachers appear to use GenAI in drafting statutory policies, streamlining administrative tasks, and aiming to reduce bureaucratic burdens (Department for Education, 2024[17]).

While self-reported survey data offer some insights into teachers' usage of GenAI, this method faces inherent limitations, notably biases arising from external pressures or social desirability that may prevent teachers from accurately reporting their AI practices. Recent research from Anthropic.ai (Handa et al., 2025[18]) analyses over four million conversations with their GenAI system and shows high reliance on AI for some professions. Among certain professional groups, including language and literature teachers, AI conversations correspond to the performance of more than 75% of their professional tasks (following the task mapping for occupations by the O*Net database of the US Department of Labor). While understanding the extent of GenAI usage among teachers is crucial, examining precisely how teachers employ GenAI, whether primarily for augmentation or automation, is equally significant. Although it is a task-dependent discussion, evidence suggests a tendency towards automating routine tasks such

as content generation (Handa et al., 2025[18]), with AI directly executing responsibilities requiring minimal teacher involvement. Although such conclusions should be interpreted cautiously, as log data alone cannot reveal how teachers ultimately engage with, or act upon AI-generated outputs (e.g. some may discard them entirely and continue their work independently), automated applications of AI in education raise critical questions regarding the broader implications for teacher agency.

Beyond the recent use cases of GenAI, the past few decades have witnessed an explosion of interest in applying AI to education, from intelligent tutoring system and automated grading tools to AI-driven decision-making tools and adaptive learning platforms. There is substantial amount of evidence that shows the positive impact of using these AI applications to support students' academic performance (Li, Gobert and Dickler, 2019[19]; VanLehn et al., 2020[20]), their affective engagement (Bosch et al., 2016[21]; Calvo and D'Mello, 2010[22]), and metacognitive development (Azevedo, Cromley and Seibert, 2004[23]; Laru and Järvelä, 2015[24]) in controlled experimental evaluations. Although these small-scale empirical studies provide valuable insights, their outcomes typically reflect carefully designed academic AI tools evaluated under controlled conditions in which teacher implementation is guided by researchers in high fidelity.

Alongside these promises, early scholarly and practitioner commentary raised flags about possible pitfalls. These concerns have been amplified by the rapid rise of GenAI, which brought AI's capabilities, along with its risks, into mainstream awareness. Therefore, most educational stakeholders recently began grappling with scenarios that once seemed futuristic, or of interest to a small group of scholars.

One core concern in these discussions is the potential erosion of teacher agency. As GenAI systems begin to handle not just typical monotonous administrative tasks of teachers but also complex pedagogical and instructional decisions, such as selecting content, assessing student work, or providing feedback, teachers might find their professional judgment marginalised by algorithmic outputs. Recent empirical research is starting to document these dynamics. For instance, a recent study by Guan, Zhang, and Gu (2025[25]) on pre-service teachers indicated that exposure to GenAI in education prompted reflections on their evolving role and anxiety about role change, highlighting the need to prepare teachers for new hybrid roles working alongside AI. Complementary evidence from interview-based research with 57 teachers across eight schools in Sweden and Australia further reveals that, rather than freeing teachers from work, GenAI often generates new forms of invisible labour while challenging their agency on the pedagogical appropriateness and social sensitivity of educational content (Selwyn, Ljungqvist and Sonesson, 2025[26]). These findings suggest that while GenAI tools have significant potential to support teachers and teaching practices, they can, in practice, redistribute and obscure teacher labour, reinforcing rather than reducing workload anxieties and concerns over teachers' professional autonomy.

A working definition of teacher agency

In this chapter, teacher agency refers to teachers' active capacity and incentives to make choices and exert influence in their professional practice. It encompasses the autonomy to make instructional decisions, the ability to adapt and innovate pedagogy, the power to shape the educational environment in accordance with their professional values and their students' needs as well as the willingness and incentives to do so. In essence, teachers have agency when they "*act rather than are acted upon*" in the educational process. Their agency is rooted in professional competence and confidence, and is often enabled or constrained by the tools and technology they use in their practice as well as the broader institutional and policy context in which they are situated. Teacher agency is not an all-or-nothing attribute; it exists in degrees and forms. Educational sociologists have described multiple forms of teacher autonomy, for example, autonomy over curriculum content, over pedagogy, over student assessment, and over professional development pathways (Frostenson, 2015[27]). A supportive school culture and policy framework can expand these autonomies, whereas top-down mandates, high-stakes accountability regimes, or technology systems making autonomous decisions on their behalf can compress them (as can be the case with AI-based educational technologies).

Teacher agency is a key concept in education since teachers' sense of agency is linked to their motivation, job satisfaction, and willingness to embrace new pedagogies or tools. When teachers feel empowered to make decisions, they are more likely to take initiative in improving their teaching and respond creatively to challenges in the classroom. Teacher agency enables teachers to adapt and update curricula, incorporate emerging real-world problems, and contextualise learning experiences to social, emotional, and relational needs of their students in ways that AI algorithms cannot. Moreover, teacher agency is tightly connected to teacher identity, that is, the sense of oneself as a professional with a meaningful mission. The introduction of GenAI into the classroom can perturb that identity, since some teachers may fear being displaced or judged by GenAI systems, while others might see GenAI as an opportunity to enhance their effectiveness.

From a learner's perspective, teacher agency has the potential to translate to richer educational interactions. A teacher with high agency will actively interpret AI-generated insights or recommendations and adapt them to the context of their students. For example, if a GenAI tool generates feedback on students' essays or produces a set of suggested prompts to foster metacognitive reflection, a teacher exercising professional agency will treat these as provisional resources (e.g. reviewing their pedagogical relevance, rephrasing or extending them to align with students' learning goals), and using them as a springboard for discussion or further inquiry. By contrast, a low-agency scenario might involve a teacher simply pasting the AI-generated feedback into the learning platform without review or contextualisation, or alternatively, disregarding the system's suggestions entirely due to mistrust or lack of confidence. Both extremes are suboptimal, and the goal is a balanced partnership where the teacher remains the orchestrator of instruction, using AI tools as informative assistants. Although most researchers and practitioners would agree with the proposed need for balance, there is little understanding regarding where this balance stands for a given teacher task, how it can be conceptualised, and how it can be operationalised. This chapter is an attempt to fill in this gap. In the following sections, the chapter dives deeper into how a theoretical conceptualisation of AI's role vis-à-vis teacher agency, introducing the conceptualisations of replacement, complementarity, and augmentation, and then proposes a five-level teacher-AI teaming framework with clear definitions and examples of each level. Although, a broad set of socio-technical, institutional, and cultural factors profoundly shape how teachers perceive, enact, and sustain agency in their interactions with AI, the five-level framework proposed here mainly examines teacher agency through the lens of AI system affordances and interface-level design considerations.

Three conceptualisations of AI in education and implications on teacher agency

This section presents three conceptual modes of integration as a framework to differentiate AI's roles and their implications for teacher agency. These modes can be considered as points along a dynamic spectrum from AI operating independently of teachers to AI becoming deeply embedded in teachers' cognitive routines to augment their competence. By delineating these, the chapter aims to clarify which approaches threaten teacher agency and which can potentially bolster it. Although the framework is applicable to all forms of AI, generative AI provides distinctive affordances that can shape and support varying levels of teacher agency. These will be discussed with examples where appropriate.

Replacement

Replacement refers to AI systems executing tasks that a teacher would typically do with an AI-driven process with minimal or no teacher intervention. Classic examples include automated grading of exams or essays, asking an AI model to create lesson plans, questions, materials, algorithmic scheduling of student practice, such as homework, or AI tutors delivering content directly to learners, automating the pedagogical practice of teachers. The main appeal of automation is efficiency and scalability. Indeed, certain labour-intensive tasks in teaching (e.g. grading multiple-choice quizzes, drafting lesson plans, generating practice problems) can be reliably automated, potentially freeing teachers' time for other or more complex work (e.g. Leiker, Finnigan and Cukurova, 2023[28]; Pea et al., 2022[29]; Roy et al., 2024[1]). Influenced by the political, managerial, or leadership level ecosystemic issues of teaching practice, currently the evidence about how teachers actually use this saved time is scarce.

The potential of productivity gains in education is indeed important. Recent research by the Education Endowment Foundation (EEF), independently evaluated by the National Foundation for Educational Research (NFER), examined the use of GenAI among 259 teachers across 68 secondary schools in England. The randomised controlled trial revealed that teachers who used GenAI, supplemented by practical guidance, reduced their lesson and resource planning time by an average of 31 percent, reducing their weekly average planning time from 81.5 to 56.2 minutes, without compromising the quality of their lesson plans and resources (Roy et al., 2024). As highlighted by the Teacher Task Force and UNESCO Global Report on Teachers (2024), education systems face compounding teacher and resource crises, especially in low- and middle-income countries. An estimated 44 million additional primary and secondary teachers are said to be needed by 2030, including 15 million in sub-Saharan Africa. This worldwide shortage of teachers is aggravated by rising attrition, as many teachers leave the profession early. Rural and remote areas have been hit hardest, where underqualified teachers often fill the gaps and multi-grade classrooms are common; 90% of secondary schools in sub- Saharan Africa face serious teaching shortages. Consequently, learning gaps are widening. Students also contend with severe shortages of education materials and quality content. In some classrooms, a single textbook must be shared by a dozen or more pupils. Much of the digital education content that could help is not in the learner's language. For instance, even though they are not representative of countries' learning resources, 92% of open education resources are in English, marginalising non-English-speaking learners (UNESCO, 2023[31]). The replacement paradigm in AI offers

opportunities to counteract these global shortages with significant productivity gains. GenAI-powered tools can supplement overburdened teaching workforces and provide instructional support to students in underserved areas. Translation and content generation driven by GenAI can expand the availability of high-quality teaching resources in local languages and for students with Special Educational Needs and Disabilities (SEND).

Nevertheless, the full automation of teacher tasks also raises concerns about the loss of teacher agency and may come at certain costs. The purpose of highlighting these challenges is not to oppose productivity gains in education, but to invite policy dialogue and careful consideration of how some of these challenges can be mitigated. Two of those concerns are the dehumanisation of education and teacher cognitive atrophy.

Figure 7.1. The replacement paradigm on the coordinates of teacher agency vs automation
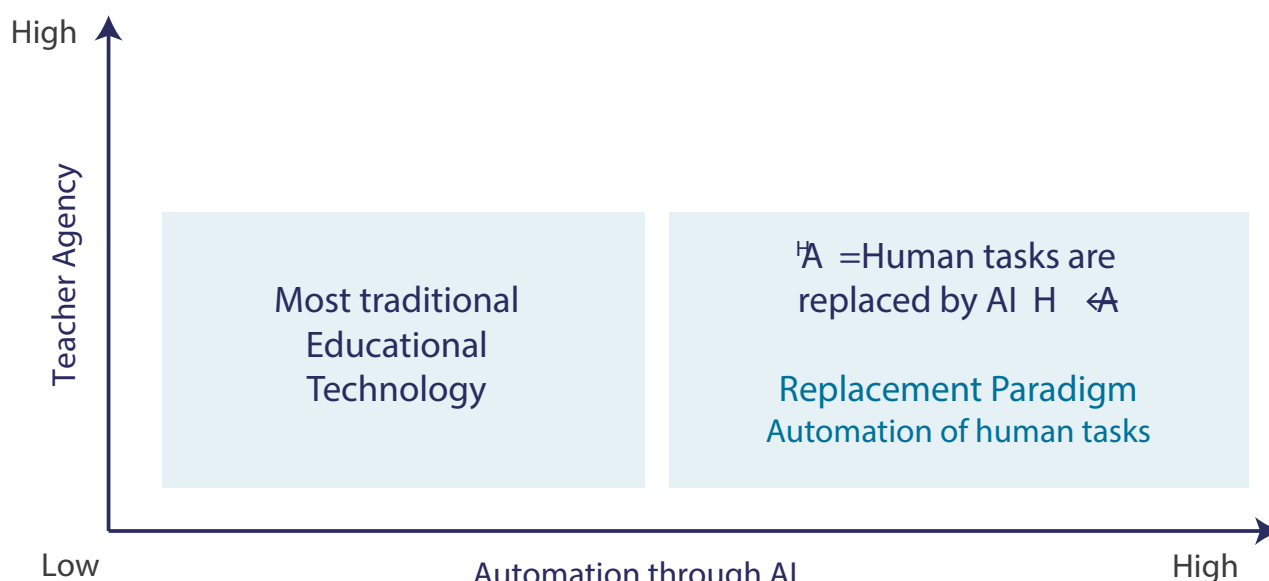
Source: Adapted from Cukurova, M. (2024[30]), *The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence.* British Journal of Educational Technology, 56(2), 469-488

Dehumanisation refers to the erosion of the human elements that are fundamental to teaching and learning. Education is an interactive process of human development. When AI is used as a substitute for teachers or peers, there is a danger that learning becomes overly mechanistic, losing the empathy and social dynamics that characterise many effective pedagogies and educational practices. Dehumanisation can manifest in teaching, assessment and feedback for example. Some may be tempted to have "teacherless schools" and provide instruction by having students working through AI-personalised curricula. However, human teachers contribute numerous intangible qualities (e.g. moral judgment, inspiration, role-modelling, the ability to build trust with other humans and mentor them) that no AI currently can replicate. Even if students progressed academically, they would miss out on collaborative learning, dialogue, and the social construction of knowledge with other humans. Educational practice often thrives on inquiry-based instruction, collaborative lab-work, group discussions, and debate, which are facets that require human presence and guidance. Similarly, feedback and evaluation from impersonal algorithms might make students feel less seen or valued as individuals. Finally, if teachers were asked to work under algorithmic scripts or performance dashboards that dictated their every move, their professional identity would likely erode and resemble more that of assembly line workers than educators. Protecting teacher agency is thus directly tied to keeping education human-centred.

Similarly, the integration of generative AI into educational environments in ways that do not allow teacher agency also raises significant questions about its impact on teachers various higher-order thinking skills, particularly through its influence on critical and reflective thinking practices. A recent experiment focusing on the cognitive cost of using a Large Language Model in the educational context of writing an essay shows that students writing without an LLM

support exhibited the strongest, widest-ranging brain activation, those using a search engine showing intermediate engagement, and those using an LLM a limited cognitive engagement (Kosmyna and Hauptmann, 2025[32]). While the evidence on the potential negative impact on GenAI users' cognitive capabilities when they use GenAI is just emerging, this may also be a risk for teachers when they use AI in the replacement paradigm.

Complementarity

Complementarity refers to AI systems functioning as supportive tools that amplify a teacher's capabilities, while the teacher remains actively involved. In the complementarity paradigm, AI and teacher work in tandem, each contributing what they do best without necessarily interacting with each other for augmentation of each other's competence. The underlying philosophy of complementarity is that leveraging AI's strengths (e.g. data processing, pattern recognition, speed, scale, no exhaustion, efficiency and time-saving opportunities etc.) to complement human strengths (e.g. relational interpretation, empathy, moral discernment, contextual judgment etc.). In this paradigm, as AI primarily processes data to present insights or operationalise instructional intentions defined by teachers. It is less about AI learning from humans and more about teachers internalising computational representations and reshaping their own mental models and professional reasoning through complementary interactions with AI. Contrary to the replacement paradigm, AI systems don't perform teachers' tasks entirely.

Figure 7.2. The complementarity paradigm on the coordinates of teacher agency vs automation
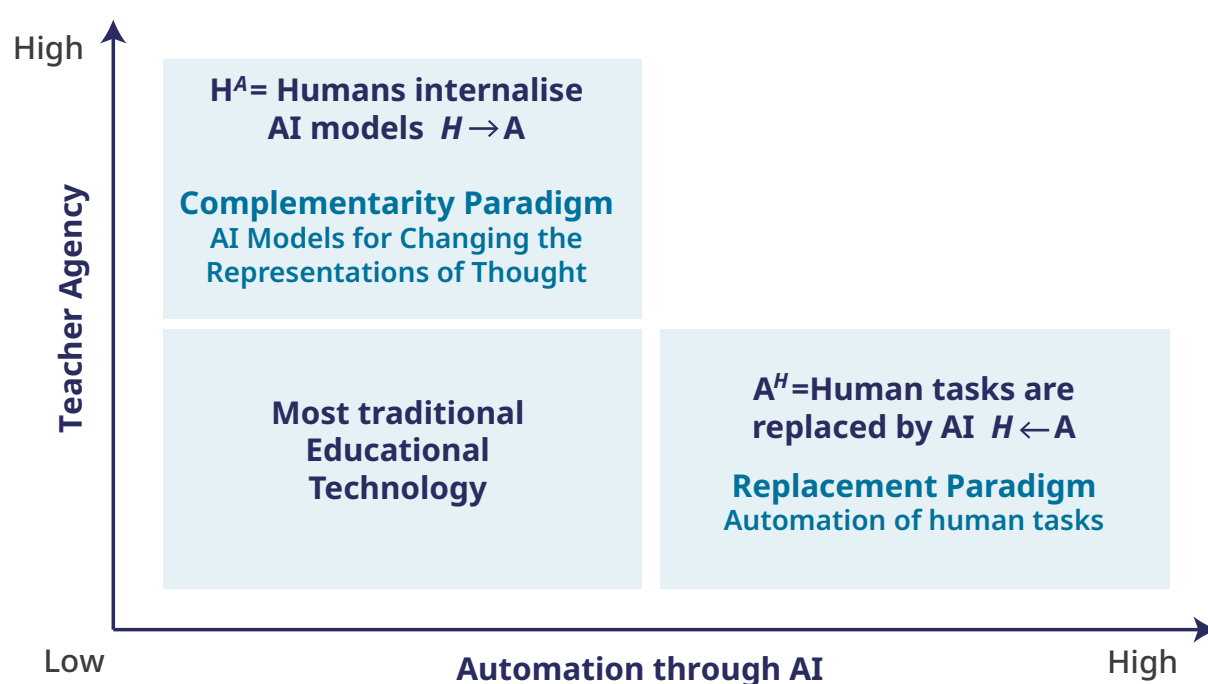
Source: Adapted from Cukurova, M. (2024[30]), *The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence.* British Journal of Educational Technology, 56(2), 469-488

If implemented appropriately, complementarity conceptualisation can further reinforce the teacher's agency and has the potential to improve human competence at a given task. For complementarity to be operationalised, we must first articulate a human competence model to be able to specifically define what aspects of a teacher competence[1] can be complemented with AI. Then, we can examine how the affordances of a given AI modelling technique or the design of a specific AI agent can interlock with each layer of that competence continuum. Holstein et al. (2020) provide a useful framework for thinking about what specific aspects of teacher competence can be complemented with the help of AI, identifying four dimensions of complementarity: complementary goal setting (e.g. when teachers set, monitor, and evaluate learning goals with AI support);

complementary perceptual input (e.g. when teachers' perception about student learning is expanded by AI); complementary actions (e.g. when teachers' actions are scaled by AI); complementary decisions (e.g. when teacher decision-making is assisted by data-informed AI recommendations). For instance, an AI model that has the capabilities of processing sensory information from students' interactions can complement teachers' situation specific skills of monitoring student interactions. On the other hand, an AI model that has the capability of tracking students' online interactions in an intelligent tutoring system can complement teachers' knowledge of their students' current level of mastery on a topic. Such a teacher-AI complementarity would have the purpose of supporting rather than supplanting teachers' integrative professional judgment. However, this framework does not provide any insights into how exactly specific aspects of teacher competence can be complemented.

Building on this notion of complementarity, it becomes essential to conceptualise how teacher-AI interactions may vary in depth and complexity. Here we propose that for any given educational task, and depending on both the specific aspects of a teacher's competence, and the specific AI affordances, teacher-AI teaming can occur at five distinct levels: transactional, situational, operational, praxical, and synergistic teaming.

Transactional teaming

Transactional teaming refers to interactions between teachers and AI that consist of discrete transactions, defined by a request-response mechanism (i.e. the teacher inputs a command, and the AI outputs a result). Teachers and AI systems' actions are perceived by one another, with each action dynamically informing and triggering a corresponding response from the other agent. At this level, AI agents can automatically perform actions based on teachers' input, often completing routine or repetitive tasks on their behalf. The core dynamic of transactional teaming is "request to execution" for task automation and efficiency; therefore, the primary goal is to enhance teaching productivity by streamlining these processes. The Srivastava et al. (2021[33])'s Smart Learning Assistance (SLA) system can be described as an exemplar of transactional teaming in teacher-AI interaction. In this system, the teacher issues a command (e.g. inputting spoken words, sign language gestures, or Braille text), and the GenAI tool automatically converts to or from speech, sign language, or Braille, thereby returning a translated result that supports communication with the teacher or peers. In doing so, the SLA tool takes over routine translation tasks, allowing the teacher to delegate discrete, repetitive conversion work, precisely the "request → execution", dynamic that defines transactional teaming.

Figure 7.3. Transactional teacher-AI teaming
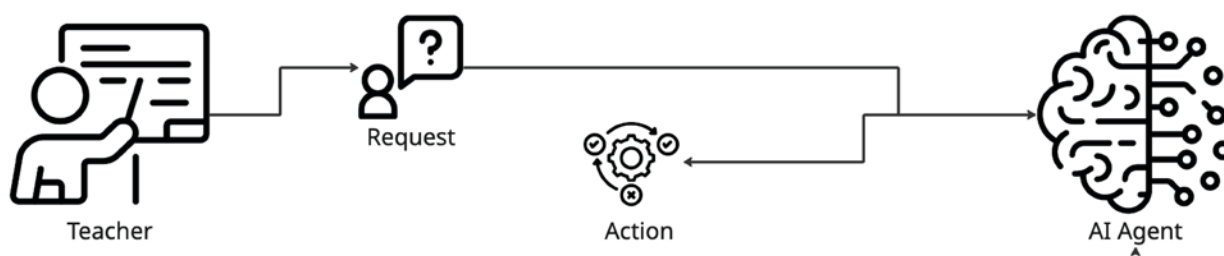
Source: Author's own work.

Situational teaming

Situational teaming refers to a form of interaction in which teachers operate based on a shared awareness of the teaching and learning context, constructed through the combined perceptions of both human and artificial agents. At this level, AI systems collect data from classroom interactions and/or learning activities in digital learning platforms or real-world classrooms through sensors, process it using underlying models, and provide educationally meaningful information to support teachers in making informed decisions and taking appropriate actions. The Hybrid Human-Agent Tutoring (HAT) platform developed by Sawaya et al. (2025[34]) can be used as an example of situational teaming. The system collects data on tutors' discourse practices, analyses it with AI models, and provides human teachers and coaches with GenAI created feedback to guide their instructional coaching sessions. This creates a shared awareness of the tutoring context, where AI highlights patterns in tutors' interactions and coaches use these insights to make informed pedagogical decisions. The core dynamic mirrors situational teaming in that teachers and AI agents co-construct context awareness that informs human action, rather than automating tasks.

Figure 7.4. Situational teacher-AI teaming
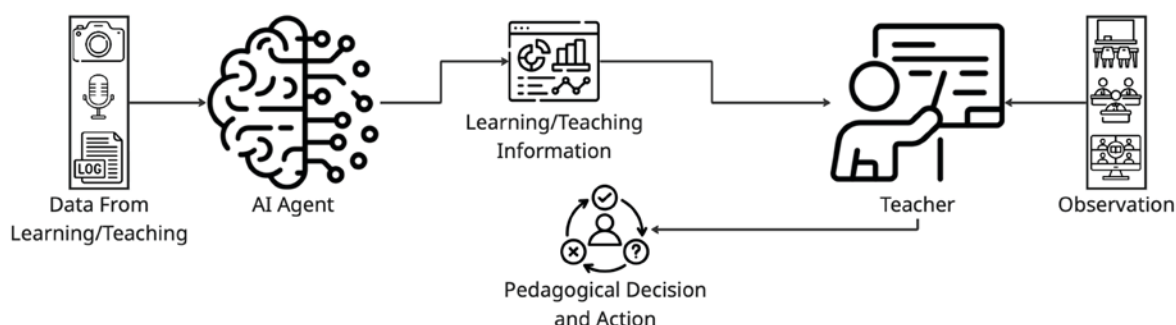
Source: Author's own work.

Operational teaming

Operational teaming involves the cooperation of planning and execution of teaching-related tasks between the teacher and the AI system. At this level, teachers provide information about the current and desired states of the teaching and learning context, articulated through intentions, instructional goals, tasks, and actionable plans. The AI system supports teachers by incorporating these goals in its decision-making to autonomously perform or assist with the goals set by teachers. This teaming enables efficient task execution aligned with the teacher's needs. However, it also requires teachers to have a comprehensive understanding of the instructional goals and the pedagogical interventions to achieve them. The Pair-Up system developed by Yang et al. (2023[35]) can be used as an example of operational teaming. In this system, teachers articulate high-level instructional goals (which can also be done with natural language using GenAI), such as when to transition students between individual and collaborative learning activities. The AI system integrates these goals into its decision-making by monitoring student progress in an intelligent tutoring system in real time, then recommending or enacting transitions for classroom practice that align with the teacher's pedagogical intentions. This cooperative planning and execution allow teachers and AI to jointly manage complex classroom orchestration tasks, with AI autonomously assisting in carrying out the instructional plans while ensuring that task execution remains consistent with the teacher's overarching objectives.

Figure 7.5. Operational teacher-AI teaming
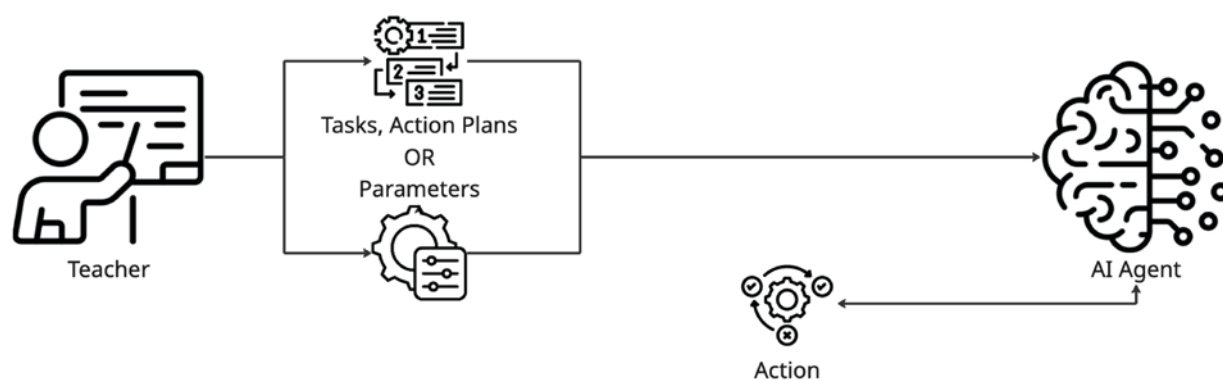
Source: Author's own work.

Praxical teaming

Praxical teaming refers to a form of interaction in which the teacher and the AI system exchange information about actions and procedures, grounded in prior experience, usage patterns, or training. This level of teaming emphasises the development of shared "understanding" and practices over time, enabling the AI to learn from the teacher's instructional habits and preferences, while the teacher adapts to the AI's capabilities and pedagogical suggestions. For example, when offering recommendations to improve teaching, an LLM-based conversational support system can

learn from teachers' feedback (both explicit and implicit) on those suggestions and adjust their underlying models accordingly. This requires teachers to possess the competence to critically evaluate the AI's suggestions, rather than accepting them uncritically and AI to have the capability to learn from teacher corrections to adjust its internal model accordingly. The machine-learning-based feedback suggestion system developed by Bernius, Krusche, and Bruegge (2021[36]) illustrates an example of praxical teaming. The system analyses students' python coding script submissions in large courses and proposes feedback suggestions that instructors can review, adapt, or reject. Over time, the AI learns from teachers' corrections and adjustments, refining its ability to generate more contextually appropriate feedback aligned with instructors' pedagogical preferences. This dynamic exchange of information, AI adapting to teachers' evaluative patterns while teachers critically assess and refine AI-generated suggestions, embodies the essence of praxical teaming, where shared practices and mutual adaptivity develop iteratively.

Figure 7.6. Praxical teacher-AI teaming
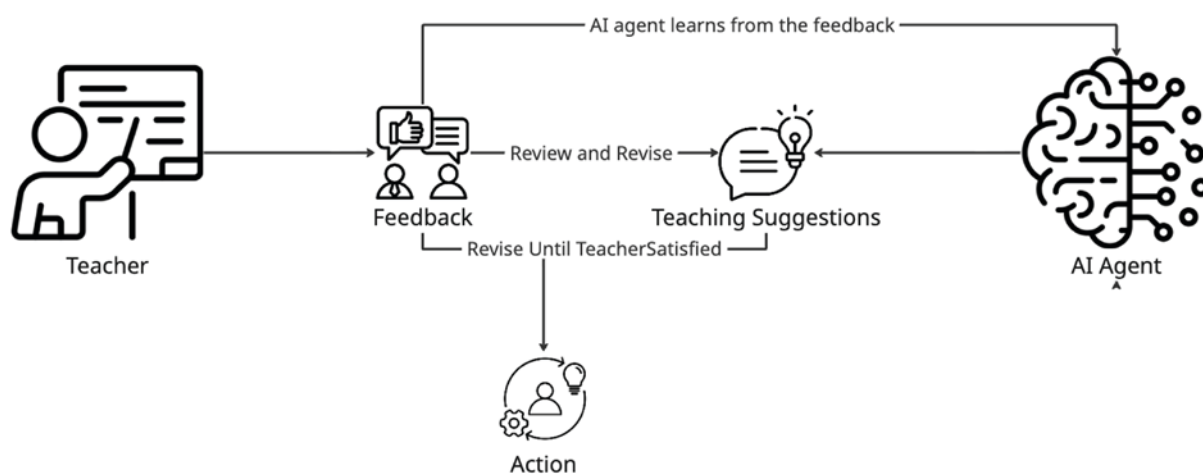
Source: Author's own work.

Augmentation

Augmentation refers to the process by which AI tools, and the new practices they enable, become woven into the internal repertoire of teachers in ways that these human-AI interactions also lead to an increase in teacher competence. This very much corresponds to the fifth level of the teacher-AI teaming framework.

Synergistic teaming

Synergistic teaming refers to a form of interaction in which the teacher and the AI system mutually enhance each other through critical evaluations, challenging to each others' suggestions and propositions with logic and evidence, and engage in solving complex problems together to move towards a shared understanding and mutual development. Effective synergistic teaming involves mutual interaction, where the AI agents and teachers evaluate each other's claims with epistemic awareness and remind one another of aspects that may have been overlooked. When this interaction is well-aligned, a form of creative resonance emerges, enabling the teacher and AI to deepen their understanding of the task and generate outcomes neither could achieve independently. Therefore, this type of teaming is conceptualised as synergistic, that is, the emergent competence is likely to exceed the maximum of individual AI or human competence at a given task (Cukurova, 2024[30]). The main limitations of praxical teaming in comparison to synergistic teaming lie in its reliance on the existing knowledge, expertise, and competence of the teacher, without necessarily pushing them beyond their current practices. As such, praxical teaming tends to converge at the maximum of the teacher's present competence, whereas synergistic teaming requires a more ambitious, mutual development between AI and teachers, a state that remains far more difficult to achieve in practice. Thus, while praxical teaming often stabilises at the ceiling of teachers' or AI's current max performance with the potential benefits of efficiency and time-savings, moving toward synergistic teaming, where AI and teachers mutually extend and transform one another's capacities, is more difficult to realise.

Figure 7.7. Synergistic teacher-AI teaming
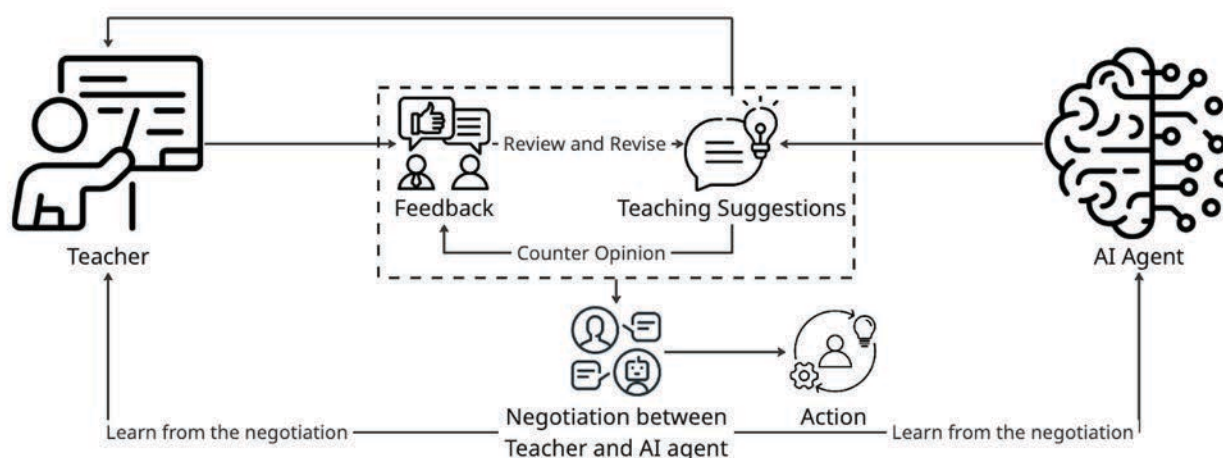
Source: Author's own work.

Augmentation through synergy

The main condition for augmentation in this sense is synergistic human-AI teaming, where human-AI combination yields emergent competence exceeding the maximum of individual AI or human competence alone at a given task.

Figure 7.8. The augmentation paradigm on the coordinates of teacher agency vs automation
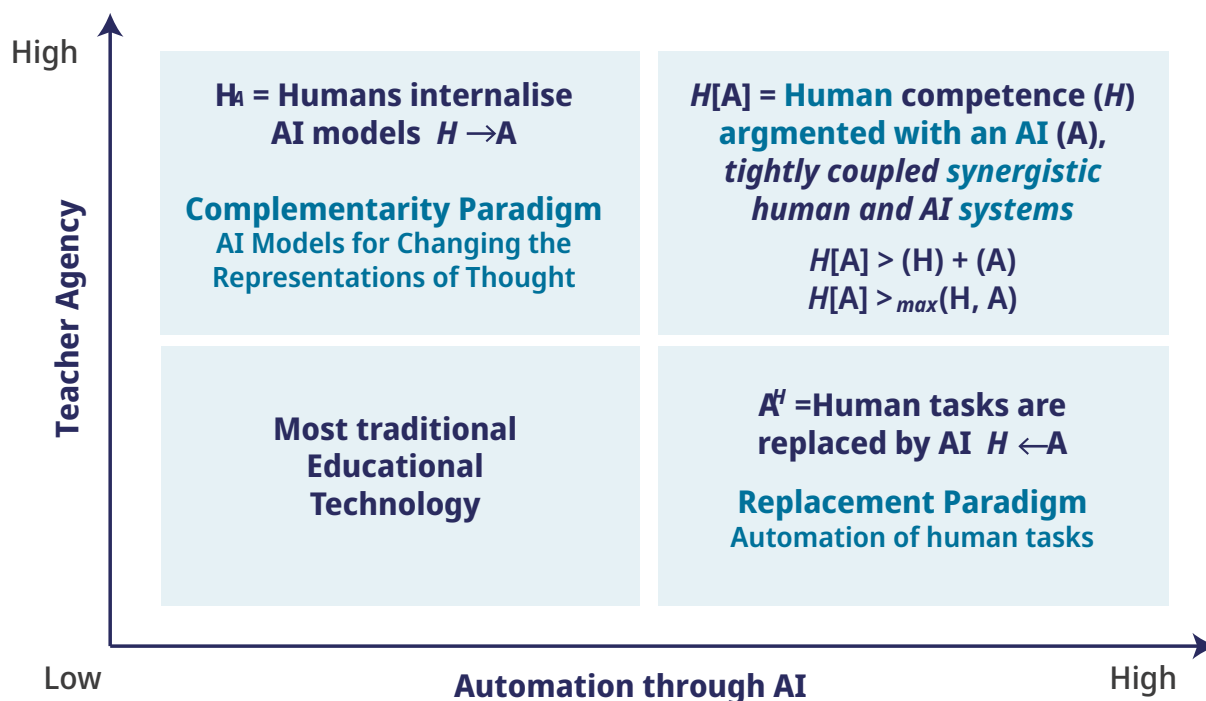
Source: Adapted from Cukurova, M. (2024[30]), *The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence.* British Journal of Educational Technology, 56(2), 469-488

However, this idea, informed by theories of distributed and extended cognition, faces significant empirical hurdles. A recent comprehensive meta-analysis of 106 experimental studies covering all sectors, with ~370 effect sizes reported, reveals that human-AI combinations underperform the better performer, human or AI alone, in 58% of the cases reviewed (Vaccaro, Almaatouq and Malone, 2024[37]), demonstrating that synergy is context-dependent and should not be assumed. Critically, synergy emerges only under specific conditions. For instance, task type matters. Synergy is more likely to occur in relatively simple content creation tasks (e.g. open-ended science education questions) but fails in complex decision-making tasks (e.g. classification of a student's emotional state given their relationship with other students). Second, when humans outperform AI alone, synergy is more likely to arise; yet when AI outperforms humans, human-AI teaming tends to degrade the performance. This indicates that synergy depends on human competencies to meta-cognitively assess when to trust AI input and when to ignore it which emphasises the importance of improving teachers and students' AI competencies to be able to protect their own agency and have a critical and informed trust in AI (see, for instance, UNESCO AI competency framework for teachers (Miao and Cukurova, 2024[11]). Third, the instructional design in which AI is interacted with is very important. The evidence from the meta-review indicates that only 3 out of 106 experiments explicitly tested predetermined subtask delegation between humans and AI to structure their interactions. These yielded non-significant synergy gains, underscoring that effective augmentation requires systematically scaffolded human-AI interactions embedded within a well-designed instructional framework.

These findings affirm that synergistic augmentation is difficult to achieve across many domains including healthcare, law, and education. They suggest that future AI implementations in education aiming to realise augmentation through synergistic interactions should prioritise task-specific AI design, development, and deployment, rather than relying on generic tools such as ChatGPT, which are not inherently educational, or human development, technologies. While such general-purpose systems can be valuable for exploratory or creative use, their outputs are not pedagogically grounded nor optimised for instructional alignment. Achieving meaningful synergy therefore requires the deliberate scaffolding of teacher-AI interactions within context-specific educational frameworks and the provision of AI competency training for teachers to help them unlock emergent forms of teacher-AI teaming. It is equally important to establish robust evaluation metrics that capture what specific aspects of teacher and AI competence are being augmented at a given task. Without these intentional design and assessment mechanisms, teacher-AI teaming is unlikely to achieve synergy and risk producing diminishing returns compared to standalone human or AI performance. The five-level framework proposed here examines the teacher agency through the lens of AI system affordances and interface-level design considerations. The system affordances discussed above therefore are necessary, yet not sufficient conditions for teacher-AI synergistic interactions. Box 7.1 presents examples for the five different AI-teacher teaming presented above.

Box 7.1. A case study example of teacher-AI teaming levels

In order to illustrate the spectrum of teacher-AI teaming in more concrete terms, here the levels are described through the case of a multimodal AI system designed to diagnose students' collaborative group challenges and group interactions in higher education contexts (Suraworachet, Zhou and Cukurova, 2025[38]). The purpose of this tool is to support teachers' situation-specific skills of monitoring students' group interactions, enabling more accurate diagnosis and timely interventions in students' group work activities.

1. **Transactional teaming**: At its most basic, the system functions as a machine-learning classification tool that automatically categorises students into high, medium, and low levels of collaborative problem-solving competence using multimodal sensor data (Spikol et al., 2018[39]). Here, the AI executes a discrete classification task in response to teacher data input of students' interaction logs, audio and videos, streamlining routine diagnostic work but not actively shaping pedagogical practice.

2. **Situational teaming**: At this level, the system integrates GenAI models to detect challenge dimensions from student dialogue (Suraworachet, Seon and Cukurova, 2024[40]) and computer vision techniques to analyse non-verbal group interactions (Suraworachet, Zhou and Cukurova, 2025[38]) and visualise insights from these

AI models on a teacher-facing dashboard. These insights create a shared awareness of classroom dynamics, aiming to complement teachers' perception of group processes and helping them recognise moments that may warrant intervention. These visualisations for instance can indicate students' active listening times in group activities, types of group interactions they are engaged in, and diagnosed challenges (i.e. cognitive, emotional, metacognitive or technical) as well as to what extent these challenges were resolved by the group.

3. **Operational teaming**: At this level, the system begins to co-operate with teachers in the planning and execution of instructional tasks by operationalising teacher-defined goals into actionable rules and parameters. Teachers specify priorities, such as fostering equitable participation or scaffolding metacognitive processes, and these are translated into threshold values, conditional rules, and automated prompts. For example, a teacher may configure the system so that if a student has not contributed during the problem-scoping phase for more than a specified number of minutes, the dashboard generates a nudge to encourage participation. Similarly, if the models detect that a group is struggling with task-specific metacognitive challenges, the system can intervene using GenAI created outputs by suggesting students revisit the activity description and clarify their immediate objectives. In this way, the dashboard functions as a co-orchestration partner, aligning its feedback and interventions with the teacher's declared pedagogical intentions while retaining the teacher's overarching authority and professional judgment.

4. **Praxical teaming**: At this level, the interaction between teachers and the system evolves into the development of shared practices, where AI system learns and adapt through sustained use. Instead of system generated GenAI feedback being delivered directly to students, the dashboard provides visualisations and tailored recommendations first to the teacher. Teachers then critically review these suggestions, revising or editing them, and may add further recommendations based on their professional judgment. These modifications are logged and processed by the AI system, which incorporates them into its knowledge base to iteratively refine its outputs. Over time, this process enables the AI to better converge with the expectations and pedagogical needs of the teacher. Nevertheless, the developmental trajectory of praxical teaming remains bounded by the teacher's existing expertise, as the AI primarily adapts to current practices rather than actively pushing teachers beyond their established competence. In contrast, synergistic teaming requires a deeper mutual transformation, where both the teacher and AI extend one another's capacities to co-create new practices and outcomes that surpass the competence of either agent alone.

5. **Synergistic teaming:** At this level, teacher-AI teaming evolves into dialogic co-construction of educational practice. An additional AI agent monitors teacher interactions with the diagnostic and feedback system, critically evaluating teacher-generated recommendations against learning sciences principles, historic interaction data, and prior teacher edits. Instead of passively accepting teacher revisions, the system engages the teacher in negotiation, prompting reconsideration of overlooked dimensions or alternative strategies. Through this dialogic exchange with a GenAI agent, both teacher and AI agents evaluate one another's claims and mutually extend their reasoning. When well aligned, this results in a form of creative resonance in which the teacher's pedagogical expertise and the AI's analytic and evaluative capacities intertwine to generate novel solutions, interventions, or instructional designs that neither could achieve alone. This is conceptualised as synergy, an emergent competence that exceeds the maximum of either human or AI performance individually.

Empirical evidence from recent studies underscores the tangible benefits of praxical and synergistic teacher-AI teaming in enhancing both teacher efficacy and student learning outcomes. For instance, in an auto grader research study, a praxical teacher-AI teaming approach was shown to reduce grading time by 44% while improving accuracy by 6% compared to manual grading (Liu and Liang, 2024[41]). Teaching assistants consistently rated the AI-assisted process as faster, easier, and more enjoyable, reporting that automation alleviated routine cognitive burdens and allowed greater focus on higher-order pedagogical reasoning. Similarly, a teacher-AI feedback co-creation study demonstrated that involving subject matter experts in GenAI-supported content authoring can yield comparable instructional quality to human-only materials, while dramatically reducing time and cognitive effort (Reza, Anastasopoulos and Pardos, 2024[42]). Involving ten mathematics experts and 358 learners, Reza et al. (2024[42]) show that an iterative human-AI approach to co-producing feedback reduced perceived workload by 50% and shortened the content development process from several months to a few hours,

while maintaining statistically significant learning gains for student. Together, these findings confirm the empirical potential of GenAI to support more advanced teaming approaches to amplify teacher productivity, improve instructional quality, and sustain learning outcomes when human oversight and agency remain central to system design.

However, both studies caution that augmentation is not without limitations. In the Liu et al. (2024[41]) study, AI performance degraded on unrepresented cases, revealing the dependence of system reliability on training data diversity and the need for ongoing human teacher verification. Similarly, participants in the Reza et al. (2024[42]) study reported occasional model unpredictability and difficulties in steering GenAI outputs, emphasising that the quality of augmentation depends on teachers' prompt literacy and capacity for meta-cognitive regulation. Moreover, lesson-specific variation in learning outcomes indicated that GenAI may underperform in certain pedagogical contexts or with particular learner profiles. Finally, neither of these studies measured improvements in teacher competence before and after their interactions with the tools to evaluate the impact of these interactions on teacher competence.

These findings collectively suggest that augmentation benefits are contingent upon interface design features, structured scaffolding of praxical and synergistic interactions, and the particular competence and motivation of teachers. Whatever the affordances of teacher-AI teaming might be, without sufficient teacher competence and motivation to engage meaningfully with these tools, the likelihood of achieving augmentation remains low. Furthermore, some teachers can indeed achieve competence gains even in their transactional teaming with AI tools, when these are used reflectively and purposefully. This also illustrates that the five levels of teacher-AI teaming are not hierarchically ordered stages of progression but rather context-sensitive modes of interaction. Different educational tasks, disciplinary demands, and institutional and local contexts may indeed benefit from different levels of teaming. Nonetheless, the design of teaming affordances that allow higher degrees of teacher agency (e.g. operational, praxical, and synergistic) represents a deliberate effort to maximise the likelihood that teacher-AI interaction contributes to further enhance teachers' agency and competence development. These higher-order forms of teaming are not inherently superior, but they are structured to provide richer opportunities for reflection, adaptation, and pedagogical transformation. In this sense, augmentation is not a property of technology alone but an emergent outcome of the dynamic interplay between teacher competence and motivation; the design of human-AI complementarity interface; and the affordances of the AI models.

Generative AI and teacher-AI teaming

GenAI marks a significant advancement regarding the AI affordances in the evolution of teacher-AI teaming levels by expanding the scope of complementarity across all levels of interaction. First of all, GenAI enables diversity and efficiency at a scale and precision previously unattainable. GenAI's capacity to generate multimodal outputs (e.g. textual, visual, auditory) allows teachers to access and adapt resources to diverse learner profiles, linguistic backgrounds, and learning needs. High performance of the state-of-the-art GenAI models in content generation creates unprecedented opportunities for teachers' transactional teaming with GenAI, allowing them to complete a wider range of tasks, from contextually appropriate lesson materials to content for students with certain impairments. Even generic GenAI systems (e.g. ChatGPT, Gemini, DeepSeek etc.), which are not designed specifically for education, can frequently support transactional and operational teaming due to their broad linguistic and creative affordances. For instance, teachers today commonly request ChatGPT to generate a quiz, summarise a text, or rephrase feedback comments. Although the interaction remains largely transactional, a request is made, and an output is returned (Handa et al., 2025[18]): they provide significant support on a wide range of content generation tasks which were simply not possible before GenAI. The teacher's agency in these interactions often lies in evaluating and adapting the GenAI response, but the cognitive exchange largely remains at the level of task automation and efficiency (Zhang et al., 2025[10]).

In situational and operational teaming, GenAI provides richer situational awareness by synthesising diverse streams of classroom data (e.g. text, speech, physiological signals, interaction logs, and visual cues) into interpretable insights that help teachers make informed, timely decisions. Unlike earlier traditional AI and analytics systems that offered static or unidimensional dashboards, GenAI can translate complex multimodal data signals into narrative explanations or alternative scenario projections, allowing teachers to be better informed about their classroom contexts through multiple perspectives. Furthermore, GenAI's language-based reasoning affordances enable it to articulate situational data processing and operational input to the model easier for co-creating plans, goals, and intentions in naturalistic dialogue rather than coding, or limited teacher authoring tools, which would be the case for traditional AI approaches.

More profoundly, the affordances of GenAI open unprecedented pathways towards praxical and synergistic teaming. Unlike traditional AI systems, which operate within fixed boundaries of prediction or classification, GenAI has the potential to engage in co-creative processes with teachers such as generating pedagogical hypotheses, critiquing lesson structures, suggesting conceptual analogies, and surfacing potential epistemic tensions in instructional design. Such a dialogic engagement, which is at the core of praxical and synergistic teaming levels, can be achieved with natural language using GenAI. GenAI also has capacity for adaptive, open-ended reasoning which can allow teachers to externalise, examine, and refine their pedagogical thinking in iterative cycles of reflection and critique. This recursive exchange has the potential to nurture professional growth, enabling teachers to question habitual practices and to explore alternative approaches that neither human expertise nor algorithmic optimisation alone could have revealed. Yet, we are only at the early stages of integrating GenAI meaningfully into teacher-facing AI tools, and the extent to which these systems can genuinely share the responsibility of meaning-making with teachers remains uncertain. It is not yet clear whether GenAI tools can co-create pedagogical meaning in a way that authentically challenges teachers' reasoning, provokes reflection, and contributes to deeper professional learning. While their dialogic affordances hold promise for more reciprocal exchanges, current implementations rarely demonstrate the capacity to push back against inappropriate or suboptimal pedagogical decisions, to question teachers' assumptions, or to propose alternative perspectives grounded in educational theory and evidence. Achieving such reflective tension, where AI systems not only assist but also constructively challenge teachers, still require advances in both the cognitive modelling of teaching expertise and the design of teacher-AI interaction interfaces, but GenAI systems provide unique opportunities to be explored in the upcoming years.

Conclusion

The future of GenAI in education will be determined not by how efficiently it helps automate tasks, but by how effectively it empowers teachers to exercise their professional judgment and expand their competence. The conceptual model proposed here (i.e. replacement, complementarity, augmentation) and the five levels of the teacher-AI teaming framework (transactional, situational, operational, praxical, and synergistic) provide a basis for policy and research to map existing teacher-facing AI systems and develop design principles to identify when GenAI systems replace, complement or augment teachers' competences. The five levels of teacher-AI teaming framework proposed here can be used to evaluate AI tools against explicit teacher-agency support criteria, and track dynamic changes from transactional to synergistic teaming interactions.

It is also important to acknowledge that since synergy requires mutual interactions of two-way information flow between AI and humans, one could also speak of the internalisation of human values into AI systems as part of this process. As teachers work closely with AI, ideally, they would influence the design and tuning of these systems (e.g. through feedback, pushback on recommendations and reasons provided for such pushback, usage choices, participatory design, etc.). Over time, AI used in classrooms should learn the priorities of good teaching for a given teacher dependent upon their pedagogical philosophy, needs and requirements (e.g. promoting inquiry, not giving away answers too easily, respecting diverse solutions, prioritising certain instructional choices over others etc.) because teachers enforce these in their interactions with the AI. In a sense, the AI system learns some of the teacher's pedagogical wisdom through these synergistic interactions. Thus, augmentation would also require appropriate spaces for the reflection of humans and space for model updates for AI. That is, teachers reflecting on how AI is affecting their practice and their students' learning, AI using logged interaction data to update its model parameters and weighs to learn from its interactions with teachers. This two-way internalisation is at the heart of human-AI co-evolution in education for synergistic augmentation; teachers shape AI just as AI shapes teaching practices synergistically.

The synergistic augmentation paradigm represents the deepest form of teacher-AI interaction, where AI is no longer a distinct and complementary "add-on" to their competence, but part of the fabric of teaching and learning. This holds the promise of truly hybrid intelligence pedagogies that leverage AI's capabilities while being guided by human wisdom in synergistic interactions that lead to an augmented competence that is greater than the maximum of an individual human or AI alone. It demands high levels of teacher agency to negotiate the human-AI relationship. In addition, teachers must remain self-aware, continually asserting human values and expertise in the loop even as they embrace new AI-empowered methods. Achieving success at this augmentation paradigm is challenging in education, but it aligns with the vision of education where AI serves as a competence augmentation tool for teachers, ultimately enriching the teaching and learning experiences.

Acknowledgements

References

Aslan, S., N. Alyuz, C. Tanriover, S. Mete, E. Okur, S. D'Mello, and A. Arslan Esme (2019), "Investigating the Impact of a Real-time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-12, https://doi.org/10.1145/3290605.3300534.   [6]

Azevedo, R., J. Cromley and D. Seibert (2004), "Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia?", *Contemporary Educational Psychology*, Vol. 29/3, pp. 344-370, https://doi.org/10.1016/j.cedpsych.2003.09.002.   [23]

Berendt, B., A. Littlejohn and M. Blakemore (2020), "AI in education: learner choice and fundamental rights", *Learning, Media and Technology*, Vol. 45/3, pp. 312-324, https://doi.org/10.1080/17439884.2020.1786399.   [12]

Bernius, J., S. Krusche and B. Bruegge (2021), "A Machine Learning Approach for Suggesting Feedback in Textual Exercises in Large Courses", *Proceedings of the Eighth ACM Conference on Learning @ Scale*, pp. 173-182, https://doi.org/10.1145/3430895.3460135.   [36]

Blömeke, S., J. Gustafsson and R. Shavelson (2015), "Beyond Dichotomies", *Zeitschrift für Psychologie*, Vol. 223/1, pp. 3-13, https://doi.org/10.1027/2151-2604/a000194.   [43]

Bosch, N., S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao (2016), D*etecting student emotions in computer-enabled classrooms*, https://pnigel.com/papers/bosch-pp-ijcai16-camera.pdf.   [21]

Calvo, R. and S. D'Mello (2010), "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications", *IEEE Transactions on Affective Computing*, Vol. 1/1, pp. 18-37, https://doi.org/10.1109/t-affc.2010.1.   [22]

Chan, K., F. Ali, J. Park, K. Sham, E. Tan, F. Chong, K. Qian and G. Sze (2025), "Automatic item generation in various STEM subjects using large language model prompting", *Computers and Education: Artificial Intelligence*, Vol. 8, https://doi.org/10.1016/j.caeai.2024.100344.   [8]

Cukurova, M. (2024), "The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence", *British Journal of Educational Technology*, Vol. 56/2, pp. 469-488, https://doi.org/10.1111/bjet.13514.   [30]

Demszky, D., J. Liu, H. Hill, S. Sanghi and A. Chung (2025), "Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement", *Computers & Education*, Vol. 227, p. 105-183, https://doi.org/10.1016/j.compedu.2024.105183.   [5]

Department for Education UK (2024), *Generative AI in education: Educator and expert views*, https://www.gov.uk/government/publications/generative-ai-in-education-educator-and-expert-views.   [16]

Department for Education UK (2024), *Generative AI in education: User research and technical report*, https://www.gov.uk/government/publications/generative-ai-in-education-user-research-and-technical-report.   [17]

Department for Education UK (2023), *Generative AI in education: Call for evidence - summary of responses*, https://assets.publishing.service.gov.uk/media/65609be50c7ec8000d95bddd/Generative_AI_call_for_evidence_summary_of_responses.pdf.   [15]

Di Mitri, D., J. Schneider and H. Drachsler (2021), "Keep Me in the Loop: Real-Time Feedback with Multimodal Data", *International Journal of Artificial Intelligence in Education*, Vol. 32/4, pp. 1093-1118, https://doi.org/10.1007/s40593-021-00281-z.   [4]

Felix, J. and L. Webb (2024), *Use of artificial intelligence in education delivery and assessment*, https://aiadvisoryboards.wordpress.com/wp-content/uploads/2024/01/use-of-artificial-intelligence.pdf.   [14]

Frostenson, M. (2015), "Three forms of professional autonomy: de-professionalisation of teachers in a new light", *Nordic Journal of Studies in Educational Policy*, Vol. 2015/2, https://doi.org/10.3402/nstep.v1.28464. [27]

Guan, L., Y. Zhang and M. Gu (2025), "Pre-service teachers preparedness for AI-integrated education: An investigation from perceptions, capabilities, and teachers' identity changes", *Computers and Education: Artificial Intelligence*, Vol. 8, https://doi.org/10.1016/j.caeai.2024.100341. [25]

Handa, K., A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. Troy, D. Amodei, J. Kaplan, J. Clark and D. Ganguli (2025), *Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations*. [18]

Holmes, W., K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. Shum, O. Santos, M. Rodrigo, M. Cukurova, I. Bittencourt and K. Koedinger (2021), "Ethics of AI in Education: Towards a Community-Wide Framework", *International Journal of Artificial Intelligence in Education*, Vol. 32/3, pp. 504-526, https://doi.org/10.1007/s40593-021-00239-1. [13]

Kosmyna, N. and E. Hauptmann (2025), *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task.* [32]

Laru, J. and S. Järvelä (2015), "Integrated Use of Multiple Social Software Tools and Face-to-Face Activities to Support Self-Regulated Learning: A Case Study in a Higher Education Context", in *Seamless Learning in the Age of Mobile Connectivity*, Springer Singapore, Singapore, https://doi.org/10.1007/978-981-287-113-8_24. [24]

Leiker, D., S. Finnigan and A. Cukurova (2023), *Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale*, https://doi.org/10.48550/arXiv.2306.01815. [28]

Li, H., J. Gobert and R. Dickler (2019), "Evaluating the Transfer of Scaffolded Inquiry: What Sticks and Does It Last?", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-23207-8_31. [19]

Liu, E. and C. Liang (2024), *Breakout grader: Reinforcement learning for scalable human-in-the-loop grading of interactive programming assignments*, https://ezliu.github.io/breakoutgrader. [41]

Miao, F. and M. Cukurova (2024), *AI competency framework for teachers*, UNESCO, https://doi.org/10.54675/zjte2084. [11]

Minn, S. (2022), "Minn, S. (2022), "AI-assisted knowledge assessment techniques for adaptive learning environments", *Computers and Education: Artificial Intelligence*, Vol. 100050. [7]

Pea, R., P. Biernacki, M. Bigman, K. Boles, R. Coelho, V. Docherty, J. Garcia, V. Lin, J. Nguyen, D. Pimentel, R. Pozos, B. Reynante, E. Roy, E. Southerton, M. Suzara and A. Vishwanath (2022), "Four Surveillance Technologies Creating Challenges for Education", in *AI in Learning: Designing the Future*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-031-09687-7_19. [29]

Reza, M., I. Anastasopoulos and S. Pardos (2024), *PromptHive: Bringing subject matter experts back to the forefront with collaborative prompt engineering for educational content creation*, https://arxiv.org/abs/2410.16547. [42]

Roy, P., H. Poet, R. Staunton, K. Aston and D. Thomas (2024), *ChatGPT in Lesson Preparation: A Teacher Choices Trial*, http://www.mmiweb.org.uk/gaiandre/literature/EEF.pdf. [1]

Sawaya, S., J. Jacobs, R. Moulder, C. Chandler, B. Milne, T. Fischaber and S. D'Mello (2025), "Improving Tutor Discourse Practices via AI-Enhanced Coaching: A Piecewise Latent Growth Curve Modeling Approach", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-98459-4_6. [34]

Selwyn, N., M. Ljungqvist and A. Sonesson (2025), "When the prompting stops: exploring teachers' work around the educational frailties of generative AI tools", *Learning, Media and Technology*, Vol. 50/3, pp. 310-323, https://doi.org/10.1080/17439884.2025.2537959. [26]

Seßler, K., M. Fürstenberg, B. Bühler and E. Kasneci (2025), "Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring", *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 462-472, https://doi.org/10.1145/3706468.3706527. [9]

Spikol, D., E. Ruffaldi, G. Dabisias and M. Cukurova (2018), "Supervised machine learning in multimodal learning analytics for estimating success in project-based learning", *Journal of Computer Assisted Learning*, Vol. 34/4, pp. 366-377, https://doi.org/10.1111/jcal.12263. [39]

Srivastava, S., A. Varshney, S. Katyal, R. Kaur and V. Gaur (2021), "A smart learning assistance tool for inclusive education", *Journal of Intelligent and Fuzzy Systems*, Vol. 40/6, pp. 11981-11994, https://doi.org/10.3233/jifs-210075. [33]

Suraworachet, W., J. Seon and M. Cukurova (2024), "Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches", *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 473-485, https://doi.org/10.1145/3636555.3636905.   [40]

Suraworachet, W., Q. Zhou and M. Cukurova (2025), "University Students' Perceptions of a Multimodal AI System for Real-World Collaboration Analytics: Lessons Learned From a Case Study", *Journal of Computer Assisted Learning*, Vol. 41/5, https://doi.org/10.1111/jcal.70103.   [38]

UNESCO (2023), *Global Education Monitoring Report 2023: Technology in education: A tool on whose terms?*, GEM Report UNESCO, https://doi.org/10.54676/uzqv8501.   [31]

Vaccaro, M., A. Almaatouq and T. Malone (2024), "When combinations of humans and AI are useful: A systematic review and meta-analysis", *Nature Human Behaviour*, Vol. 8/12, pp. 2293-2303, https://doi.org/10.1038/s41562-024-02024-1.   [37]

van den Berg, G. and E. du Plessis (2023), "ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education", *Education Sciences*, Vol. 13/10, p. 998, https://doi.org/10.3390/educsci13100998.   [3]

Van Lehn, K. (2011), "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems", *Educational Psychologist*, Vol. 46/4, pp. 197-221, https://doi.org/10.1080/00461520.2011.611369.   [2]

VanLehn, K., C. Banerjee, F. Milner and J. Wetzel (2020), "Teaching Algebraic Model Construction: A Tutoring System, Lessons Learned and an Evaluation", *International Journal of Artificial Intelligence in Education*, Vol. 30/3, pp. 459-480, https://doi.org/10.1007/s40593-020-00205-3.   [20]

Yang, K., V. Echeverria, Z. Lu, H. Mao, K. Holstein, N. Rummel and V. Aleven (2023), "Pair-Up: Prototyping Human-AI Co-orchestration of Dynamic Transitions between Individual and Collaborative Learning in the Classroom", *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* pp. 1-17, https://doi.org/10.1145/3544548.3581398.   [35]

Zhang, A., Y. Gao, W. Suraworachet, T. Nazaretsky and M. Cukurova (2025), *Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students*, https://doi.org/10.48550/arXiv.2504.10961.   [10]

8 Transitioning from general-purpose to educational-oriented Generative AI:

Maintaining teacher autonomy

Paraskevi Topali*, Alejandro Ortega-Arranz**, Inge Molenaar*

*NOLAI, National Education Lab AI, Radboud University, the Netherlands

**Universidad de Valladolid, Spain

This chapter proposes a teacher-centred approach for transitioning from general-purpose generative Artificial Intelligence (GenAI) to education-oriented AI that emphasises contextual relevance, pedagogical grounding, and professional autonomy. Drawing on Human-Centred Design principles and Molenaar's Human-AI Automation Model, the chapter conceptualises two levels of teacher autonomy. It then presents insights from the life cycle of ideating and co-creating a GenAI prototype that supports instructional design and enactment, illustrating how teacher involvement can support the development of AI tools that serve pedagogical goals and real classroom needs.

Introduction

Advancements in Artificial Intelligence (AI) are impacting everyday life in a variety of domains, including education, healthcare and industry (Holmes and Tuomi, 2022[1]; Varghese and Chapiro, 2024[2]). In the educational sector, the current and rapid (r)evolution of generative AI (GenAI) promises more accurate and personalised learning support, automation of some teaching routines and augmentation of teachers' pedagogical actions (Holmes et al., 2022[3]; Zhai et al., 2021[4]). At the same time, the use of AI tools involve potential risks to human-led teaching, such as lack of human interaction between students and teachers, threats to teacher autonomy or data privacy and ethical issues (Baidoo-Anu and Owusu Ansah, 2023[5]; Mørch and Andersen, 2023[6]).

Against this backdrop, it is important to distinguish between general-purpose AI tools and those AI tools specifically designed for education (European Commission, 2025[7]). Educational AI tools, such as intelligent tutoring systems or adaptive learning technologies, are developed with the explicit aim of supporting teaching and learning, offering functions like personalised tutoring, automated feedback, and curriculum-aligned content generation. On the other hand, general-purpose AI, often using GenAI-powered tools as chatbots and conversational agents, addresses a wide range of tasks across domains. This distinction underscores that while the responsible integration of AI into education should be approached with care, particular attention is needed when adopting general-purpose tools, as they are not designed with pedagogical goals in mind (Molenaar et al., 2025[8]).

General-purpose AI, such as ChatGPT or Claude, has increasingly found its way into educational settings. Yet, as general-purpose tools they are not designed for educational use. Significant concerns with these tools are their lack of:

1. connection with the educational context and curriculum
2. grounding in pedagogical and learning theories
3. personalisation to learners' needs
4. support of teacher autonomy.

Connecting the characteristics of the educational context (e.g. delivery setting, learning objectives and activities, teaching and assessment methods) with the technology is critical for ensuring that teaching genuinely supports learning. These characteristics are reflected in both instructional and learning design processes. Instructional design provides systematic guidance for translating curriculum and subject knowledge into coherent learning pathways and aligning objectives, methods, and assessments (Reiser, 2001[9]). Learning design, in turn, emphasises the creation of meaningful learning experiences grounded in pedagogical theory, accounting for prior student knowledge, motivation, collaboration, and the multiple ways learners interact with content, teachers, and peers (Laurillard, 2013[10]). However, general-purpose AI is not sensitive to differences in course structures, teaching methods, evaluation procedures, and content curricula, risking that the AI output can be irrelevant or confusing. For example, a history teacher may struggle to ensure that GenAI-generated explanations follow the national curriculum's emphasis on critical source analysis, while a mathematics teacher may find that GenAI responses skip over foundational concepts needed by students who lack prerequisite knowledge (Mørch and Andersen, 2023[6]).

Effective learning also relies on pedagogical principles that guide how knowledge is delivered, practiced, and assessed. General-purpose AI tools lack such grounding and typically cannot be configured to match with the pedagogical intentions of a course (e.g. scaffolding knowledge step by step, fostering inquiry-based learning). As a result, they produce content that may seem accurate or plausible but does not necessarily promote deep understanding or meaningful learning. While GenAI tools excel at providing immediate, direct answers to user prompts, the learning process itself often benefits from strategically delayed and less direct feedback, since meaningful learning is rooted in inquiry, not just in receiving the correct answers. Moreover, general purpose AI offers limited personalisation and adaptation to learners' individual needs. While GenAI may appear adaptive, it typically lacks genuine learner modelling, which takes account of prior knowledge or enables appropriate responses to affective, motivational, or cognitive differences among students. Finally, general-purpose AI challenges teacher autonomy over the educational triangle that connects the teacher, the learning environment, and the student. That is, by mediating interactions within this triadic relationship, AI risks shifting the balance of agency away from teachers and learners toward algorithmic systems whose operations and decision-making processes are largely opaque. This can lead to a reconfiguration of pedagogical control, where instructional decisions, feedback, and even the pacing of learning lacks grounding in human interpretation, empathy, and contextual understanding. Such a shift raises critical questions about autonomy, accountability, and the preservation of human-centred education in technologically mediated learning environments.

Building on the abovementioned context, we raise a critical question for research and practice: *How can general-purpose AI tools be effectively transitioned into educational AI that connects to the educational context in classrooms, aligns with pedagogical theories, supports personalisation to learners' needs and supports teachers' autonomy?* Addressing this question is key to ensuring that AI not only generates content but also meaningfully enhances teaching and learning. Prior research has stressed the importance of Human-Centred Design (HCD) in the co-creation of technological solutions to achieve a careful consideration of instructional and learning design aspects, learning theories and teacher perspectives. These approaches actively position the educational stakeholders (e.g. teachers, curriculum designers) as co-design partners to achieve a synergy between their needs, the technological innovations and the pedagogical context (Buckingham Shum, Ferguson and Martinez-Maldonado, 2019[11]; Dimitriadis, Martínez-Maldonado and Wiley, 2020[12]). Nevertheless, existing literature reviews on Human-Centred Design in AI note that its adoption in actual cases is still scarce and mainly regards stakeholder involvement in initial brainstorming on the users' needs rather than their actual participation in the design and development of tools (Topali et al., 2025[13]; Zawacki-Richter et al., 2019[14]).

We discuss the use of Human-Centred Design to enhance teacher autonomy, i.e. the degree of professional freedom granted to teachers to make decisions about learning, assessment and the tools available to mediate these processes (Molenaar, 2021[15]). Within this frame we explore how to enhance teacher autonomy via Human-Centred Design with teachers participating as:

1. Co-design partners during the design of an AI tool (e.g. interface, tool features, infrastructure) expected to be integrated for a particular educational purpose

2. Co-orchestration partners where the teachers plan, execute and reflect on each lesson and apply the tool to their own teaching by deciding, for example, how the educational tasks can be divided between the teacher and the AI.

This chapter introduces a teacher-centred approach to the transition from generative AI as a general-purpose tool to educational AI, and stresses the importance of maintaining connection to the educational context, pedagogical objectives and autonomy of teachers during this transition. The chapter draws upon the Human-AI Automation Model proposed by Molenaar (Molenaar, 2021[15]) to examine how teachers can be meaningfully positioned as co-actors in the design and integration of GenAI technologies. Building upon this theoretical foundation and the limited evidence on Human-Centred Design in AI, we offer empirical insights into the development lifecycle of educational GenAI systems, with teachers as co-designers, following Human-Centred Design principles (i.e. gathering initial teachers' needs, eliciting co-design requirements and resulting in co-development). We conceptualised and designed a prototype to support teacher instructional/learning design (e.g. feedback design) and classroom enactment (e.g. capturing student-generated GenAI analytics), according to teacher needs and the current limitations of GenAI tools. Our takeaways aim to highlight the key role of educators as co-design partners to ensure that GenAI tools support pedagogical goals, classroom needs and teacher autonomy during the learning process.

Related work

Within the landscape of Technology-Enhanced Learning, researchers stress the need to adopt Human-Centred Design approaches in the design of technological innovations, so that the end-products meet the user needs (e.g. teachers, students) (Giacomin, 2014[16]). Human-Centred Design places people's needs, values, and rights at the core of digital design (Shneiderman, 2020[17]). As far as educational-oriented AI is concerned, Human-Centred Design views AI-driven tools as a means of empowering learners and educators, supporting efficiency, active learning, critical thinking, and creativity (Topali et al., 2025[13]). The approach also offers frameworks, such as Value Sensitive Design, that explicitly integrate human values into technology design (Friedman et al., 2013[18]). Yet, there are limited examples of adopting Human-Centred Design in the design of AI and GenAI solutions (Topali et al., 2025[13]). Examples include studies by Holstein, McLaren, and Aleven (2019[19]; 2019[20]), who positioned teachers as co-designers of a wearable AI tool to augment students' monitoring in K-12 classrooms. Likewise, Lister et al. (2021[21]) followed a participatory approach to design a virtual agent with visual-disabled students assisting them within the context of distance learning.

When it comes to general-purpose GenAI, there are a few studies that implemented Human-Centred Design at the initial design stage by understanding stakeholders' perspectives in K-12 education. For instance, Han et al. (2024[22]) interviewed primary school teachers and students, and found that GenAI could be beneficial in personalising the learning experiences and providing instant feedback to them. The authors indicated concerns over data authorship, lack of critical thinking in the case of hallucinations and students' and teachers' autonomy. Likewise, Hays, Jurkowski, and Sims (2023[23]), Kaplan-Rakowski et al. (2023[24]), Laak and Aru (2024[25]) and Monteiro et al. (2024[26]) shed light on teachers' viewpoints on the use of ChatGPT for educational purposes. In most studies, teachers reported that ChatGPT might be beneficial for students, but they regretted their lack of control on the answers given, and had concerns about the privacy of the data and who can analyse it and about the lack of contextualisation. For instance, Prestridge, Fry, and Kim (2024[27]) interviewed ten secondary school teachers to understand the potential added value of GenAI in their courses and they stressed the importance of considering the different course contexts to use GenAI meaningfully. These studies also discussed the need for teachers' professional development, both to provide guidance on how to use GenAI for education and as a means to address feelings of fear and replacement by AI.

The above studies provide important insights about teacher and student needs regarding the use of GenAI for teaching and learning, such as the importance of teacher control over the GenAI generated learning content. However, most studies simply focus on teachers' general perceptions about the use of only ChatGPT via surveys with Likert scale items, or short-answer formats (e.g. Hays, Jurkowski and Sims, 2023[23]; Monteiro et al., 2024[26]). They do not extract design guidelines for the development of educational-oriented GenAI tools and do not position the stakeholders as co-partners of such development and integration processes. One exception is the study by Han et al. (2024[22]), which limits its focus to a GenAI writing tool for primary education. Table 8.1 summarises the current state of the literature in this regard. In all studies, to the best of our knowledge, there is a lack of evidence that Human-Centred Design processes informed the design of a GenAI system based on participants' requirements and actively involved them as co-designers.

Table 8.1. Characteristics of studies applying Human-Centred Design principles in the design of GenAI-based solutions

Reference	Involve teachers	Involve students	Educational level	Study current GenAI uses	Data collection method	Generate design guidelines
Kaplan-Rakowski et al. (2023[24])	Yes	No	K-12	No	Survey	No
Han et al. (2024[22])	Yes	Yes	Primary	No	Interviews/ Workshop	Yes
Monteiro et al. (2024[26])	Yes	No	Secondary	No	Survey	No
Hays et al. (2023[23])	Yes	No	K-12	No	Survey	No
Laak and Aru (2024[25])	Yes	No	K-12	Yes	Survey	No
Prestridge et al. (2024[27])	Yes	No	Secondary	Yes	Interviews	No

Source: Author's own work.

Autonomy vs. automation in AI-driven educational tools

Given the limited application of the full cycle of Human-Centred Design in general-purpose GenAI, we deem that GenAI tools designed without teachers' pedagogical considerations may inadvertently hinder their autonomy, since teachers have to deal with AI-generated outputs that do not align with their pedagogical intentions. Indeed, prior research indicated that while ChatGPT and other GenAI tools are nowadays used by most students in higher education (Xiaoyu, Zainuddin and Hai Leng, 2025[28]), they limit teacher control during classroom practices (Chapwanya, 2025[29]). For instance, if a teacher proposes that their students use ChatGPT to help create an essay, the teacher may not be able to monitor what the students are asking the chatbot, what feedback is given to them and/or when such prompts or answers need to be corrected, thus leaving aside teacher knowledge/expertise from the learning situation.

Although initially introduced to discuss how educational AI affects teacher control, the Human-AI automation model (Molenaar, 2022[30]) can also be applied to understand different levels of automation in learning situations using GenAI systems. This model articulates the transition of control between teacher and intelligent technology through six levels of automation during the teaching and learning practice (Figure 8.1), i.e. from 'teacher has full control' (left), to 'technology has full control' (right). Projecting the model into GenAI usage, in the second level (Teacher Assistance), we have cases where teachers have full control over the learning situation (e.g. teaching methods, feedback on course assignments) and GenAI tools propose additional information, explanations, examples, or text snippets for the teacher to implement in their lessons, improving teachers' existing instructional design. In the third level of the model (*Partial Automation*), teachers give part control of specific tasks to GenAI. For instance, GenAI highlights common student errors and provides an overview to students highlighting which errors they have made. The teacher can then use this list in a class discussion of errors and elaborate on how to resolve them. In the next levels of the model, advancing towards the right side, the AI takes almost full control largely acting independently. For instance, GenAI tools such as platforms like Synthesis Tutor[1] where GenAI generates content, feedback, and assessments dynamically for each student, without expecting any teacher intervention.

An apparent limitation of this model is that it does not integrate the entire instructional cycle in which teachers plan, execute, reflect on, and revise their lessons. The model focuses mainly on actions performed during teaching, while it is equally important to consider teacher autonomy and control during course design and refinement. In the planning phase, critical instructional and learning design decisions are taken about sequencing, scaffolding, and alignment with learning objectives, pedagogical tasks and teaching methods. In the refinement process, teachers continually improve their lessons.

Figure 8.1. Human-AI automation model

Teacher controls	Teacher has full control	Teacher monitors technolodgy	Teacher monitors incidentally, but can resume control at all time	Teacher control and monitoring is not required for specific tasks	
	Technology provides supportive information	Technology controls specific tasks	Technology signals when teacher control is needed Technology controls broader set of tasks	Technology requests teacher control Technology controls most tasks automatically	Technology contols all tasks automatically
TEACHER ONLY	**TEACHER ASSISTANCE**	**PARTIAL AUTOMATION**	**CONDITIONAL AUTOMATION**	**HIGH AUTOMATION**	**FULL AUTOMATION**

Anne Horvers and Inge Molenaar, Adaptive Learning Lab

Source: Molenaar, (2022[30]), Towards hybrid human-AI learning technologies", *European Journal of Education*, Vol. 57/4, pp. 632-645, https://doi.org/10.1111/ejed.12527.

Let us consider the case of a GenAI-feedback tool: during the instructional cycle, the teacher specifies the assessment criteria, the type of feedback that should be emphasised (e.g. formative comments on argumentation rather than grammar), and the depth or timing of responses that are appropriate for their students. Based on this instructional plan, the functioning of the GenAI can be refined. Now the general-purpose AI tool is embedded in an instructional cycle, which predetermines how the GenAI tool operates. This allows for a division of labour in which the teacher retains complete control over the instructional plan and functioning of the GenAI, while the GenAI supports the teacher with high automation during the enactment of the lesson, generating feedback for student submissions in real time without the teacher's direct intervention. In this scenario, teacher autonomy is exercised during the orchestration cycle of the instruction (in this context, orchestration refers to the general management of the learning activities). This means that the teachers can adjust the system requirements of the GenAI tool given their educational context to ensure that they continue to determine the pedagogical framing of the feedback, while automation during enactment reduces teacher workload and ensures consistency.

Drawing upon these reflections, the Human-AI Automation model presented above can be extended to understand teachers' autonomy within GenAI-powered systems with respect to design and orchestration decisions. Teacher autonomy is not only exercised through control of AI during the course enactment, but also *during the course design and instructional cycle.* Accordingly, we conceptualise two different levels of autonomy where teachers act as: 1) co-design partners during the AI tool design (e.g. voicing their needs regarding the interface, tool features, and tool infrastructure) and 2) co-orchestration partners in instruction cycle where they plan, execute and reflect on each lesson and apply the tool to their own teaching, deciding for example how the educational tasks will be divided between the teacher and the AI.

The development of a GenAI prototype using design-based research and participatory approaches

Reflecting on these different levels of autonomy for teachers and the constraints in the general-purpose AI tools, we aimed to explore how we can support the transition from general AI technologies into education-oriented tools. The study presented in this chapter is part of a project addressing the following research question: *To what extent can we enhance teacher autonomy by positioning them as co-design tool partners in the transition from general-purpose AI tools into educational AI?* To answer the research question, we followed Human-Centred Design, by giving teachers the possibility to voice their needs when it comes to GenAI use, as well as their pedagogical goals and classroom practices. This emphasis on teacher input supports the development of GenAI as a co-design partner rather than a one-size-fits-all assistant.

This section first presents the process of understanding teachers' and students' needs in GenAI usage and then the ideation and low-fidelity development of a GenAI chatbot with extended functionality to foster teacher autonomy for secondary education. To guarantee that the GenAI system satisfies the learning and teaching needs, we followed a Design-based Research (DBR) methodological approach (Amiel, 2008[31]) (Figure 8.2). Accordingly, we further divided the research question into the following sub-questions that were addressed in each design-based research phase (see Figure 8.2):

1. How do teachers and students use GenAI?
2. What limitations and challenges do teachers and students face when using GenAI?
3. How do teachers perceive the requirements for GenAI tools identified from prior interviews?
4. How do teachers perceive the added value and potential adoption of the proposed GenAI tool features in their educational contexts?

Figure 8.2. The design-based research process
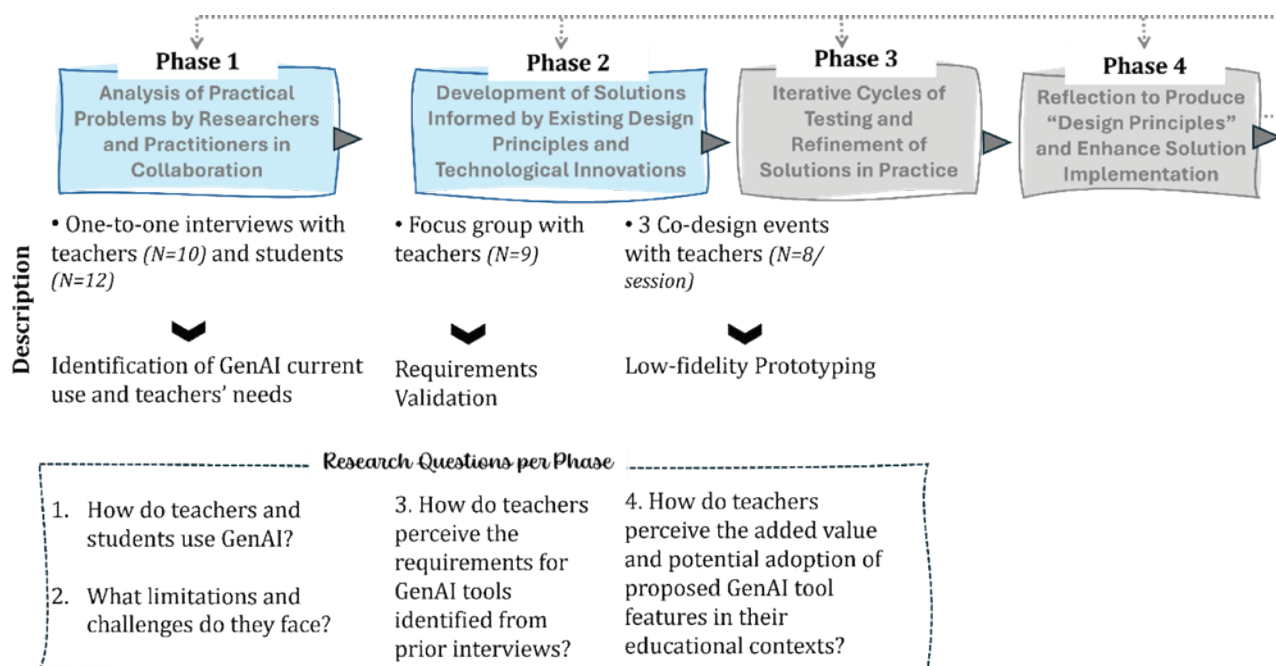
Note: The blue colour depicts the steps already taken
Source: Authors' own work.

Our research design followed a qualitative phenomenological process (Creswell, 2013[32]) to get from: 1) general problem understanding on GenAI usage (questions 1 and 2, Figure 8.2) to 2) requirements' validations of teachers' main needs (question 3, Figure 8.2) and 3) low-fidelity prototyping of a GenAI system satisfying such needs (question 4, Figure 8.3).

During Phase 1, we conducted a set of face-to-face, semi-structured, one-to-one interviews with ten secondary school teachers and 12 students. In Phase 2, we conducted a focus group with nine teachers to reflect upon teachers' requirements (as collected from Phase 1) for using GenAI for secondary education. These data collection approaches were chosen to help us understand in-depth teachers' and students' needs when using GenAI. Lastly, in Phase 2 we also conducted three co-design events with eight teachers, each time working with the requirements gathered from the focus group towards the design of a low-fidelity prototype of a GenAI system (see Table 8.4). During Phase 2, teachers worked with three different scenarios for the prototype addressing different teaching moments (i.e. course design, course enactment, after course reflection and assessment) taking into account the complexity and nature of the teaching process.

Prerequisites for participants to participate in this study were the previous use of GenAI tools for formal teaching and learning purposes. Teachers often face difficulties in connecting the course learning design and their teaching needs with the desired data-driven information about the student (Mangaroska and Giannakos, 2019[33]). Thus, we employed a set of techniques to better guide our teachers in their role as co-designers and support them during this process. For example, we conducted interviews and a focus group to understand teachers' current use of the existing GenAI tools. Additionally, we followed the 'superpower' approach proposed by Holstein, McLaren and Aleven (2019[19]; 2019[20]), asking teachers about the 'superpowers' that an ideal GenAI tool could support. During the co-design events we used "speed-dating" processes and prototype simulation exercises, to discuss the use of GenAI in relation to the actual learning scenarios of the teachers.

The data sources employed in the interviews were the stakeholders' recordings about the perceived GenAI challenges, added value and actual use cases, answers to a profiling questionnaire and artifacts (post-it notes with additional ideas). In the focus group, the data sources regarded mainly a profiling questionnaire and the generated artifacts (post-it notes with challenges and superpowers). Content analysis of the collected data was employed utilising inductive coding, i.e. categories derived from participants' answers. For instance, we extracted the following categories based on participant answers on their current use of GenAI:

1. GenAI for replacing current learning and teaching tasks [Replacement]
2. GenAI for complementing current learning and teaching tasks [Complementarity]
3. GenAI for supporting learning both for teachers and students [Learning].

Information about the stakeholder groups and their characteristics is presented in Annex 8.A.

Identification of current GenAI use and teacher needs [phase 1]

Teachers

At the initial interviews, the majority of the teachers (N=8) reported ChatGPT as their main GenAI tool and two teachers reported using other tools such as Snippet for code generation or Microsoft Copilot, which they found more accurate than ChatGPT. Half of the teachers (N=5) used these tools for course design (e.g. for outlining the course structure), and the other half during course enactment, either asking students to use it to conduct learning assignments or as annotation tool used by the teachers to comment and reflect upon student answers. Specifically, most teachers (N=8) reported using GenAI tools for replacement purposes (see Table 8.1, category "replacement", A and B) to conduct orchestration tasks, such as the creation of learning materials, which previously had to be done manually. Several teachers use GenAI tools to conduct educational tasks more effectively (N=3) like introducing GenAI tools as learning options for students to support their learning activities (see Table 8.1, category "complementarity", A and B). Finally, some teachers (N=3) use GenAI as a learning tool to enhance their teaching methods (see Table 8.1, category "Learning", A and B). The teacher group thus represented a variety of views.

Students

Most students (N=9) reported ChatGPT as the main GenAI tool used, and three of them reported the use of other tools, such as Wombo Dream and DALL-E, for generating photos and artwork; or Deep AI and Microsoft Copilot for text generation. Also, most students (N=9) used GenAI at home for assignment preparation, while only three used GenAI either at home or at school following the teacher's recommendations and guidance. The use of these tools was related to a wide variety of topics: History (N=4), Computer Science (N=2), Geography (N=1), Economics (N=1), Literature (N=1), English (N=1), Sex Education (N=1), Physics (N=1). Unlike the teachers, most students described using GenAI as a resource for gathering information for their assignments (N=8, Table 8.2 [Complementarity] C, D).

In several cases GenAI was found to act as a partner for students to test their knowledge, to provide explanations or to practice and improve their writing skills (N=4, Table 8.2 [Learning] C, D). At the same time, three students stated they used GenAI tools by copy-pasting their outputs to prepare their assignments (Table 8.2 [Replacement] C, D). In that case, there is cognitive offloading as AI is replacing the work done previously by the students themselves.

Table 8.2. Selected excerpts of evidence related to participants' use of GenAI.

Category	Informants	Excerpt of Evidence
[Replacement]	Teachers	A. " I wanted to create 17 questions about the topic of 'friendship', and manually it is time-consuming" B. "I use GenAI a lot to create assessment rubrics"
	Students	C. "I add the requirements to Deep AI and it turns them into text. I simply copy-paste the result" D. "I use it to do my assignments. I know that this is not the proper way to proceed but I normally copy-paste the information that ChatGPT gives me"
[Complementarity]	Teachers	A. "In my Arts course we use many tools to do visual work. I showed Dall-E to my students and asked to make drawings with the help of AI" B. "It is easier to capture lesson the findings"
	Students	D. "It [ChatGPT] gives me ideas better than searching simply on the Internet. You can be for half a day without having any idea, while with ChatGPT you have ideas right away" E. "I had to do an assignment for school for geography. I didn't really know what to write about it and Copilot provided me with a lot of information"
[Learning]	Teachers	A. "I was looking how I could use active learning approaches and learn what they are and that's why I also used ChatGPT to help me" B. "It helps me to pick up new teaching methods"
	Students	C. " I use it to learn, because it can clarify quite a lot of things. I can prepare better for the test weeks" D. "I prompt information to Deep AI, and it writes it in a beautiful way that helps me to improve my skills"

Source: Author's own work.

When it comes to GenAI-associated fears (which can guide the design of the new system), most of the teachers' concerns (N=16) were related to *students' cognitive offloading* and the negative impact on learning (see Table 8.3, [Cognitive Offloading] A, B). Another GenAI disadvantage was related to an overreliance on the *AI-generated outcome.* Many teachers (N=13) were concerned about students becoming over-dependent on AI without questioning the results (see Table 8.3, [Overreliance] A, B) and on how they themselves might also rely too much on the GenAI output (N=7) (see Table 8.3, [Overreliance] C). Additionally, eight teachers connected such overreliance with the *lack of quality of AI answers* given the hallucinations that an AI model can produce (see Table 8.2, [Overreliance] B). Furthermore, many teachers (N=8) were concerned about *how GenAI impacts their own role and control in the teaching process.* Two teachers expressed their worry regarding the student-teacher relationship (see Table 8.3, [Teacher Replacement] A-C). Lastly, many teachers (N=7) were concerned about how to monitor students' use of GenAI (see Table 8.3, [Monitoring] A-C).

Reflecting the teachers' concerns, most of the students (N=7) indeed characterised GenAI as their personal "24-hour teacher" (Table 8.3, [Teacher Replacement] D, E) and five students expressed their unquestionable trust in its output (Table 8.3, [Overreliance] D). Also, a few students admitted *being sceptical about how GenAI use affects their cognitive development* and their work evaluation. Concretely, four students mentioned that the use of GenAI may hinder their learning growth because they tend to simply use the output with no further thought or work on their part (see Table 8.3, [Cognitive Offloading] C, D).

Table 8.3. Selected excerpts of evidence related to participants' ideas about GenAI pitfalls

Category	Informants	Excerpt of Evidence
Cognitive Offloading	Teachers	A. "I am afraid students make something, such as the assignments, out of nothing, because they need to think very creatively outside of what is given to them and they just stay on what ChatGPT is giving to do them" B. "I am afraid students have stopped thinking or brainstorming; this impact the development of critical thinking"
	Students	C. "The way it's used now, by me among others, affects our performance, because our skills are undeveloped. You don't build anything, someone else does all the work" D. "I think that in some cases you shouldn't use it. Sometimes we copy everything literally. Then I just think it's not smart, because we become lazier"
Overreliance	Teachers	A. "They [students] do not use other resources and uncritically they are copying everything" B. "ChatGPT is going to hallucinate, and you do not realise that, both us and especially the students" C. "I ask myself whether to trust the GenAI output or it is not trustable at all?"
	Students	D. "I like ChatGPT because it is something that you don't really have to check. I find it very reliable"
Teacher Replacement	Teachers	A. "Help! how can I guarantee my autonomy?" B. "I am afraid there will be collision between student and teacher, because the expert is considered to be ChatGPT" C. "I am not sure how to foster the student-teacher relationship within the AI era"
	Students	D. "The teachers are the ones normally providing information. But the teachers are not always available. So, it is easier and faster to use ChatGPT" E. "It is like a teacher for me. You can actually ask all the questions you don't understand, and it is always there for you. And for me, that's basically the same as a teacher"
GenAI Monitoring	Teachers	A. "There is the problem that you cannot see how much it has been created by AI and how much by the students" B. "I want to follow more students' progress using GenAI" C. "I do not know how to check how my students used GenAI"

Source: Author's own work.

Validation of requirements [phase 2]

While teachers and students had identified potential benefits of GenAI tools such as their usefulness for assignment support and efficiency, the interviews indicated that when GenAI is not pedagogically oriented, it can decrease teachers' awareness and capacity to exert control over the students' learning progress, i.e. it can decrease teacher autonomy during the instructional cycle. A loss of autonomy can potentially impact effective and contextualised teaching, especially important in primary and secondary education. Consequently, the results helped in the identification of design requirements for GenAI systems in education to foster teacher autonomy (see Figure 8.3).

Tune the GenAI output.

Teachers raised the need to control the GenAI output in order to enhance students' cognitive skills (e.g. "I am afraid students have stopped thinking or brainstorming, this impacts the development of critical thinking"). Currently, students (and teachers) use general-purpose GenAI models which are likely to not be contextualised. Thus, the envisioned tool should provide teachers the opportunity to train the models with their own documentation, and to place some "controls" on the responses these models provide to students. For instance, teachers might increase

the level of hallucinations to increase students' critical thinking when using these tools, or the depth and timing of the GenAI's output. Han et al. (2024[22]) also mentioned the need to create GenAI tools that will permit teachers to maintain their agency and control by finetuning options in the GenAI system (see also Cukurova, 2026[34]; Baker et al., 2026[35]).

Monitor GenAI use.

Tracking students' interaction with the GenAI tool was an important aspect raised by the teachers to assess whether these tools support the learning progress or merely replace the task completion (e.g. "I want to follow the students' progress more closely using GenAI"). Students also reported bad practices of copy- pasting the GenAI output directly (e.g. "I know that this is not the proper way to proceed but I normally copy-paste the information that ChatGPT gives me"). Monitoring the student interactions with GenAI tools permits the identification of potential knowledge gaps and the shaping of pedagogical informed interventions, readjusting the learning objectives, the lesson plan and the GenAI use (Hoogland et al., 2016[36]). Thus, the desired GenAI tool should permit teachers to follow student interactions with GenAI and provide pedagogical guidance accordingly.

Provide configurable options.

Teachers employ GenAI tools for replacement and complementarity purposes based on the nature of different educational tasks. Therefore, there was agreement that it is desirable to let teachers define the level of autonomy they desire with the tool, as proposed by Molenaar (2021[15]). For instance, some teachers might prefer to programme semi-automatic reactions when students overuse the GenAI system; others might prefer to get an alert, and others may not care about such issues.

Figure 8.3. Overview of the envisioned educational GenAI system

Note: Overview of the envisioned system according to the retrieved requirements: R1: Tuning the GenAI Output; R2: Monitoring GenAI use; R3: Providing configurable options. During the Co-Design Phase the teachers provide configurable options to semi-automate different educational tasks and train the GenAI model according to their course documents, curriculum etc. During the Co-Orchestration Phase, students interact with the different GenAI models (e.g. ChatGPT models, Co-pilot) according to the prior teacher configurations and teachers can monitor student interactions (i.e. GenAI Analytics) and get alerts of student progress.

Source: Authors' own work.

Low-fidelity prototyping [phase 2]

The design requirements derived from the interviews led to the development of a low-fidelity prototype under three different scenarios (see Table 8.4 and below). The GenAI could be finetuned by developers to better support classroom use while maintaining teacher autonomy. This prototype was then used in other co-design sessions with additional teachers to confirm the elicited requirements and to modify it according to their preferences and needs.

Prior to these sessions, the prototypes were fed with fictional data supporting the different design requirements described before (e.g. GenAI analytics about students who asked for more exercises, students who copy-pasted the given answers to course assignments). During the sessions, teachers interacted with the prototype simulating three scenarios at different teaching moments: pre-course design configuring and contextualising the GenAI chatbot; course monitoring and reflection on the GenAI analytics captured from the interactions of the students with the chatbot; and after-course reflection and assessment of students' submissions and of the employed teaching methods.

Afterwards, teachers were asked to complete several surveys with both close and open-answer questions to assess the integration of the prototype, its usefulness, its usability and potential adoption in their regular practice. Further information about the low-fidelity prototype and one of the co-design sessions can be found in Ortega-Arranz et al. (2025[37]).

Each scenario (see Table 8.4) aimed to address different teaching moments (i.e. course design, course enactment, after course-reflection and assessment), taking into account the complexity and nature of the teaching process. This was also about filling gaps: a lot of AI tools mainly focus on course enactment and offer limited educational tools supporting teachers in course design and after-course assessment or automatically generating learning tasks and feedback interventions that are context aware (Delgado Kloos et al., 2025[38]; Topali et al., 2025[39]).

Table 8.4. Prototype description under three use scenarios

Scenario	User	Phase	Description	Figure
1	Teachers	Course design	The prototype enables teachers to configure the behaviour of the system when the students interact with the chatbot. This includes contextualising the answers with documentation, and configuring the answers according to different teaching methodologies such as the desired feedback depth (e.g. providing hints versus direct answers).	Figure 8.4 left
	Students	Course enactment		Figure 8.4 right
2	Teachers	Course enactment and Post-course reflection	The prototype allows teachers to monitor all the interactions between the students and the chatbot. The prototype provides (and alerts) teachers with students' "GenAI Analytics". Such GenAI analytics include the topic and frequency of the students' prompts, and potential copy-paste behaviours.	Figure 8.5
3	Students	Course design	The prototype serves as a tool for learning and instructional preparation based on the course design, the activity characteristics and the learning objectives. It is a feedback partner for the teachers and provides feedback interventions for different circumstances, timing and foci, for example, in project-based assignments.	Figure 8.6

Source: Author's own work.

Figure 8.4. Screenshots of the low-fidelity prototype for Scenario 1
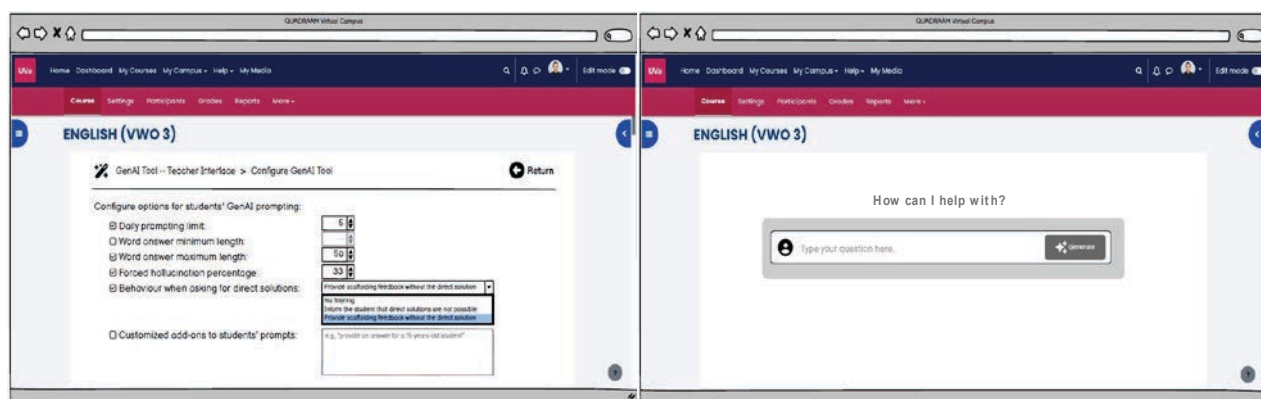
Source: Author's own work.

Figure 8.5. Screenshots of the low-fidelity prototype for Scenario 2
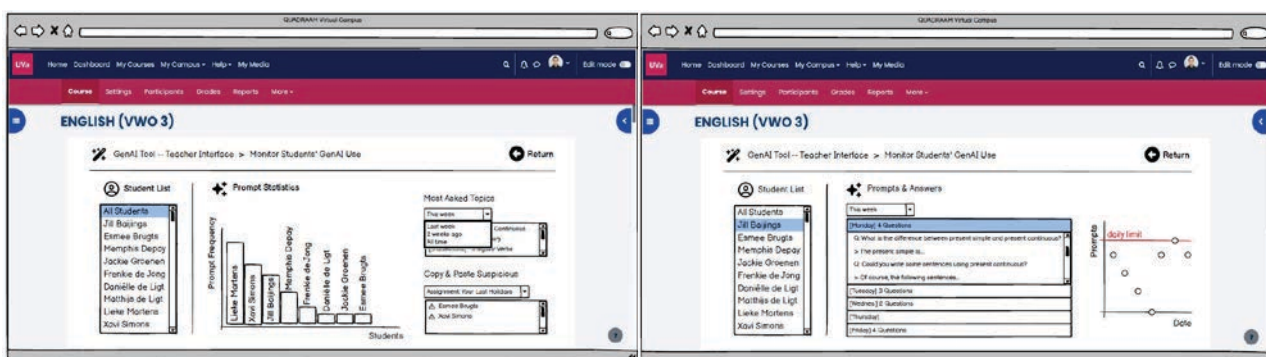
Source: Author's own work.

Figure 8.6. Screenshots of the low-fidelity prototype for Scenario 3
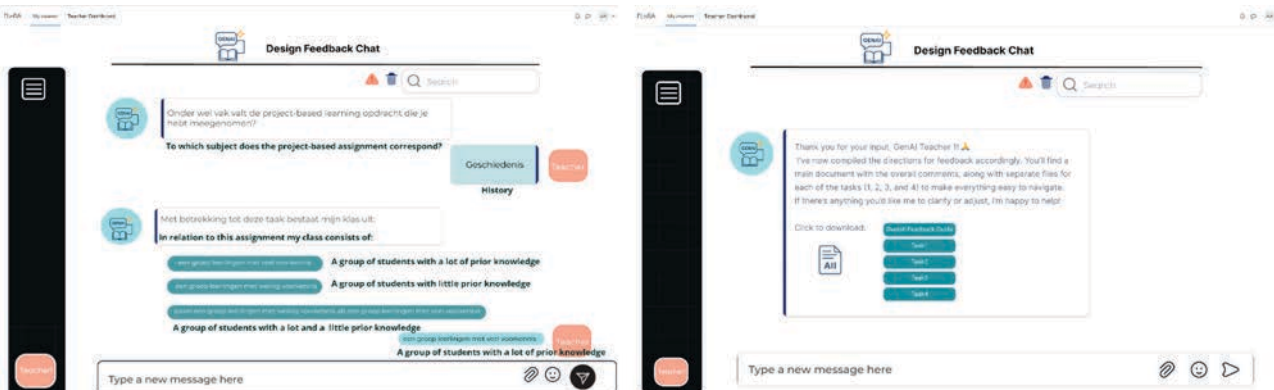
Source: Author's own work.

First insights on the low-fidelity prototype

The teachers actively contributed to shaping the prototype according to real-case scenarios and their own teaching contexts. They offered valuable feedback on what felt promising as well as areas that need further refinement. Overall, their impressions were positive, and the collaborative process highlighted both the opportunities and challenges in co-designing GenAI for educational contexts.

A key outcome of the GenAI co-design approach was *that teachers wanted to remain actively present as co-orchestration partners with the GenAI tool throughout the instruction process.* This co-orchestration unfolds through a deliberate division of labour over time. During the lesson planning and preparation phase, teachers set the pedagogical parameters within which the GenAI operates, defining learning objectives, instructional strategies and assessment criteria. This ensures that the system's functions remain grounded in the teacher's pedagogical judgment and contextual understanding of the learners. During the lesson enactment phase, the GenAI can execute predefined tasks, such as monitoring student progress or providing adaptive feedback, allowing teachers to redirect their attention toward more advanced pedagogical responsibilities, including facilitation of critical discussion, individualised mentoring, and emotional support.

Table 8.5. Selected excerpts of evidence related to participants' ideas about GenAI pitfalls.

Category	Informants	Excerpt of Evidence
Insights	Teachers	A. "It would help me to see what exactly the students do and where they might still need support B. "You could see if students can master the material with the right questions" C. ¨I think that students are more focused when working, then I can also give more targeted feedback¨
Personalisation	Teachers	A. "By getting insights about how the students use the tool, you can give better feedback and directions to the individual students." B. "It provides tools to better coach individual students, or to address possible misconceptions and ambiguities in the learning material in a more targeted way?"
Autonomy	Teachers	A. "It allows me to monitor many more individual students and therefore also guide them better. That tool makes me a more competent teacher" B. "I can control better a learning situation that includes GenAI" C. "Because it helps with awareness and goal-oriented use of GenAI, within a controlled environment. "
[Complexity]	Teachers	A. "Although it doesn't become easier due to the new options, it does become better" B. "It gives a clear and insightful overview of the prompts and how GenAI is used, but at the same time it results in increased workload" C. "It gives you a lot of insight, but at the same time it does not make your job faster"

Source: Author's own work.

Regarding concrete findings, while we explored the three scenarios, in this section we will only focus on Scenario 1 on helping teachers monitor the student-GenAI interactions. Here are the identified findings:

- **The prototype could help teachers to understand their students' moments of progress and adjust the feedback accordingly.**

  Seven teachers mentioned that such a tool would support their effectiveness by gaining insights on student progress, the methods they are applying and the moments they are struggling. It would help them know when and what type of support to provide (Table 8.5 [Insights]). Two teachers stressed the added value of personalising the support to individual student needs (Table 8.5 [Personalisation]).

- **The prototype could enhance teachers' feeling of autonomy and control over AI.**

  Five teachers noted that the envisaged tool could foster their autonomy and increase their level of control within learning situations involving GenAI tools (Table 8.5 [Autonomy]). During the co-design session they proposed further ideas for GenAI analytics to support their level of pedagogical oversight of their students, such as "an overview of cohorts of students that are and are not efficient in working with AI", "group report on students who work together using GenAI", "aspects that are changing notably about the student progress while they are using GenAI for a period. If you could get that in a file you could also share and re-use", "a report on GenAI prompt use or progress throughout the week".

- **Providing more insights to teachers could add more complexity.**

  Four teachers discussed the added value of the tool in terms of awareness of student progress compared to the extra complexity or workload that it adds (Table 8.5 [Complexity]). Several teachers stated that the tool might add cognitive load or time to their tasks, yet it would also offer insights to understand their students and help them stay in control over AI.

Discussion

During our Design-Based Research process, we examined how teachers and students actually use GenAI, what challenges they face and how they envision an ideal GenAI tool. Their main challenges concerned overreliance on AI outputs, limited critical evaluation, and uncertainty about appropriate use of GenAI. Teachers emphasised the importance of and need for context-sensitive, transparent, and customisable GenAI tools that complement rather than replace their professional judgment.

Concretely, we found that teachers use GenAI for three different purposes, but especially for replacement tasks such as lesson planning, content creation or teaching recommendations. Secondary school students use GenAI in assignments that require text-generation to get further explanations in STEM-related lessons, to get additional information on a given topic and to improve their writing skills. These results are in line with the ones obtained by Laak and Aru (2024[25]) regarding GenAI use cases for both teachers (e.g. recommendations on teaching methods) and for students (e.g. GenAI seemed more helpful in some courses rather than in others). Moreover, our study suggested that in practice teachers and students desire different AI use and automation levels based on the nature of the learning and the teaching tasks. Likewise, Brandão et al. (2024[40]) described that different activities can lead to different GenAI use; GenAI can serve for automating trivial tasks and it can be used as a critical partner for cognitively demanding activities. This aligns with Cukurova et al. (2019[41]) and Cukurova (2026[34]), who emphasise that AI in education should augment human capabilities rather than replace them, a finding echoed in our own results.

In our interviews, teachers' concerns on GenAI were focused on student cognitive offloading for core learning processes and on GenAI overreliance without developing students' critical thinking. Similarly, prior studies (Klopfer et al., 2024[42]; Zhai, Wibowo and Li, 2024[43]) highlighted that an overreliance on AI can undermine students' critical thinking. Buckingham-Shum (2024[44]) discussed the benefits of short-term productivity that GenAI offers in minor tasks and how its uncritical use within the learning practices may impact foundational learning skills, e.g. critical thinking. Further research is needed on how to potentially reshape the teaching and assessment processes questioning what new skills and processes are important to boost in the AI age (Rethink, 2024[45]; Miao and Shiohira, 2024[46]; Miao and Holmes, 2023[47]).

At the same time, many teachers expressed concerns about the impact of GenAI on their own role in the teaching process. Such concerns have been acknowledged as well by (Qin, Li and Yan, 2020[48]; Renz and Hilbig, 2020[49]) who detected several barriers to the adoption of AI, including the lack of teachers' skills and technological competences, the potential loss of teaching autonomy and teachers' agentic role within the learning process and data ownership. In line with these fears, in our study more than half of the students reported perceiving GenAI as a 'second teacher', always available to provide instant answers and feedback. Chan and Tsi (2024[50]) and Giannakos et al. (2024[51]) stressed the need to foster human expertise which encompasses students' emotional and contextual knowledge as well as pedagogical/didactical proficiency. This is essential to provide more holistic learning interventions. Framing the teachers' and students' GenAI use and concerns within the Human-AI automation model, we observed that teachers' current practices corresponded to high levels of automation in our sample, where GenAI often operates as an autonomous agent. In such cases, teachers have limited ability to control, adapt, or critically mediate the GenAI outputs to align with their pedagogical goals or specific classroom contexts. This dynamic positions teachers more as end-users of pre-configured GenAI tools than as decision-makers exercising their pedagogical considerations.

The evidence gathered from the Prototype in Scenario 2, which allowed teachers to configure how students interact with GenAI, helped to address several of the limitations of general-purpose GenAI tools discussed above, such as the lack of educational contextualisation, personalisation, pedagogical foundation and teacher autonomy. First, teachers reported that such a tool would allow them to monitor students' progress, identify learning difficulties, and adjust feedback accordingly, thus emphasising the value of personalisation to individual learners. Second, several teachers highlighted that the prototype enhanced their sense of autonomy and control, as it enabled them to interpret and act on data within their pedagogical judgment, reinforcing rather than diminishing their professional autonomy. In our next steps, we plan to engage in the co-design of the GenAI tool in connection with teachers' lesson plans and classroom activities, with the aim of contextualising the tool to specific curriculum and pedagogical needs.

Our study indicated that teacher autonomy is not a static condition but can be fostered via participation in both tool design and classroom co-orchestration with GenAI. We would like to acknowledge that an important dimension of teacher autonomy lies also in teachers' decision-making actions, e.g. to reject or override GenAI suggestions, such as discarding outputs that do not align with their pedagogical context, However, our analysis in this chapter centres on the structural forms of autonomy embedded in the design and orchestration of GenAI systems. When teachers are actively involved in the tool design phase, they help embed their pedagogical values, contextual knowledge, and ethical considerations into the system's parameters, ensuring that AI aligns with curricular intentions and classroom realities. When teachers are acting as co-orchestration partners during the instructional cycle, the foundation for a division of labour (with GenAI) over time is established: teachers determine the pedagogical tasks and when and how automation can happen.

Consequently, these two levels create a matrix (see Figure 8.8). A GenAI system may have *low AI automation in enactment* but still limit teacher autonomy if its design is predefined. Another GenAI system might support *high automation during enactment,* but still allow high teacher autonomy since its elements and aspects are co-designed. Within this dynamic, the role and autonomy of teachers can evolve in multiple directions. In some scenarios, GenAI systems risk replacing teachers by taking over core instructional functions such as content delivery, assessment, and feedback without permitting them to control or monitor the GenAI output, thereby marginalising human judgment and pedagogical expertise. In other scenarios, GenAI may complement teachers by handling some instructional tasks, enabling teachers to focus on higher-order teaching activities such as critical discussion and socio-emotional support. In some scenarios, GenAI could augment teachers, enhancing their capabilities and insights by performing tasks and generating learning insights that were previously difficult or impossible to achieve, such as real-time identification of the misconceptions of multiple individual students. This spectrum from replacement to augmentation highlights the ethical and pedagogical imperative to design and orchestrate GenAI systems that reinforce the central role of teachers in shaping meaningful learning experiences (Ley et al., 2025[52]; Molenaar, 2024[53]; Cukurova, 2026[34]) .

Figure 8.7. Matrix on teacher autonomy levels as co-participants during tool design and course enactment.
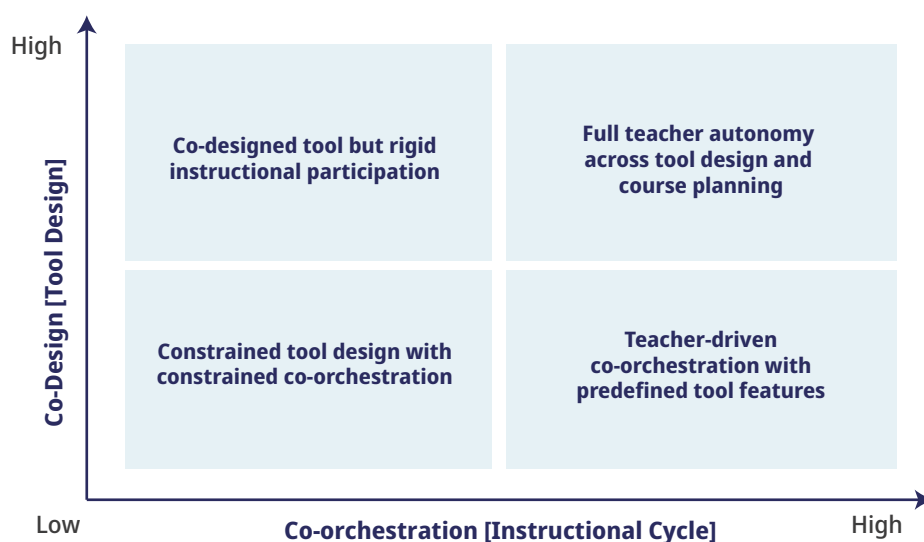
Source: Author's own work.

Conclusions

The acceptance and adoption of GenAI-powered systems in formal educational activities have to balance automation with teacher autonomy. Current GenAI tools rarely account for the autonomy teachers require in earlier stages of course design. Without meaningful involvement of teachers in the design of these systems, GenAI risks reinforcing predefined models of teaching and learning that may conflict with instructional and learning design principles.

This chapter discussed the importance of maintaining teacher autonomy during the design and use of GenAI in educational settings. The chapter showed how Human-Centred Design and Design-Based Research approaches can help to transition from general-purpose GenAI to educational GenAI, highlighting the need to involve stakeholders as co-design and co-orchestration partners when developing and using such systems. Human-Centred Design and Design-Based Research offer promising approaches to address this challenge. Human-Centred Design ensures that teachers and students are engaged as active participants in the design of educational technologies, while Design-Based Research involves iterative cycles of development and testing in authentic contexts.

We expanded the Human-AI automation model proposed by Molenaar (2022[30]) to frame teacher autonomy in designing and using GenAI systems in education. Accordingly, we distinguish between two key forms of teacher autonomy in the context of AI-supported education. The first involves teachers acting as *co-design partners* in the creation of the AI tools themselves. In this role, they contribute their insights and express their needs regarding elements such as the user interface, functional features, and underlying infrastructure, ensuring that the tools are purposefully aligned with educational goals. The second form positions teachers as co-orchestration partners throughout the instructional cycle. Here, teachers plan, enact, and reflect on each lesson while meaningfully integrating the AI tool into their teaching practice. This level of autonomy allows them to determine how instructional tasks are distributed between themselves and the AI, shaping how the technology supports learning in their own classrooms. Our study shows, using tangible evidence, how teachers and students can be involved in the design of educational GenAI systems and what kinds of systems meet their expectations. We hope it can inspire policy-makers and EdTech developers and companies.

Acknowledgements

Note

1.   https://www.synthesis.com/tutor, Last Access October 2025

References

Amiel, T. (2008), "Design-Based Research and Educational Technology: Rethinking Technology and the Research Agenda", *Educational Technology & Society*, Vol. 11(4), pp. 29-40, https://doi.org/10.1590/S0325-00752011000100012.   [31]

Baidoo-Anu, D. and L. Owusu Ansah (2023), "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning", *Journal of AI*, Vol. 7/1, pp. 52-62, https://doi.org/10.61969/jai.1337500.   [5]

Baker R.S., X. Liu, M. Shah, M. Pankiewicz, Y.J. Kim, Y. Lee and C. Porter (2026), "Generative AI as a teaching assistant", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [35]

Brandão, A., L. Pedro and N. Zagalo (2024), "Teacher professional development for a future with generative artificial intelligence – an integrative literature review", *Digital Education Review*, Vol. 45, pp. 151-157, https://doi.org/10.1344/der.2024.45.151-157.   [40]

Buckingham Shum, S., R. Ferguson and R. Martinez-Maldonado (2019), "Human-Centred Learning Analytics", *Journal of Learning Analytics*, Vol. 6/2, https://doi.org/10.18608/jla.2019.62.1.   [11]

Buckingham-Shum, S. (2024), *Generative AI for Critical Analysis: Practical Tools, Cognitive Offloading and Human Agency*.   [44]

Cai, N. (ed.) (2021), "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020", *Complexity*, Vol. 2021/1, https://doi.org/10.1155/2021/8812542.   [4]

Chan, C. and L. Tsi (2024), "Will generative AI replace teachers in higher education? A study of teacher and student perceptions", *Studies in Educational Evaluation*, Vol. 83, https://doi.org/10.1016/j.stueduc.2024.101395.   [50]

Chapwanya, O. (2025), "Exploring the teacher's role amid rising generative AI: An activity theory analysis in further education", *Studies in Technology Enhanced Learning*, Vol. 4/3, https://doi.org/10.21428/8c225f6e.4d227fbd.   [29]

Creswell, J. (2013), *Research Design: Qualitative, Quantitative and Mixed Approaches*, SAGE Publications Inc., https://doi.org/10.2307/1523157.   [32]

Cukurova, M. (2026), "A Conceptual Framework for Teacher-AI Teaming in Education: Harnessing GenAI to Enhance Teacher Agency", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [34]

Cukurova, M., C. Kent and R. Luckin (2019), "Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring", *British Journal of Educational Technology*, Vol. 50/6, pp. 3032-3046, https://doi.org/10.1111/bjet.12829.   [41]

Delgado Kloos, C., J. Asensio-Pérez, D. Hernández-Leo, P. Moreno-Marcos, M. Bote-Lorenzo, P. Santos,... B. Tabuenca (2025), "GENIE Learn: Human-Centered Generative AI-Enhanced Smart Learning Environments", *Proceedings of the 17th International Conference on Computer Supported Education*, pp. 15-26, https://doi.org/10.5220/0013076000003932.   [38]

Dimitriadis, Y., R. Martínez-Maldonado and K. Wiley (2020), "Human-Centered Design Principles for Actionable Learning Analytics", in *Research on E-Learning and ICT in Education*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-64363-8_15.   [12]

European Commission (2025), *Ethical guidelines for educators on using artificial intelligence*, https://education.ec.europa.eu/focus-topics/digital-education/action-plan/ethical-guidelines-for-educators-on-using-ai.   [7]

Friedman, B., P. Kahn, A. Borning and A. Huldtgren (2013), "Value Sensitive Design and Information Systems", in *Philosophy of Engineering and Technology, Early Engagement and New Technologies: Opening Up the Laboratory*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-007-7844-3_4.   [18]

Giacomin, J. (2014), "What Is Human Centred Design?", *The Design Journal*, Vol. 17/4, pp. 606-623, https://doi.org/10.2752/175630614x14056185480186.   [16]

Giannakos, M., R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo,... B. Rienties (2024), "The promise and challenges of generative AI in education", *Behaviour & Information Technology*, Vol. 44/11, pp. 2518-2544, https://doi.org/10.1080/0144929x.2024.2394886.   [51]

Han, A., X. Zhou, Z. Cai, S. Han, R. Ko, S. Corrigan and K. Peppler (2024), "Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1-17, https://doi.org/10.1145/3613904.3642438.   [22]

Hays, L., O. Jurkowski and S. Sims (2023), "ChatGPT in K-12 Education", *TechTrends*, Vol. 68/2, pp. 281-294, https://doi.org/10.1007/s11528-023-00924-z.   [23]

Holmes, W., J. Persson, I. Chounta, B. Wasson and V. Dimitrova (2022), *Artificial intelligence and education - A critical view through the lens of human rights, democracy and the rule of law*, Council of Europe.   [3]

Holmes, W. and I. Tuomi (2022), "State of the art and practice in AI in education", *European Journal of Education*, Vol. 57/4, pp. 542-570, https://doi.org/10.1111/ejed.12533.   [1]

Holstein, K., B. McLaren and V. Aleven (2019), "Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity", *Journal of Learning Analytics*, Vol. 6/2, https://doi.org/10.18608/jla.2019.62.3.   [19]

Holstein, K., B. McLaren and V. Aleven (2019), "Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-23204-7_14.   [20]

Hoogland, I., K. Schildkamp, F. van der Kleij, M. Heitink, W. Kippers, B. Veldkamp and A. Dijkstra (2016), "Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations", *Teaching and Teacher Education*, Vol. 60, pp. 377-386, https://doi.org/10.1016/j.tate.2016.07.012.   [36]

Kaplan-Rakowski, R., K. Grotewold, P. Hartwick and K. Papin (2023), "Generative AI and Teachers' Perspectives on Its Implementation in Education", *Journal of Interactive Learning Research*, Vol. 34/2, pp. 313-338, https://doi.org/10.70725/815246mfssgp.   [24]

Klopfer, E., J. Reich, H. Abelson and C. Breazeal (2024), "Generative AI and K-12 Education: An MIT Perspective", *An MIT Exploration of Generative AI*, https://doi.org/10.21428/e4baedd9.81164b06.   [42]

Laak, K. and J. Aru (2024), "Generative AI in K-12: Opportunities for Learning and Utility for Teachers", in *Communications in Computer and Information Science, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky,* Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-64315-6_49.   [25]

Laurillard, D. (2013), *Teaching as a Design Science*, Routledge, https://doi.org/10.4324/9780203125083.   [10]

Ley, T., M. Cukurova, J. Edwards, A.-C. Falhs, S. Järvelä, R. Kasepalu, R.,... Q. Zhou (2025), "Teaching with AI: The Role of Teachers in the Hybrid Intelligent System", in *Lecture Notes in Computer Science, Two Decades of TEL. From Lessons Learnt to Challenges Ahead*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-032-03873-9_3.   [52]

Lister, K., T. Coughlan, I. Kenny, R. Tudor, R and F. Iniesto (2021), "Taylor, the Disability Disclosure Virtual Assistant: A Case Study of Participatory Research with Disabled Students", *Education Sciences*, Vol. 11/10, p. 587, https://doi.org/10.3390/educsci11100587.   [21]

Mangaroska, K. and M. Giannakos (2019), "Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning", *IEEE Transactions on Learning Technologies*, Vol. 12/4, pp. 516-534, https://doi.org/10.1109/tlt.2018.2868673.   [33]

Miao, F. and K. Shiohira (2024), *AI competency framework for students*, UNESCO, https://doi.org/10.54675/JKJB9835.   [46]

Miao, F. and W. Holmes (2023), *Guidance for generative AI in education and research*, UNESCO, https://doi.org/10.54675/ewzm9535.   [47]

Molenaar, I. (2024), *Human-AI collaboration in education: The hybrid future,* Inaugural Speech RU, Radboud University, Nijmegen.   [53]

Molenaar, I. (2022), "Towards hybrid human-AI learning technologies", *European Journal of Education*, Vol. 57/4, pp. 632-645, https://doi.org/10.1111/ejed.12527.   [30]

Molenaar, I. (2021), *Personalisation of Learning: Towards Hybrid Human-AI Learning Technologies*, OECD Publishing, Paris, https://doi.org/10.1787/589b283f-en.   [15]

Molenaar, I., D. Baten, I. Bárd and M. Stevens (2025), "Artificial Intelligence and Education: Different Perceptions and Ethical Directions", in (Ed.), Smuha, N.A. , *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, Cambridge University Press, https://doi.org/10.1017/9781009367783.017.   [8]

Monteiro, F., P. Souza, M. da Silva, J. Maia, W. da Silva and D. Girardi (2024), "ChatGPT in Brazilian K-12 science education", *Frontiers in Education*, Vol. 9, https://doi.org/10.3389/feduc.2024.1321547.   [26]

Mørch, A. and R. Andersen (2023), *Human-Centred AI in Education in the Age of Generative AI Tools*, https://ceur-ws.org/Vol-3408/short-s2-08.pdf.   [6]

Ortega-Arranz, A., P. Topali and I. Molenaar (2025), "Configuring and Monitoring Students' Interactions with Generative AI Tools: Supporting Teacher Autonomy", *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 895-902, https://doi.org/10.1145/3706468.3706533.   [37]

Prestridge, S., K. Fry and E. Kim (2024), "Teachers' pedagogical beliefs for Gen AI use in secondary school", *Technology, Pedagogy and Education*, Vol. 34/2, pp. 183-199, https://doi.org/10.1080/1475939x.2024.2428606.   [27]

Qin, F., K. Li and J. Yan (2020), "Understanding user trust in artificial intelligence-based educational systems: Evidence from China", *British Journal of Educational Technology*, Vol. 51/5, pp. 1693-1710, https://doi.org/10.1111/bjet.12994.   [48]

Reiser, R. (2001), "A history of instructional design and technology: Part II: A history of instructional design", *Educational Technology Research and Development*, Vol. 49/2, pp. 57-67, https://doi.org/10.1007/bf02504928.   [9]

Renz, A. and R. Hilbig (2020), "Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies", *International Journal of Educational Technology in Higher Education*, Vol. 17/1, https://doi.org/10.1186/s41239-020-00193-3.   [49]

Rethink, A. (2024), *The Metacognition Revolution, AI is Playing a Central Role in Reshaping How We Learn,* https://www.theatlantic.com/sponsored/google/metacognition-revolution/3924/.   [45]

Shneiderman, B. (2020), "Human-Centered Artificial Intelligence: Three Fresh Ideas", *AIS Transactions on Human-Computer Interaction*, Vol. 12 (3), pp. 109-124, https://doi.org/10.17705/1thci.00131. [17]

Topali, P., C. Haelermans, I. Molenaar and E. Segers (2025), "Pedagogical considerations in the automation era: A systematic literature review of AIEd in K-12 authentic settings", *British Educational Research Journal*, https://doi.org/10.1002/berj.4200. [39]

Topali, P., A. Ortega-Arranz, M. Rodríguez-Triana, E. Er, M. Khalil, and G. Akçapınar (2025), "Designing human-centered learning analytics and artificial intelligence in education solutions: a systematic literature review", *Behaviour & Information Technology*, Vol. 44/5, pp. 1071-1098, https://doi.org/10.1080/0144929x.2024.2345295. [13]

Varghese, J. and J. Chapiro (2024), "ChatGPT: The transformative influence of generative AI on science and healthcare", *Journal of Hepatology*, Vol. 80/6, pp. 977-980, https://doi.org/10.1016/j.jhep.2023.07.028. [2]

Xiaoyu, W., Z. Zainuddin and C. Hai Leng (2025), "Generative artificial intelligence in pedagogical practices: a systematic review of empirical studies (2022–2024)", *Cogent Education*, Vol. 12/1, https://doi.org/10.1080/2331186x.2025.2485499. [28]

Zawacki-Richter, O., V. Marín, M. Bond and F. Gouverneur (2019), "Systematic review of research on artificial intelligence applications in higher education – where are the educators?", *International Journal of Educational Technology in Higher Education*, Vol. 16/1, https://doi.org/10.1186/s41239-019-0171-0. [14]

Zhai, C., S. Wibowo and L. Li (2024), "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review", Smart Learning Environments, Vol. 11/1, https://doi.org/10.1186/s40561-024-00316-7. [43]

Annex 8.A. Description of participant demographics in the study

Figure 8.8 depicts the demographics of the teachers who participated in the different sessions. Despite all of them previously using GenAI tools for their teaching practice, 63.2% of the teachers reported limited AI competence and only 36.8% feel confident enough with GenAI. Students (N=6 male, N=6 female), also coming from the Netherlands, were between 12-14 (N=7) and 15-17 (N=5) years old.

Figure 8.8. Teachers' gender, years of experience and courses expertise as collected during phase 1 and phase 2
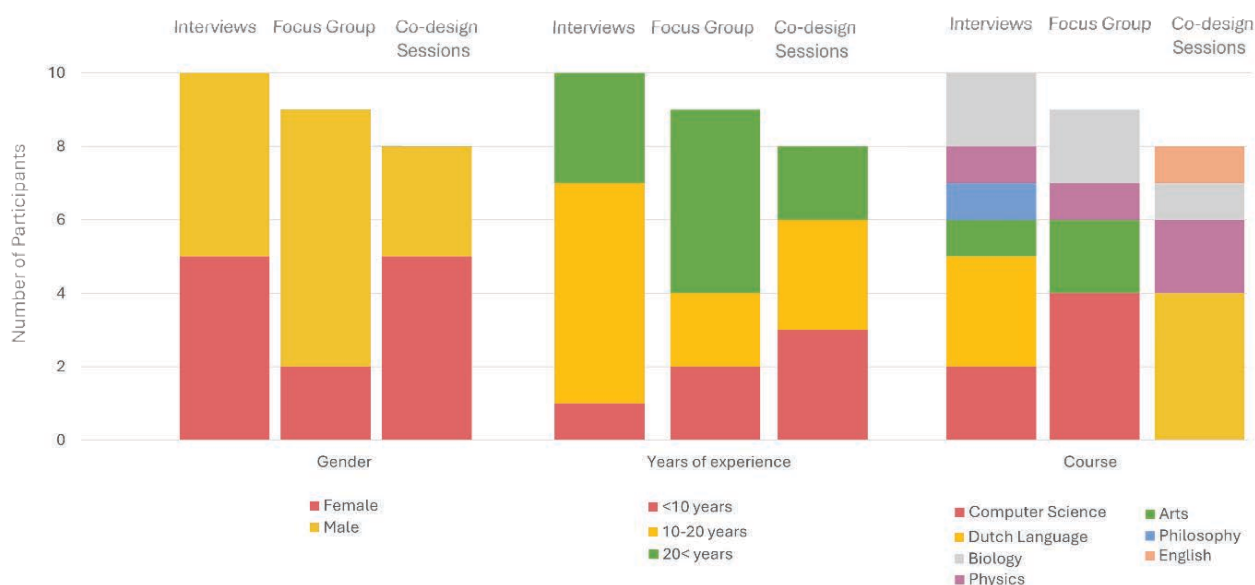
Source:Authors' own work.

9 Generative AI as a teaching assistant

Ryan S. Baker, Xiner Liu, Mamta Shah, Maciej Pankiewicz, Yoon Jeon Kim, Yunseo Lee, Chelsea Porter
University of Pennsylvania, United States

This chapter examines how AI Teaching Assistants driven by generative AI can expand and support the traditional roles played by human Teaching Assistants (TAs) and instructors in higher education. AI TAs are designed to automate routine administrative tasks, deliver real-time student support, and provide timely formative feedback at scale in ways that complement and extend human capabilities. Rather than replacements for human TAs, we argue for a model of augmentation AI TAs provide support and feedback that is otherwise infeasible for human instructors and TAs alone: support and feedback that is immediate, personalised, detailed, and available around the clock. The presentation and discussion of JeepyTA, a generative AI system, helps ground these ideas. We also discuss key design and implementation considerations.

Introduction

Much of the recent focus on generative artificial intelligence (GenAI) in education has considered it as a tool used individually either by a student or an instructor, through a commercial off-the-shelf chatbot designed to be an assistant. This role has considerable value (see Gašević and Yan, 2026[1]; Li and Hu, 2026[2]), but GenAI can be used in several other fashions, as shown in this report. This chapter considers the role of GenAI in supporting instructors and Teaching Assistants (TAs). TAs serve as a foundational support structure within colleges and universities, bridging the gap between students and academic staff and helping to sustain the quality of instruction in a wide range of disciplines. By leading discussion sections, grading assignments, answering questions, and offering individualised guidance, TAs play a vital role in shaping how students experience their courses. This dual role as both an intermediary and a mentor underscores the significance of the work that TAs do, not just in managing course logistics but in advancing student learning, engagement, and success. As higher education evolves – facing increasing enrolment pressures, budget constraints, and shifts toward online or hybrid delivery – the role of TAs is likely to expand. At the same time, TAs are themselves students, balancing these teaching responsibilities with their own scholarly activities, which can lead to time conflicts, uneven support for learners, and burnout. These challenges raise important questions about how universities can sustainably leverage the benefits TAs provide while addressing the real human limitations of time, expertise, and scalability.

Recent advances in GenAI offer a promising avenue to complement TAs' efforts, while retaining the human element that underpins great teaching. GenAI-driven "teaching assistants," powered by large language models (LLMs) and other advanced technologies, have the potential to handle repetitive administrative tasks, deliver targeted learning support, and provide immediate feedback to students in a way that human TAs alone simply cannot. By examining the evolving role of TAs and exploring how AI can enrich and extend their capabilities, this chapter seeks to highlight both current practices and new horizons for more equitable, accessible, and impactful teaching support. We conclude with a discussion of implications, including for policy.

Teaching assistants

The teaching assistant role

A Teaching Assistant (TA) in higher education (in countries where this role exists) is typically a graduate (Master's and PhD) or advanced undergraduate (Bachelor's) student who supports the primary instructor in delivering course content and assisting students. A TA's responsibilities may vary by institution and discipline, but they generally include facilitating small-group discussions, answering student questions, and supplementing the main instructor's efforts to create an effective learning environment (Park, 2004[3]).In recent years, as more courses have tended to partly or fully move online, TAs also play a large role in managing discussion forums, answering student questions and supporting discussions there (Wadams and Schick-Makaroff, 2022[4]). By handling a portion of the teaching and administrative workload, TAs play a critical role in making large or complex courses more manageable, thus enhancing the overall educational experience for both academic staff and students.

The origins of Teaching Assistants can be traced back to the late nineteenth century, when growing student enrolments and expanding research expectations prompted universities to look for ways to extend instructional capacity (Chism, 1998[5]). In these early stages, TAs often served as informal helpers to more senior academic staff, assisting with tasks like grading or lab supervision as part of their own apprenticeship in academia. Over time, and particularly after World War II when higher education systems expanded rapidly, the role of TAs became increasingly formalised. Universities began creating structured programs that provided clearer job responsibilities, training, and professional development opportunities, reflecting the recognition that TAs could significantly enhance both teaching and learning. This evolution laid the foundation for the modern TA role, in which graduate and advanced undergraduate students are systematically integrated into the educational process (Park, 2004[3]).

Teaching Assistants carry out a range of tasks that collectively support both the instructor and students in higher education settings. Their responsibilities can be broadly categorised into instructional support and student engagement, though these two areas naturally overlap (Park, 2004[3]). By taking on these roles, TAs help to foster an environment that promotes understanding, participation, and continuous feedback - key components of effective learning (Hattie and Timperley, 2007[6]; Chi and Wylie, 2014[7]).

In terms of instructional support, TAs frequently lead discussion sections, tutorials, or laboratory sessions, serving as facilitators who bridge theory and practice (Park, 2004[3]). In these smaller and often more interactive settings, TAs clarify course material, demonstrate practical techniques, and encourage student participation. By adapting teaching methods to the needs of specific groups of students, TAs help maintain a dynamic and inclusive classroom atmosphere. Another critical element of a TA's role involves assessing student work. TAs often grade assignments, quizzes, and exams under the supervision of the lead instructor (Marshman et al., 2018[8]). This process typically includes reviewing submissions, providing constructive feedback, and highlighting areas for improvement - ideally, guiding students to developing and demonstrating deeper understanding (Marshman et al., 2018[8]).Grading student work not only eases faculty workload but can also offer TAs valuable, instructor-scaffolded experience in evaluating academic performance, helping TAs to develop deeper understanding of student thinking (Karim, Maries and Singh, 2018[9]).

As for student engagement in their learning process, in many modern courses – particularly those with hybrid or fully online components – TAs serve as key points of contact on discussion forums (Wadams and Schick-Makaroff, 2022[4]). By responding to questions, facilitating conversation, and sharing clarifications from the instructor, they help maintain an active and supportive online learning community. This work often extends to moderating peer-to-peer exchanges, ensuring that discussions remain on topic and respectful. In addition, TAs often hold regular office hours and meet with students to allow students to seek in-depth explanations, review feedback, or discuss academic challenges (Park, 2004[3]). These mechanisms often provide learning support beyond what instructors can offer just-in-time and on-demand, particularly for research-active senior academic staff or large courses.

Teaching Assistants provide a *range of benefits* to higher education institutions, more senior academic staff, and students. Universities often find that using TAs is a cost-effective means of managing large course enrolments while still providing individualised support to students, a topic of constant interest to administrators when university budgets are under pressure. For more senior academic staff, TAs offer substantial advantages by relieving some of the workload associated with teaching, grading, and administrative duties. By delegating tasks such as discussion facilitation, assignment feedback, and routine course management, faculty members can devote more time to developing innovative curricula, advancing their research agendas, and mentoring students (including the TAs) at higher levels. In addition, TAs often introduce diverse perspectives or novel approaches to instruction, encouraging

a collaborative environment in which both senior academic staff and TAs refine teaching strategies (Begley et al., 2019[10]). Finally, students also benefit significantly from the involvement of TAs. In many cases, TAs are more available to answer questions outside of regular class times, and their support on online discussion forums can be accessed asynchronously, providing a flexible option for students who need extra help. TAs' relative proximity to the student experience - whether by age, academic journey, or shared disciplines - can also result in a peer-mentorship-like atmosphere during office hours and informal interactions. As such, TAs' greater relatability and students' perception that they are more understanding (Kendall and Schussler, 2012[11]) can ease anxieties and foster a sense of community, ultimately enhancing the overall learning experience.

However, several challenges have been noted for current practices involving Teaching Assistants. Regarding instructional support, many TAs may for example lack pedagogical training or skill (Nicklow, Marikunte and Chevalier, 2007[12]). This lack of formal preparation can undermine the quality of their instruction, as they may be uncertain about how to present information clearly. Moreover, TAs sometimes adopt a surface-level approach to grading, focusing on relatively simplistic aspects of correctness rather than attempting to provide feedback that guides students toward deeper conceptual understanding (Marshman et al., 2018[8]). Compounding these issues is the fact that TAs typically possess less subject-specific expertise than full academic staff (for example, see Karim et al., 2018[9]), which can limit their ability to answer complex questions or provide advanced guidance.

Furthermore, TAs often face challenges around workload and time constraints. Many TAs must balance teaching responsibilities with personal academic obligations, such as coursework, research projects, and preparation for required examinations (Park, 2004[3]). Some TAs may find it difficult to invest the necessary time in class preparation, grading, or providing substantive feedback to students. This overload can also lead to high stress and exhaustion, reducing their effectiveness as a TA while also impairing their other work and personal success (Berta and Pembridge, 2019[13]).This is exacerbated by the uneven training and faculty support provided to TAs. Some instructors involve TAs extensively in designing lesson plans, assessment rubrics, or instructional materials, while others may provide only minimal training and mentorship (Begley et al., 2019[10]). Furthermore, many TAs do not have access to teaching mentors other than the instructor. This lack of support can leave TAs uncertain about expectations or best practices, making it harder for them to support students without spending large amounts of their time.

Hence, while TAs fulfil crucial roles in supporting learning and engagement, and help bridge the gap between senior academic staff and students, there are several aspects in which current practices are not optimal for either TAs or students.

Review of AI teaching assistants (AI TAs) and their use

Could generative AI (GenAI) technologies improve things for both TAs and students? The emergence of GenAI models has created a great deal of enthusiasm for the potential of a wide range of educational benefits. One area of rising interest has been the creation of AI Teaching Assistants–tools that extend the capabilities of human TAs. AI Teaching Assistants (AI TAs) use computational methods – in many recent cases GenAI (Sajja et al., 2024[14]; Yetişensoy and Karaduman, 2024[15]; Liu et al., forthcoming[16]) but, before that, machine learning and previous-generation natural language processing (NLP) (Khosravi et al., 2021[17]; Shermis and Burstein, 2003[18]) – to perform tasks that were typically in the purview of human TAs, although in many cases beyond what was feasible for human TAs. Their scope has included streamlining routine administrative tasks, providing targeted learning support or rapid formative assessment, and empowering human TAs and instructors with information and insights about their students, enhancing the overall effectiveness of instructional delivery. While they sometimes take over what used to be human tasks, these systems are not intended to replace human educators, but instead to provide support 24/7 and free up valuable time for TAs and instructors to focus on more complex, critical, and high-impact aspects of teaching.

Over the past few decades, the use of AI in educational technology has evolved from simple automation tools – like basic quiz generators and grading scripts (see discussion in Brooks, 2023[19]) – to advanced AI systems that can process language and consider context, allowing them to respond to students' questions and needs in real time. The previous generation of intelligent tutoring systems and question answering systems could offer sophisticated support (Nye, Graesser and Hu, 2014[20]; Goel and Polepeddi, 2018[21]), but was highly expensive to author (Aleven et al., 2006[22]), often necessitating intense focus on only a single aspect of adaptivity (Baker, 2016[23]). The contemporary use of GenAI, sometimes combined with previous-generation machine learning, creates the potential for a qualitative leap forward in functionality and sophistication, at much lower development cost.

These technologies, when used to complement instructors, can take on repetitive tasks – such as answering common questions and some parts of the assessment of student work – thereby providing immediate, round-the-clock support to learners and eliminating the bottleneck that often occurs when TAs or instructors

are not available – for instance, for an online learner working from a different time zone than the instructors. At the same time, human TAs and instructors can allocate their expertise to higher-order pedagogical activities, such as facilitating in-depth discussions, offering mentorship, and providing customised feedback for unusual cases and learning challenges. This synergy ultimately helps institutions maintain quality education at scale, addresses the labour-intensive aspects of teaching, and ultimately supports instructors in finding time for high-value personal interaction with learners. In the following section, we will discuss some of the ways that AI TAs can support learners, human TAs, and instructors.

Key functions and use cases of AI-based teaching assistants (AI TAs)

Administrative and logistical support is an area of responsibility for current human TAs that is easy and fairly non-controversial to replace. By automating more administrative processes such as course enrolment or the monitoring of completion of assignments, instructors can devote more time to pedagogical planning and personalised student engagement (Ahmad et al., 2022[24]). AI TAs can also manage course communications by sending out timely reminders for assignments, examinations, and events, ensuring students remain informed and minimising the risk of missed deadlines (Pardo et al., 2018[25]). Streamlining these tasks can allow human instructors and TAs to focus on higher-level teaching responsibilities, such as curriculum development and individualised feedback. Beyond routine administrative tasks, AI systems can support course logistics by monitoring student participation and promptly alerting instructors to potential engagement and performance issues (Wise and Jung, 2019[26]; Khosravi et al., 2021[17]), as well as distil insights from discussions on course forums for instructors.

Another potential area of application for AI TAs is *instructional support*. AI Teaching Assistants can provide supplementary explanations or resources tailored to students' individual needs. For example, if a student expresses confusion about a specific topic, these systems can supply targeted materials, such as a textual explanation, brief video tutorials, interactive modules, or suggested readings (Sajja et al., 2024[14]; Essel et al., 2022[27]; Yetişensoy and Karaduman, 2024[15]). Such a system can provide more attention to customising learning for a given student than would be feasible for even the most dedicated human Teaching Assistant. By engaging in follow-up discussion, an AI TA based on a chatbot can support a student in ways that would be infeasible with a static resource. In addition, just as current GenAI can recommend resources to a learner, it can also assist educators with content curation and lesson planning, suggesting how to communicate topics more effectively (Karpouzis et al., 2024[28]).

Furthermore, AI TAs can support instructors in *evaluating* students' progress, particularly when it comes to formative assessment. Summative assessment through AI has been used in some applications but still needs to meet a higher bar for reliability and fairness (Bulut and Beiting-Parrish, 2024[29]). Formative assessments to inform instructors or support learners can be used safely due to the lower stakes. There is a long history of using automated assessments; decades of work assessed learners with automated quizzes and multiple-choice items (Brooks, 2023[19]) and a previous generation of NLP afforded short-answer grading (Gao et al., 2024[30]) and automated essay scoring (Shermis and Burstein, 2003[18]). Previous work also enabled the generation of new items, through procedural templates for instance (Kurdi et al., 2019[31]). However, the advent of generative AI has made it possible both to generate new items in sophisticated, tailored ways (Yadav and Tseng, 2023[32]; IIITD et al., 2024[33]; Sajja et al., 2024[14]) and to offer much more detailed, rich feedback on complex artifacts created by students (Chiang et al., 2024[34]; IIITD et al., 2024[33]; Stahl, Biermann and Wachsmuth, 2024[35]; Teng et al., 2024[36]). Automated announcement tools can then deliver personalised updates to students, supporting both performance and self-regulation (Lim et al., 2021[37]). By making it feasible to offer detailed feedback in a much timelier fashion, students are likely to revise their understanding and adapt in ways that align with the course goals (Graham, 2023[38]).

AI Teaching Assistants embedded in course platforms can effectively serve as a first point of contact for students, promptly addressing frequently asked questions about the syllabus, assignment deadlines, and other logistical concerns (Liu et al., forthcoming[16]). By referencing a structured knowledge base, these systems can also respond to content-related inquiries, offering supplementary explanations or clarifications (Liu and M'Hiri, 2024[39]; Maiti and Goel, 2024[40]; Miroyan et al., 2025[41]; Teng et al., 2024[36]; Liu et al., forthcoming[16]). Questions can be asked in an external platform (Alsafari et al., 2024[42]; Teng et al., 2024[36]), within the course discussion forum (Liu et al., forthcoming[16]; Miroyan et al., 2025[41]; Haaben, 2024 (Unpublished Manuscript)[43]), a learning management system (LMS) (Maiti and Goel, 2024[40]), or in the context of a learning activity itself (Lee et al., 2023[44]). These tools can save considerable time for instructors and human TAs (Sinha, 2024[45]; Miroyan et al., 2025[41]).

More importantly, the immediate, round-the-clock availability of these tools supports learners who may need help outside of conventional office hours; for example, one study of an AI TA embedded into a discussion forum found that students received responses significantly more often on weekends with the AI TA than during the previous (only human TA support) semester (Liu et al., forthcoming[16]). While such a system cannot respond to all student

queries, it can provide support in many cases. In other cases where a query requires more nuanced interpretation or context–such as complex conceptual misunderstandings or unique personal circumstances–AI TAs can escalate the matter to human TAs, thus ensuring students receive appropriate and thorough support. This triaging function can help manage the flow of incoming questions, reducing the volume of simpler queries that human TAs and instructors must handle. As a result, educators are freed to spend more time providing personalised feedback, guiding higher-level discussions, and engaging students in meaningful academic interactions. While this type of question answering functionality was available even before the advent of LLMs, it required considerable engineering (e.g. Goel and Polepeddi, 2018[21]) compared to the relative ease of deployment now possible.

Key considerations

One of the key steps to moving these types of advancements from one-off research projects to scalable solutions benefitting a large number of learners is ensuring they integrate seamlessly with existing educational infrastructures. Many AI-based tools of this type so far require learners and instructors to use separate platforms rather than being integrated directly into their primary learning management systems (LMS) or discussion forums. This lack of integration or interoperability can create a fragmented user experience, requiring additional sign-ins, duplicating data entry, and making it harder to track student progress across multiple systems. In contrast, compatibility with widely used LMSs (e.g. Canvas, Moodle, Blackboard) and discussion forum platforms (Piazza, Discourse, phpBB, vBulletin, Flarum) would allow AI TAs to seamlessly access course materials, participation records, and student performance data. Such interoperability not only streamlines the user experience but also supports richer analytics and more effective, personalised interventions, ultimately strengthening the teaching and learning process.

Another key step for making these systems usable at scale will be efforts to engineer the human-computer interactions of these systems to facilitate their use by busy human TAs and instructors. Currently, the process of integrating course resources varies in complexity between tools, and the degree of uptake can vary considerably between instructors (Maiti and Goel, 2024[40]). There are several ways to accomplish this, including shared folders, access to learning management systems as a simulated student, or tools for uploading resources - but whichever approach is chosen, it must be low-effort for human beings. It should also be easy to continually update these resources, as changes to course materials and syllabi will often occur within a semester and across semesters for courses that are offered on a regular basis.

In addition, onboarding and even training is needed for the human TAs and instructors who will collaborate with an AI TA. They will need to understand enough about how the system works, what it can do, and what its limitations are, to ensure that they implement it effectively in their courses. By clearly communicating which queries or tasks the AI TA should handle versus those that call for human expertise, institutions can maintain quality control while maximising efficiency. Over time, incorporating continuous improvement and feedback loops can further refine these boundaries. For instance, platforms should support human TAs and instructors in regularly reviewing the AI TA's responses to student inquiries. Student and instructor feedback collected through short surveys or mining forum discussions can also highlight areas where the AI TA might be underperforming or producing confusing or inaccurate information. Supporting instructors in checking and refining the system will help to ensure that content remains accurate, relevant, and aligned with educational objectives, reduce instructor frustration, and increase the likelihood of long-term sustained use.

Case study analysis: The JeepyTA platform in universities

JeepyTA is an example of a course-specific, AI-driven Teaching Assistant designed to integrate with existing classroom and online practices. Developed by the Penn Center for Learning Analytics at the University of Pennsylvania (UPenn) and launched in Fall 2023 (Liu et al., forthcoming[16]), JeepyTA utilises a multi-turn conversational architecture of large language models (LLMs) and is not bound to a specific model – it can be configured to run on many LLMs (e.g. GPT, Llama, or DeepSeek). In courses where JeepyTA has been used, a recent OpenAI GPT model has been used (starting with GPT-3.5 Turbo, moving to GPT-4, GPT-4-Turbo and GPT-4o).

JeepyTA has been used in various scenarios: to deliver responses to logistics questions, contextually respond in discussions based on provided course materials, provide targeted feedback to written assignments and coding problems, and to serve as a brainstorming partner. As of Spring 2025, JeepyTA has been deployed across 16 sessions of 14 courses at three higher education institutions in the USA (with deployments also running later in 2025 in Singapore and Colombia). This widespread adoption reflects the growing interest in AI systems that can free human instructors and TAs from repetitive logistical duties, while still delivering responsive, round-the-clock support for learners.

To ensure course alignment, JeepyTA is primed with instructors' chosen reference materials, including syllabi, textbooks, readings, and past instructor feedback. These resources are embedded in the system's knowledge base through a retrieval-augmented generation (RAG) workflow: newly uploaded documents are converted into embedding vectors, enabling JeepyTA to retrieve contextually relevant information via semantic search. In doing so, JeepyTA can address administrative queries – such as answering date-related questions from the syllabus – while simultaneously leveraging contextual readings to stimulate in-depth discussions on course-specific topics. Additionally, through collaboration with instructors, in several cases, JeepyTA's prompts have been iteratively refined to better address specific learning objectives. Separate models have been employed to automate decisions on whether responses appear immediately or await instructor approval, providing finer control over JeepyTA's engagement in forum discussions. Finally, JeepyTA's behavior can be customised by category of tasks, giving instructors the flexibility to choose which topics or discussion forum categories it responds to and with what level of human supervision (human-in-the-loop).

Across courses, JeepyTA has been integrated into the open-source Flarum platform, appearing as a forum user distinctly marked as an AI Teaching Assistant. Through a Progressive Web App, the forum is accessible on mobile devices, allowing students and instructors to stay engaged on-the-go. In addition to traditional email notifications, users of a mobile app can receive push alerts – such as when JeepyTA responds or specifically mentions them – ensuring timely updates and facilitating faster interaction within the discussion forum.

Answer logistics questions

One of JeepyTA's primary functions is answering logistics questions about the course. At the beginning of the semester, JeepyTA can handle enrollment-related inquiries, including prerequisites, add/drop deadlines, and options for changing course registration. When students need accommodations, JeepyTA directs them to official university guidelines and relevant support services. It also provides information on class schedules, classroom locations, and changes due to holidays or special events. When a course has multiple sections, JeepyTA helps students confirm where they need to attend.

To support coursework, JeepyTA clarifies submission guidelines for assignments, specifying required file formats, submission portals, and deadlines. JeepyTA also assists with technical aspects of online learning platforms when required by the course. It helps students log into external platforms used by instructors (for instance, for video discussions) and provides information like login codes, platform access links, and usage instructions. If students experience submission errors or other technical problems, JeepyTA offers guidance in many cases without needing to involve the instructor.

JeepyTA helps students understand grading policies by explaining how grades are calculated based on rubrics, weighted components, and participation requirements. It also assists in interpreting feedback from instructors and TAs and guiding students on resubmissions, appeals, or grade disputes (see Figure 9.1). When students need access to course materials, JeepyTA provides links to lecture slides, reading repositories, and virtual meeting links, ensuring they have the necessary resources.

With recent updates, JeepyTA can remember instructor responses and announcements on recurring topics. If students ask about schedule changes, assignment deadlines, or policy updates, JeepyTA provides the latest information. This reduces confusion and keeps students informed without requiring instructors to repeat themselves.

It is worth noting that JeepyTA's ability to answer logistics questions depends on the information instructors choose to provide. It does not generate responses based on general knowledge but instead pulls from course-specific details that instructors input. If a detail was not provided, JeepyTA directs students to the human TA (if available for the course) and instructor or official course documents rather than guessing or giving incomplete information.

JeepyTA's performance in answering logistics questions is not always perfect, as some student inquiries may go beyond what is covered in the course materials. Therefore, instructors can choose to edit JeepyTA's responses at any time instead of only choosing between fully accepting or discarding them. This option is especially useful when combined with the feature that allows instructors to review JeepyTA's response before it becomes visible to other students (as explained below). This flexibility allows instructors to keep useful parts, make quick edits, and provide students with accurate information while reducing effort.

Figure 9.1. JeepyTA guiding students in making up for part of the assignment
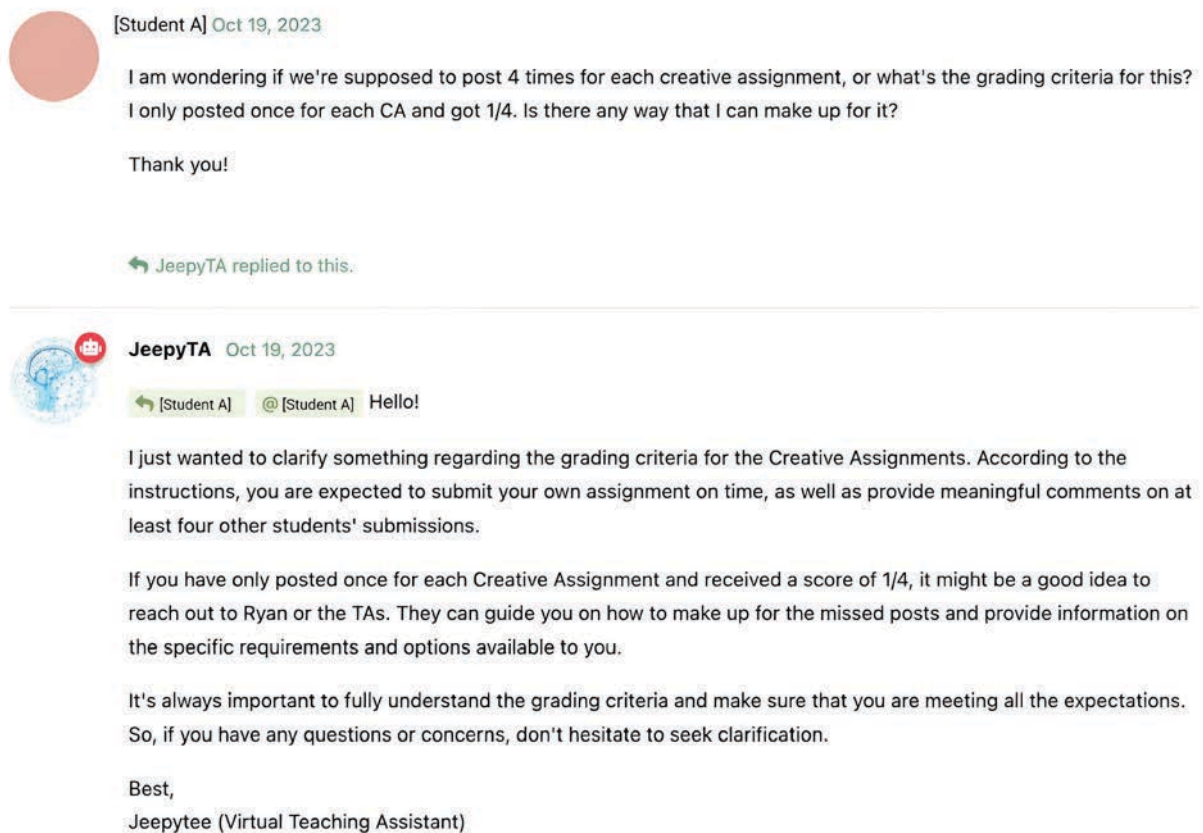
Source:Authors' own work.

Provide feedback to student essays

In several courses, JeepyTA provides feedback on student essays based on the grading rubric specified for the assignment. This consists of both higher-level conceptual elements and aspects of writing. In terms of higher-level conceptual elements, JeepyTA evaluates essays on the goals of the assignment, such as whether (for example) the student has appropriately discussed the needs of stakeholder groups, whether the student has made arguments in terms of theories discussed in class, or whether the limitations of a proposed solution have been concretely detailed (see Figure 9.2). When students submit drafts, JeepyTA acknowledges what they do well and highlights their strengths.

JeepyTA also offers feedback on more mechanical aspects of writing such as argument clarity, evidence use, structure, and writing quality (see Figure 9.2). In addition, JeepyTA comments on lower-level details such as language use, unclear phrasing, grammar mistakes, and wordiness. In these cases, it suggests revisions that can preserve the student's original intent.

GPT models are trained to provide generic responses that apply across many scenarios, which can make their default feedback vague or overly general. To prevent this, JeepyTA was instructed to "provide actionable insights rather than shallow suggestions". This small detail within prompt engineering makes a difference in helping students receive concrete guidance that improves their revisions.

If students need clarification, they can ask follow-up questions, and JeepyTA refines its guidance based on those questions. Instructors can also adjust JeepyTA's feedback settings to focus on specific aspects of writing or emphasise areas where students generally struggle the most.

Before asking JeepyTA to provide feedback on essays, its responses are first tested on a set of sample essays and the output is reviewed with instructors. This step helps confirm that the feedback aligns with the pedagogical goals of the course. When necessary, the prompts are refined based on the instructor's suggestions in the review process. This process helps JeepyTA provide comments that are clear, relevant to the assignment, and focused on the aspects instructors consider most important. It also creates an opportunity to catch cases where the LLM's default knowledge base produces inaccurate info, such as in cases where much of the content on the web reflects an incorrect understanding of a specific technical point.

In some cases, the prompt specifies a particular tone to shape the feedback style. For example, JeepyTA can be instructed to provide concise and direct feedback or take a more encouraging and supportive tone. This allows the feedback to align with the way instructors and TAs typically communicate with students about their writing. Additionally, past feedback from previous course offerings, along with de-identified student essays from matching pairs, are included in some cases as a reference for JeepyTA. JeepyTA does not use the content of past essays as a source for feedback but instead looks at these examples to follow the structure, level of detail, and key focus areas that instructors and TAs have emphasised. This helps make the feedback more useful to students by reflecting the expectations and priorities set in previous iterations of the course.

Figure 9.2. JeepyTA providing feedback on the first step in a student essay assignment – the essay prospectus.
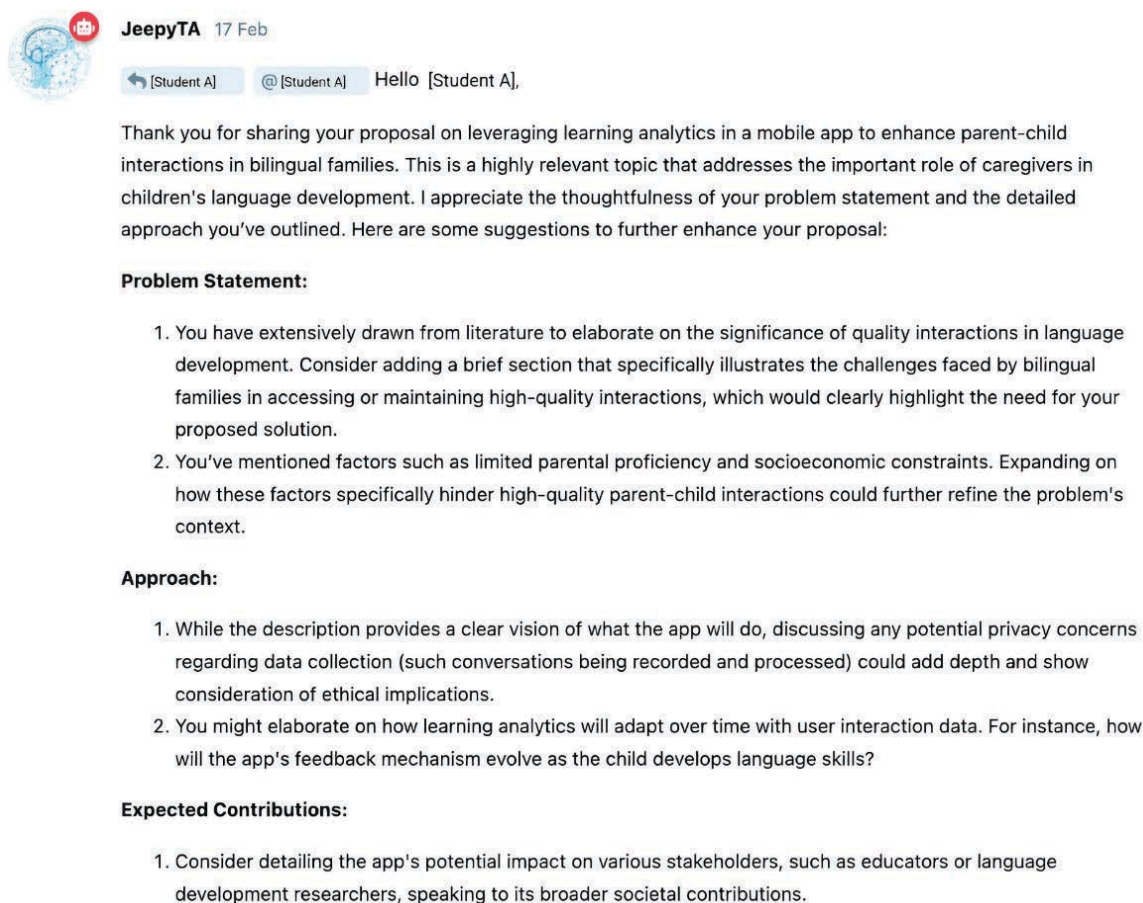
Source:Authors' own work.

Respond to student reflections/questions on the reading/lecture

JeepyTA is also capable of responding to student reflections and questions on the course readings and lectures, offering additional clarification, prompting further thinking, and connecting ideas across course materials. When students share reflections, JeepyTA acknowledges their contributions by reinforcing key ideas from the readings or connecting their insights to broader course themes. If a student raises a question about a concept, theory, or method, JeepyTA provides a response by summarising relevant arguments, explaining terms, or pointing to sections of the readings that address the issue (see Figure 9.3). When a reflection introduces an interesting perspective or critique, JeepyTA may pose follow-up questions to encourage further discussion. To maintain consistency between JeepyTA's responses and the course content, JeepyTA was specifically instructed to first reference the course materials, with specific materials selected based on their similarity score to the student's query, while replying. JeepyTA was also prompted to use course-specific language as defined by instructors before the start of the semester in its responses. If a student's question is not closely related to the course, JeepyTA may be instructed to rely on its knowledge base to respond.

Instructors or TAs can modify the visibility settings of JeepyTA's responses at any time during the semester. If preferred, JeepyTA's responses can be flagged for instructor review before being shared with the students. This option can be turned on or off at any time during implementation. It is also available for selected categories, such as only for answering logistics questions. This helps prevent the provision of incorrect or misleading information, a particular issue in subject areas where misconceptions are highly present on the web and therefore also in the LLM knowledge base. Instructors or TAs can review flagged responses. If a response is inaccurate, they can discard it and reply directly. If the response is mostly correct but needs refinement, they can edit it before posting. When a response is accurate and well-structured, the instructors or TAs can approve it as is. This additional layer allows JeepyTA to provide timely support while making sure students receive information that is accurate, relevant, and aligned with the course objectives.

Figure 9.3. JeepyTA explaining a key detail about an algorithm
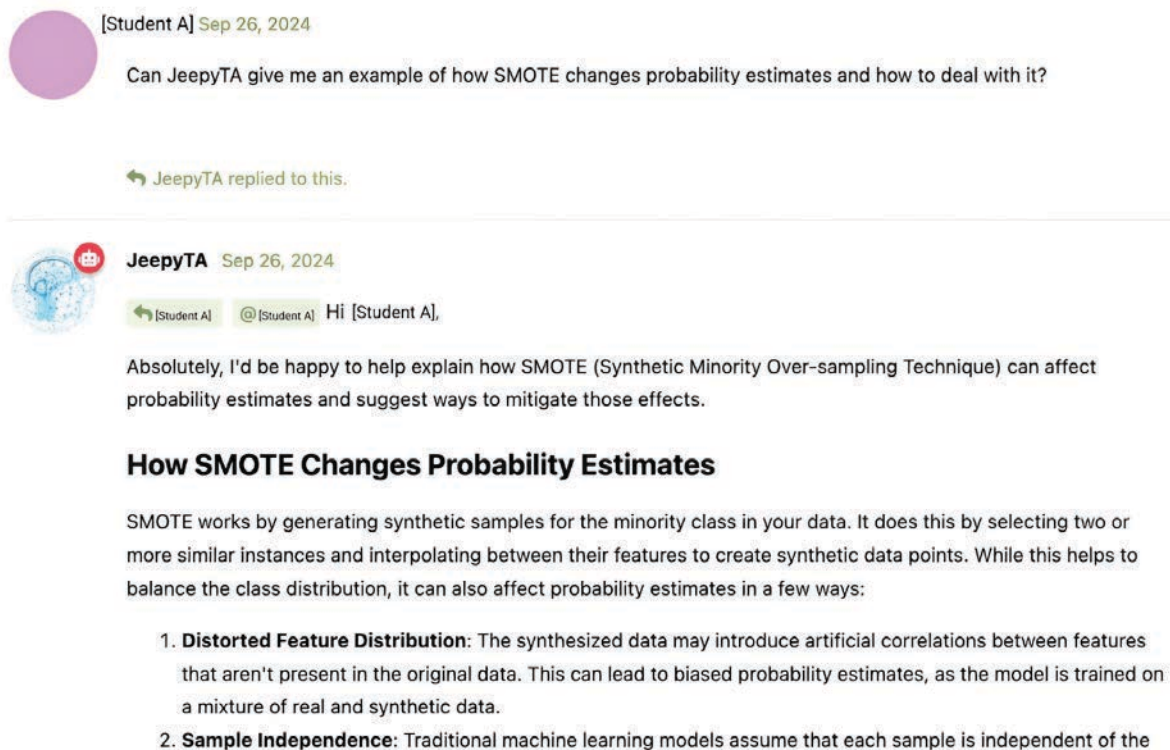
Note: SMOTE is an algorithm used for data pre-processing.
Source:Authors' own work.

Provide debugging support to programming code

For courses involving programming (but not focused on learning to programme), JeepyTA also provides debugging support to students working on assignments by analysing their code and identifying potential errors. When students submit code snippets or describe issues they encounter, JeepyTA reviews the logic, syntax, and structure to pinpoint common mistakes. It then suggests corrections or improvements to resolve the errors (see Figure 9.4).

JeepyTA is generally encouraged to use its existing programming knowledge to solve coding issues while following course-specific conventions or practices. In an Educational Data Mining course, for example, student-level cross-validation is required because this method evaluates how well a model generalises to unseen students. A general LLM chatbot may default to recommending a simple train-test split, a technique that would be acceptable in other contexts but is not the method needed in this course. To prevent this, JeepyTA is instructed to prioritise debugging support based on course materials, assignment requirements, and instructor guidelines rather than relying on broadly used techniques that may not be appropriate in the course context. When addressing programming errors, it refers to the course's preferred approaches and explains why they are used and how they differ from other methods.

In cases where code produces unexpected output, JeepyTA offers strategies for troubleshooting, such as adding print statements, checking variable values, or breaking down complex functions into smaller, testable parts. If students describe the problem rather than submitting code, JeepyTA suggests debugging techniques based on the nature of the issue and guides them through potential causes and solutions.
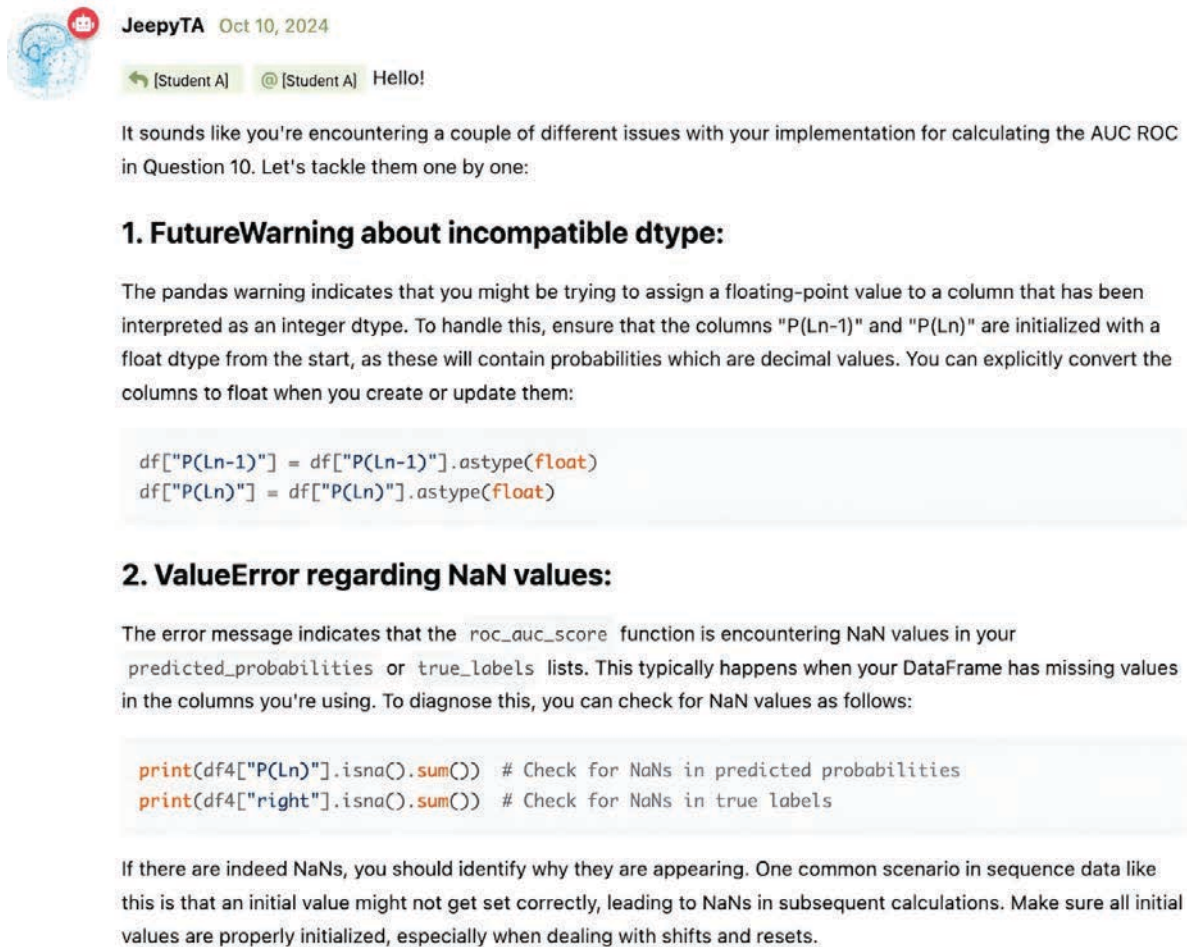
If students describe the problem vaguely or provide incomplete context, JeepyTA asks follow-up questions to clarify the issue before offering suggestions. For example, if a student says, "The code isn't working," without specifying the error message or expected output, JeepyTA prompts them to provide more details, such as the error message received, the intended function of the code, or the steps they have already tried.

When providing programming code support, JeepyTA's prompts are designed to avoid simply providing the correct code but instead focus on helping students understand how to diagnose and fix errors themselves. The prompt design encourages students to learn from their own debugging process, which creates opportunities for them to build confidence and capacity to read errors, trace code, and solve problems independently rather than rely on being given the exact fix.

JeepyTA has provided debugging support for two courses across two semesters, but its effectiveness in identifying errors has been limited in some cases. One possible reason is that it did not have access to the datasets students were working on in either implementation, which makes it difficult for JeepyTA to verify data structures, variable values, or dataset-specific errors. On the other hand, in other cases, it has caught unusual mistakes (such as a student using the symbol \ instead of |) and typos which can be difficult for instructors and human TAs to see in a lengthy programme (see Figure 9.4). Even when JeepyTA is unable to pinpoint the exact cause of issues, it has helped students clarify their problems and suggests general debugging strategies. This still reduces the time instructors or TAs need to spend guiding students through the initial steps of troubleshooting.

Figure 9.4. JeepyTA explaining the errors the student encountered while solving a programming problem and giving advice for diagnosis

JeepyTA Oct 10, 2024

↩ [Student A]   @ [Student A]   Hello!

It sounds like you're encountering a couple of different issues with your implementation for calculating the AUC ROC in Question 10. Let's tackle them one by one:

1. FutureWarning about incompatible dtype:

The pandas warning indicates that you might be trying to assign a floating-point value to a column that has been interpreted as an integer dtype. To handle this, ensure that the columns "P(Ln-1)" and "P(Ln)" are initialized with a float dtype from the start, as these will contain probabilities which are decimal values. You can explicitly convert the columns to float when you create or update them:

```
df["P(Ln-1)"] = df["P(Ln-1)"].astype(float)
df["P(Ln)"] = df["P(Ln)"].astype(float)
```

2. ValueError regarding NaN values:

The error message indicates that the `roc_auc_score` function is encountering NaN values in your `predicted_probabilities` or `true_labels` lists. This typically happens when your DataFrame has missing values in the columns you're using. To diagnose this, you can check for NaN values as follows:

```
print(df4["P(Ln)"].isna().sum())  # Check for NaNs in predicted probabilities
print(df4["right"].isna().sum())  # Check for NaNs in true labels
```

If there are indeed NaNs, you should identify why they are appearing. One common scenario in sequence data like this is that an initial value might not get set correctly, leading to NaNs in subsequent calculations. Make sure all initial values are properly initialized, especially when dealing with shifts and resets.

Source: Authors' own work.

Summarise the discussion posts for each week

JeepyTA has also been applied to generate summaries of discussion forum conversations. When the use case was first introduced in Spring 2024, summaries were provided only to instructors and TAs to give them an overview of the student discussions. As of Spring 2025, in some courses, these summaries are accessible to all students on the forum. After each weekly discussion, JeepyTA summarises key themes, groups discussions into meaningful themes and recurring arguments, and identifies important questions from the discussion forum. In doing so, JeepyTA credits students who introduced specific points in order to give the instructor and TAs a sense of specific student participation as well as overall trends (See Figure 9.5).

The purpose of these summaries is not to replace reading or participating in discussions but to provide an additional layer of support in organising and reflecting on what was discussed. Students are still expected to engage in the full conversation, but the summary can help identify patterns, highlight areas of agreement and disagreement, and uncover questions that might require further discussion. In other words, instead of replacing direct engagement, the summary function can serve as a tool to make the overall direction of conversations more accessible.

Figure 9.5. JeepyTA summarising the weekly discussion within the forum

JeepyTA 30 Jan Edited

↰[Teacher A] @[Teacher A] Hello [Teacher A],

It's wonderful to see such dynamic and insightful engagement with this week's readings. Let's synthesize the themes and questions that have emerged from the discussions.

Common Themes

1. **Applications and Suitability of QE**: Several students, including [Student B and Student C] , highlighted QE's potential across various fields—such as education, social sciences, healthcare, and online discourse—due to its ability to combine qualitative depth with quantitative analysis. A common refrain is its effectiveness in contexts rich with complex data patterns, like collaborative learning and identity exploration.

2. **Methodological Decisions**: [Student D and Student E]   discussed the numerous key methodological choices that researchers face, ranging from data segmentation and sampling to coding and defining units of analysis. This theme ties into the broader challenge of balancing qualitative and quantitative demands.

3. **Transparency and Standards**:   [Student F]   emphasized the necessity for transparency in QE methodology to bolster trustworthiness and reproducibility. This includes clear documentation of process, terminology, and coding decisions. This call for transparency aligns with the need for conceptual clarity mentioned by several participants.

4. **Integration and Reflexivity**: Multiple commenters noted the central role of critical reflexivity in QE, as [Student G and Student H] pointed out. This is seen as vital due to the centrality of the researcher in guiding study outcomes through their interpretations.

5. **Quantification of Qualitative Data**:   [Student I]   have voiced intrigue regarding how qualitative data transitions into quantitative formats. This reflects the core of QE's dual methodology and sparks curiosity about practical implementations and visualizations.

Frequently Raised Questions

- How does data get sorted and make it past certain phases of screening once it passes identification?

Source:Authors' own work.

Suggest ideas for essays and support brainstorming

In a Games and Learning course offered in Spring 2024, JeepyTA was used as a brainstorming partner in two play journal assignments, where students consulted with JeepyTA to propose educational uses for classic games and Minecraft (Shah et al., 2024[46]). A play journal is a structured reflective assignment in which students document and analyse their gameplay experiences to critically engage with classic and contemporary video games by examining their design, narrative, and educational potential. These journals help students develop knowledge of a game through both direct and vicarious experiences, providing a baseline understanding that enables them to propose meaningful educational applications. This structured reflection also prepares students with foundational knowledge of the game, which they can draw upon when consulting with JeepyTA – thus streamlining the labour-intensive process of game analysis and educational integration. While recent initiatives have aimed to make the repurposing of commercial entertainment games for education more accessible (Foster and Shah, 2020[47]), the process still requires significant human effort. This poses challenges for students new to game-based learning and for instructors with limited resources to support student experimentation through direct experience (e.g. playing the game, learning through trial and error) or vicarious experience (e.g. watching YouTube videos, learning from colleagues). JeepyTA offers a solution by enabling users with varying levels of technological, pedagogical, and content knowledge to generate ideas, helping them explore how a game can be adapted for specific educational contexts and learning goals. Students can be guided in developing familiarity with a game and using it in dialogue with JeepyTA (See Figure 9.6).

Figure 9.6. JeepyTA supporting brainstorming and idea generation in a "games and learning" course

JeepyTA Apr 16, 2024

↩ [Student A]   @ [Student A]   Based on the key elements you mentioned, it's clear that the visual style and sandbox nature of Minecraft have impacted your gaming experience. Here are a few suggestions for exploring the application of Minecraft in education based on these elements:

1. Visual Design: While the pixel style may not be appealing to everyone, it can still be leveraged for educational purposes. Consider exploring how Minecraft's visual design can be used to enhance creativity and encourage artistic expression. This could involve designing and building pixel art structures or recreating famous works of art within the game.

2. Sandbox World and Freedom: Minecraft's open-ended gameplay provides opportunities for self-directed learning and problem-solving. You could explore how this aspect of the game can support inquiry-based learning or project-based learning in the classroom. Encourage students to set their own goals and explore topics of interest within the game, allowing them to develop critical thinking and problem-solving skills.

3. Collaborative Play: Minecraft's multiplayer mode allows for collaborative play experiences. Consider exploring how Minecraft can be used to foster collaboration and teamwork among students. Assign group projects or challenges within the game where students have to work together to achieve a common goal. This can help develop communication skills, teamwork, and cooperation.

4. Content Creation: Minecraft's world-building capabilities can be utilized to enable students to create their own educational content. For example, students can build historical landmarks or recreate scenes from literature. This hands-on approach allows students to actively engage with the content and provides opportunities for deeper understanding and retention of knowledge.

I hope these suggestions are helpful! Let me know if you have any further questions or if there's anything else I can assist you with.

Source: Authors' own work.

Personas in a course

In a course on cultural foundations for teaching and learning offered in Fall 2024, JeepyTA interacted with students through personas, discussing scenarios and stories related to their cultural experiences with teaching and learning.

The persona prompts for JeepyTA consisted of three main components: persona description, situation, and instruction. The persona description section defines the persona's name and role (e.g. Felipe, a teacher educator), the context in which the persona operates (e.g. teaching elementary education majors), and the personal and cultural background that reveals their funds of knowledge, such as home language, family activities, cultural rituals, and hobbies (Gonzalez, Moll and Amanti, 2006[48]). The situation section contains the persona's role in the interaction (e.g. providing feedback, answering questions, mentoring), the task or topic being discussed (e.g. reviewing drafts on a specific subject), and the participants in the conversation (e.g. the persona interacting with a preservice teacher). Finally, the instruction section provides information regarding the tone of voice (e.g. formal, informal, supportive, critical), the level of detail required in responses (e.g. detailed feedback with personal experiences), and specific elements to include, such as relevant examples.

For each persona defined on the forum, a separate sub-forum category was created using the persona's name. In these categories, JeepyTA responded while acting as the corresponding persona. Students were informed about the personas and instructed that JeepyTA would post there as the persona indicated by the category name.

Since the main goal of this course is to situate novice preservice teachers in culturally relevant and sustaining teaching pedagogies (Ladson-Billings, 1995[49]; Paris, 2012[50]), the persona descriptions intentionally emphasise the cultural and linguistic backgrounds to avoid generating general responses that do not centre specific groups of learners. Thus, the personas were developed based on the lived experiences of four individuals who were members of historically underrepresented groups whom the team interviewed (e.g. Mexican American, Hmong American). These individuals also reviewed the draft persona descriptions to ensure that their identities were accurately portrayed and to avoid racial essentialisation (Omi and Winant, 2018[51]).

For instance, one persona, 'Claire,' who identified as Hmong-American, shared personal stories related to food, spirituality, and family history, providing preservice teachers with concrete examples that fostered a deeper understanding of Hmong learners. Additionally, the personas guided preservice teachers in understanding concrete ways to incorporate funds of knowledge into classroom settings. For example, the persona 'Felipe' suggested specific strategies for adapting class materials to align with Mexican-American families, such as incorporating family tree activities and introducing home craft projects (see Figure 9.7). These approaches offered practical and culturally responsive methods for connecting multicultural perspectives to classroom instruction.

Generate a discussion prompt to start the weekly discussion

One use case being piloted in Spring 2025 is the generation of discussion prompts to start weekly discussions based on the assigned readings. The goal is to provide a foundation for meaningful discussions while maintaining relevance to the course objectives.

When generating the discussion prompt, JeepyTA extracts core arguments, methodologies, and debates from the week's readings. It identifies recurring themes, unresolved questions, or contrasting viewpoints and frames them in a way that encourages meaningful discussion. If students are expected to relate the reading material to their own research, JeepyTA includes questions that prompt reflection on personal experiences or future applications. If the goal is to explore methodological issues, it focuses on the strengths, limitations, and assumptions underlying the methods presented in the readings (see Figure 9.8).

Some prompts JeepyTA suggests may not be immediately useful. For example, it has generated questions that are overly broad and more at the level of the entire course than a specific week's content. Other discussion questions generated may be too complex and require extensive background knowledge or additional explanation before students can engage with them. A highly detailed methodological critique, for example, may be difficult to address within the scope of a discussion forum. Therefore, instructors and TAs have reviewed all of the discussion prompts before making them visible to students. JeepyTA's suggestions have provided a starting point that allows instructors and TAs to refine the wording, adjust the focus, or simplify overly technical questions to improve understandability. As such, JeepyTA does not replace instructor or TA expertise in orchestrating discussions, but rather, it streamlines the process by offering an initial draft that helps structure each week's forum.

Figure 9.7. JeepyTA acting as a Mexican American persona, "Felipe", bringing a specific persona to recommendation

JeepyTA Nov 22, 2024

↩[Student A]   @[Student A] To incorporate 'funds of knowledge' effectively into your design project for teaching 5th-grade students, you need to first ensure that you are thoroughly familiar with the cultural, social, and economic backgrounds of your students. This understanding will allow you to tailor your project to reflect and utilize the knowledge and skills that students bring from their home environments. Here are a few detailed suggestions based on my experiences as a student:

1. **Cultural Relevance:** When I was in 5th grade, classroom materials that included or were relevant to Mexican culture could have helped me engage more fully with the content. For your project, consider how you can integrate examples, case studies, or stories from the students' own cultures. For instance, if teaching a history lesson, include historical figures and events from the cultures represented in your classroom.

2. **Family and Community as a Resource:** In my upbringing, family played a crucial role. Involving families in the learning process can greatly enrich classroom learning. You might design projects or assignments that require students to interview family members or involve them in building a family tree, which teaches social studies and personal history.

3. **Language and Communication:** Recognize the linguistic assets students bring into the classroom. Using students' first languages when possible can aid understanding and show respect for their linguistic background. For example, providing bilingual resources or allowing students to present work in their first language can make learning more accessible and engaging.

4. **Skills from Home:** Many students develop specific skills through their home life—be it cooking, mechanical repairs, or other crafts. Projects that allow students to demonstrate and apply these skills can validate their non-academic knowledge and show its relevance to school subjects. For example, if a student's family is involved in construction, a math project could involve calculating materials needed for a model house.

Source:Authors' own work.

Figure 9.8. JeepyTA suggesting discussion questions to start off a weekly discussion on coding qualitative data in the "Quantitative Ethnography and Epistemic Network Analysis" course

JeepyTA 13 Feb Edited

Hello dear class! I have some engaging themes, topics, and guiding questions based on the readings for Week 6:

1. **Developing a Coding System:**

   ○ What are the challenges and benefits of developing your own coding system versus using a predefined one (e.g., what David mentioned in the QE book: inductive/bottom-up and deductive/top-down) in QE research?

   ○ How do the readings suggest balancing automation (like using AI tools) with human expertise when coding qualitative data?

2. **The Role of Codes in QE:**

   ○ What is the relationship between codes and the research questions in QE studies? How do the readings illustrate this relationship? How do you see this relationship in your only work?

3. **Ethical Considerations in Coding:**

   ○ What ethical considerations should researchers keep in mind while developing a coding scheme?

4. **Code Consistency and Integrity:**

   ○ According to Shaffer & Ruis (2020), how do researchers ensure "fairness" in the coding process across different datasets or studies? Give it an example of what it might look like in your work.

These themes and questions aim to create a robust dialogue around the complexities and innovations in coding for quantitative ethnography. I hope they inspire deep and meaningful discussions in your class!

Best,
Jeepy.

Source:Authors' own work.

Evaluation of JeepyTA

JeepyTA was first implemented in Fall 2023 and covered use cases described above. To understand how students viewed the virtual teaching assistant, we distributed a voluntary end-of-semester survey, approved as exempt by the university's Institutional Review Board, and clearly stated that participation was optional and would not affect grades. Students provided informed consent before answering 13 multiple-choice questions. The survey asked about specific aspects of JeepyTA's performance, including how quickly and accurately it responded to questions, how clearly and professionally it communicated, and how well it supported student learning, development, and motivation, compared to a human TA. Response options ranged from 1 ("Human TA is significantly better") to 5 ("AI TA is significantly better"). We used two-sample t-tests to check whether the average scores for each question differed from the neutral midpoint of 3. This allowed us to measure whether students viewed JeepyTA as better or worse than a human TA. Results showed that students rated JeepyTA as comparable to a human TA in several areas, including the speed and clarity of its replies, the accuracy and professionalism of its responses, its ability to support learning without giving away answers, and the overall usefulness and quality of its feedback. However, students rated it lower in three areas: offering useful ideas, supporting student development, and motivating students.

We also evaluated JeepyTA's impact on when responses were available to students. In two identical courses offered one year apart, taught by the same professor at the same institution and involving comparable student groups, a difference emerged in response times to students' queries. During the earlier term, when JeepyTA was not in use, the instructor and the TAs posted 153 responses, with a median response time of 7.09 hours. In the following term of Fall 2023, after JeepyTA was introduced, course staff posted 136 responses, and the median response time dropped to 2.23 hours, a statistically significantly lower time.

Among 89 posts where JeepyTA attempted to answer student inquiries, 22 responses received approval. The AI generated replies in approximately 40 seconds, and course staff approved them within an average of 38 minutes. Because JeepyTA was able to handle these queries, even manually written responses were faster. In the JeepyTA semester, the median human response time was 4.14 hours, statistically significantly faster than the 7.09-hour median time observed in the prior semester. This suggests that JeepyTA improved the efficiency of even fully-human responses, likely because the instructor and TAs had more time to address tasks that deserved more of their attention.

Another of JeepyTA's intended purpose was to support instructors and TAs outside standard working hours. In the term prior to its introduction, course staff posted 62% of replies outside regular U.S. business hours (after 5 p.m. and before 9 a.m.). In the term when JeepyTA was available, this proportion was 60%, not statistically significantly different. However, there appeared to be a difference in the proportion of responses posted during weekends. In the prior semester, instructors posted 10% of all replies on weekends. After JeepyTA was introduced, this number increased to 29% of the total number of replies (including JeepyTA posts approved or edited by the instructors). This increase was statistically significant, suggesting that course staff were better able to focus their work time even on weekends, following the introduction of JeepyTA.

Work is currently ongoing to study the impact of JeepyTA's feedback on student essays (3b). In our initial work, we have found that semester-on-semester, students receive statistically significantly higher assignment grades (according to an independent grader) after receiving JeepyTA's feedback – going from an average of 64% of students receiving an A or an A+ on their final submission to 95% of students receiving an A or A+. In follow-up work, we are investigating whether students specifically fix the issues identified by JeepyTA in that same essay, and whether they make the same mistakes in subsequent essays (including in a different class where JeepyTA is also offered).

A study on JeepyTA's brainstorming support functionality (3f) (Shah et al., 2024[46]) revealed that JeepyTA helped students generate a higher volume of ideas (averaging 2.78 per student compared to 1.7 for student-led ideation) and increased the production of fully formed, detailed concepts. However, this came with important trade-offs: JeepyTA-driven ideas often showed thematic overlap, with common suggestions like "teamwork and collaboration" appearing across multiple students' work, while student-driven ideas exhibited greater diversity (cf. Doshi and Hauser, 2024[52]). In analysing students' descriptions of their process of using JeepyTA, the researchers identified five distinct patterns: 36% of students found the suggestions insightful and aligned with their own ideas, 18% acknowledged JeepyTA's role in idea generation but did not say if it was actually helpful in doing so, 18% referenced using JeepyTA but did not say how/if its use was connected to their proposed ideas, 18% made no reference to JeepyTA, and 13% critically evaluated and built upon JeepyTA's recommendations. Notably, 68% of students proposed multiple educational applications spanning knowledge types essential for 21st-century learning (Kereluik et al., 2013[53]): meta-knowledge (collaboration, problem-solving), foundational knowledge (mathematics, history, computer science), and humanistic knowledge (digital citizenship, ethical awareness). Especially novel recommendations emerged when students integrated insights from readings, game experiences, and their domain expertise, suggesting that JeepyTA works best as a brainstorming tool when students already possess foundational knowledge they can build upon (Nasiar, forthcoming[54]).

Other case studies of AI-augmented TAs

Example from India

At the *Indian Institute of Technology Kanpur,* for instance, researchers have piloted AI augmented TAs for an introductory computer science course, where student-to-instructor ratios are often too high for human instructors or TAs to provide individual guidance at scale (Ahmed, 2025[55]). In response to this situation, the team integrated an AI agent into Prutor, a web-based programming platform used by students to submit solutions for C programming assignments. When a student's programme failed to pass the instructor-defined test cases, they could request assistance by clicking a "Get Help" button within the platform. This action triggered a feedback request that compiled four key inputs and sent them to GPT-4 Turbo via an internal API: the problem description, the student's buggy code, the test case results, and an optional message written by the student to describe their confusion or ask a specific question. Using this information, GPT-4 Turbo generated targeted feedback linked to specific lines in the student's code that highlights the exact locations of potential errors and explaining what may have gone wrong. The output was routed to a centralised dashboard, where human TAs could review the AI's draft response, make edits, add notes, or reject it entirely before sending the final feedback back to the student through the same interface. The AI agent was evaluated against two other conditions: one in which human TAs provided all feedback without any AI support, and another in which feedback generated by GPT-4 Turbo was sent directly to students without human review. Researchers examined how these types of feedback methods affected feedback quality (measured through expert evaluations), TA efficiency

(measured through response times recorded in system logs), and student performance (measured by whether final code submissions passed all instructor-defined test cases). In addition, students rated each piece of feedback based on helpfulness, clarity, and timeliness using built-in rating tools on the platform. Although AI-generated feedback was often rated favourably by students – particularly for its detailed explanations – these positive perceptions did not consistently lead to improved performance. Students receiving AI-assisted or fully automated feedback were not significantly more likely to complete the assignments successfully, and in many cases, manual TA feedback led to faster and more effective problem resolution.

One issue observed by the researchers was that some TAs using AI-generated feedback forwarded it without making necessary corrections, even when the output contained inaccuracies or hallucinations. This appeared to reflect a tendency among certain TAs to rely too heavily on the AI drafts instead of critically evaluating their quality. In contrast, TAs working without AI support often provided responses which highlighted the immediate next step or pinpointed the specific source of the error. As a result, students in the manual TA group were, in some cases, able to resolve issues more efficiently, despite often receiving shorter and less detailed feedback.

Example from Czechia

Another example comes from Czechia, where a GPT-3-based conversational chatbot named Alex was deployed in a university-level English course (Polakova and Klimova, 2024[56]). Alex is a web-based application that combines multiple AI models: GPT-3 is used to generate natural language responses, while Gramformer and T5 are applied to detect and correct grammatical errors. At the beginning of each chat session, GPT-3 generates a topic-specific opening question based on preselected weekly themes. Students then respond freely in English. Their input is processed by Gramformer and T5 to identify grammatical mistakes. When an error is detected, the chatbot enters a correction phase, during which GPT-3 provides a corrected version of the sentence along with an explanation in natural language. The chatbot also allows users to rate the feedback as either "good" or "bad". Over a four-week period, students engaged with Alex in simulated dialogue sessions and received real-time feedback. Though the topic of each session starts with a fixed opening, users can take the conversation in any direction. To constrain the types of interactions, users were limited to one session per day and three per week, with each session requiring a minimum of 1 000 characters of typed input. After completing the four-week programme, students filled out a questionnaire survey about their experience. Analysis of student feedback revealed that learners responded positively to several aspects of the chatbot. Students noted that the chatbot asked clear, easy-to-understand questions and responded quickly, which helped keep the conversation at a natural pace. Many also appreciated the flexibility to practice outside the classroom. The option to access conversation practice on their own helped them gain confidence. According to the survey, 88% of students reported they did not feel stressed using the chatbot, and several commented that the experience felt like chatting with a real person. In terms of learning gains, pre- and post-tests focusing on grammar and vocabulary showed measurable improvements. Upper-intermediate students improved their tests scores from about 59% to 75%, while advanced students increased from 80% to 90%.

At the same time, students also reported several limitations of the chatbot that affected their overall experience. Some participants encountered technical problems, such as system lags and incomplete replies from the chatbot, which disrupted the flow of conversation. Others pointed out that certain responses felt repetitive or too limited in variation, which reduced the usefulness of later sessions. Survey results also showed that the chatbot failed to increase motivation or encourage continued use: 74% of students did not feel more motivated to learn English as a result of using the chatbot, and 79% said they would rather use other tools like Duolingo or talk to native speakers instead.

Though different in design and pedagogical goals, these examples share JeepyTA's aim of delivering coursealigned, scalable support. Even though these AI-powered teaching assistants differ from JeepyTA in how they function, the courses they support, the tasks they were assigned, and in the regional infrastructure of the learning environment, similar benefits and challenges can be noted.

Reflection on generative AI as teaching assistants: implications and policy recommendations

The emergence of generative AI (GenAI) in education signals a shift in the professional landscape of education, decentralising some traditional instructional roles and prompting a rethinking of what it means to teach. Historically, expertise in education has been concentrated in instructors and (to a lesser extent) human TAs, who provide guidance, feedback, and assessment. However, as GenAI systems become increasingly capable of tutoring, scaffolding learning, and responding to student needs in real time, the boundaries of these roles are being redrawn.

Rather than replacing human expertise, AI TAs built using GenAI demand that we rethink how instructional work is distributed–not just between instructors and human TAs, but across AI-enhanced systems. This shift mirrors what Shaffer, Nash, and Ruis (2015[57]) describe as the reconfiguration of professional expertise in response to new tools and technologies. As they argue, professionalisation is not static; it evolves when new ways of knowing and working emerge. In the case of education, AI's growing presence means that instructors must develop new competencies – not only in teaching content, but in orchestrating AI-enhanced learning environments, interpreting AI-generated insights, and ensuring alignment between AI feedback and pedagogical goals. As such, policy should emphasise training of instructors and human TAs to work effectively with AI.

It is also critical to ensure that GenAI integration does not drive pedagogy but rather supports it in meaningful ways. The Technological Pedagogical Content Knowledge (TPACK) framework (Koehler, 2009[58]) provides a useful lens for understanding this challenge. Educators must consider how GenAI interacts with both content knowledge (what is being taught) and pedagogical strategies (how it is taught). Without thoughtful integration, there is a risk that AI could push education further toward efficiency-based models, where rapid feedback and automated assessments replace deeper engagement with complex ideas rather than finding an optimal combination of deeper learning and more automated activities where appropriate (Mishra, Warr and Islam, 2023[59]). Prioritising efficiency and engagement over meaningful understanding, may increase the quality of student work and student experience in the short-term but may not benefit the student in the longer term.

Mishra and colleagues' work also highlights the need to move beyond mere adoption of AI tools to meaningful integration into learning experiences. The presence of AI TAs does not inherently improve education; their effectiveness depends on how they are aligned with broader learning goals. Educators must take an active role in shaping AI's function within courses, ensuring it complements and enhances human-centred teaching practices rather than supplanting them. Therefore, we recommend against designing of AI Teaching Assistants in ways that replace humans and fully automate all learning activities, and the adoption of cost-cutting measures that create pressure to eliminate human TAs should also be avoided. Beyond reducing the quality of instruction, reduced funding for Teaching Assistants would also decrease opportunities for economically disadvantaged individuals who rely on Teaching Assistant positions as a pathway into academia, ultimately reducing the pipeline of talented scholars into research and scholarship.

Henriksen and Mishra's (2024[60]) work on practical wisdom further reinforces this perspective, emphasising that experienced educators bring a form of professional knowledge that AI cannot replicate - one rooted in ethical decision-making, contextual understanding, and reflective practice. As AI transforms the nature of knowledge in education, teachers must ensure that human judgment, adaptability, and social-emotional insights remain at the core of teaching. This highlights the need for educators to approach GenAI critically, leveraging its strengths while maintaining the core humanistic elements of teaching and mentorship.

Reprofessionalisation in this context is therefore not just about preparing educators for an AI-integrated classroom–it is about ensuring that humans and AI systems work in sync to advance student learning experiences and outcomes. Just as previous technological shifts reshaped the teaching profession, GenAI requires a reimagining of teacher preparation, assessment design, and professional collaboration. The goal is not merely to integrate AI, but to define new models of expertise in an AI-augmented educational ecosystem–one in which human and AI agents work together to support meaningful learning experiences while maintaining a commitment to equity and effectiveness.

Implications for student assessment

Assessment practices in higher education serve multiple purposes: they inform students about their progress, provide instructors with actionable insights to guide instruction, and certify learners' competence. The emergence of AI Teaching Assistants (AI TAs) such as JeepyTA adds new opportunities for formative assessment, while creating possibilities for summative assessment which merit some caution. In both cases, careful design and policy guardrails are necessary to ensure that AI TAs enhance rather than undermine the educational process.

A chief benefit of AI TAs is in supporting formative feedback, where fast, specific, and individualised guidance can promote deeper learning. This feedback can be provided both immediately and 24/7. Students who work late at night, study remotely, or juggle other responsibilities often cannot attend regular office hours or wait for TAs to become available. An AI TA can fill this gap by providing immediate, round-the-clock feedback, easing the pressure on human TAs and making support more equitable.

Historically, automated assessments (e.g. quizzes, short-answer grading) have helped identify student misconceptions and encouraged targeted practice. Generative AI now expands these possibilities by supporting more complex tasks– from essay drafts to coding projects. In general, an AI TA can assess a broader range of competencies than a typical

human TA, supporting shifts in assessment from assessing what students know to their conceptual understanding and their process of solving problems. Importantly, using AI for assessment also opens the possibility of assessing student learning through a broader range of artifacts – including annotations, reflections, peer feedback, conversations, and other forms of evidence – enabling a more holistic and nuanced view of learning that extends beyond traditional measures.

They can also provide sophisticated, multi-dimensional assessment. As seen in the JeepyTA platform discussed earlier, AI TAs can offer feedback on essay organisation, argument clarity, and conceptual rigor, referencing course rubrics and standards to align with the instructors' goals. This kind of timely, actionable commentary can help learners iterate more quickly, moving from basic correctness checks toward higher-order thinking and reflection (Hattie and Timperley, 2007[6]; Chi and Wylie, 2014[7]). If these systems are designed carefully to align with the course-specific materials and rubric (as JeepyTA has been primed to do), it is possible to reduce the risk of misleading feedback that the instructor will disagree with–although it still happens, just as a human TA could also provide feedback that an instructor disagrees with. Furthermore, the use of personas (as discussed above) or carefully designed prompting can offer students feedback from different perspectives, highly relevant in some disciplinary areas but difficult for a single human TA or instructor to provide.

With AI TAs taking on time-consuming tasks such as answering routine questions, reviewing initial drafts, or providing rubric-aligned suggestions, human TAs and instructors are increasingly able to reallocate their time toward more pedagogically meaningful and relationship-centred activities. These include leading in-depth discussions that challenge students to think critically, working directly with individuals or small groups to support their academic progress, meeting individually with students to support their academic and professional growth, and developing activities that promote academic agency and collaboration. Academic staff can now invest more energy in synthesising performance patterns across student submissions, identifying emerging misconceptions, and making ongoing improvements to assignments or assessments based on observed trends and student needs. These higher-order instructional practices remain difficult for AI systems to replicate, though they can support these tasks in various ways. Rather than displacing humans, then, AI TAs can create space for them to focus on tasks that require interpreting student thinking in context, applying subject-matter expertise, and exercising instructional judgment.

Another possible concern with the provision of formative feedback by AI TAs is over-scaffolding, where learners rely so heavily on AI-generated suggestions that their final products no longer represent independent work. Particularly if students can seek several rounds of feedback, or obtain lower-level writing suggestions, the sophistication of today's generative AI can obscure the boundary between a student's own efforts and AI-provided content. To avoid issues of this nature, careful consideration of design is needed, which can be supported by policy funding the establishment of guidelines for how much scaffolding is appropriate in different situations and research on methods for producing the benefits of scaffolding while avoiding over-scaffolding. Nonetheless, on the whole, the benefits of providing formative feedback through an AI TA seem – if carefully designed – do outweigh the risks, and policy should encourage higher education to move forward in using AI TAs for formative feedback in ways that can improve student learning.

Greater concerns are present for more summative assessment, such as assigning final marks on assignments. Automated essay scoring has a history dating back decades (Shermis and Burstein, 2003[18]), offering consistency and quick turnaround but often criticised for focusing on superficial textual features. Still, the use of automated essay scoring offers many lessons in how automated grading can be psychometrically validated and used appropriately in ways that support higher education institutions in offering high-quality consistent grading at lower human cost. More advanced, generative AI-based approaches can analyse content in greater detail than most earlier methods, but can be prone to systematic biases, a lack of transparency in scoring, treating inaccurate but widely believed misconceptions as true, and even in some cases have "hallucinations" of information not in the original student work. For high-stakes decisions such as course grades, even small error rates can have consequences for student outcomes and perceptions of fairness. Moreover, incorporating AI into summative grading can amplify existing equity concerns. If an AI TA's underlying model has been trained on data reflecting cultural or linguistic biases, students from underrepresented backgrounds may be inadvertently penalised. The presence of generative AI in summative grading therefore necessitates robust validation across diverse student populations and consideration of whose perspectives are centred within assessment (Lee, 1998[61]) and even in the evaluation of the fairness of assessments (Randall, 2023[62]). Policymakers and institutional leaders must establish policies that ensure that any summative use of AI-based scoring be supported by transparent procedures, documented reliability metrics (including evidence that there is limited or no algorithmic bias (Baker, Hawn and Lee, 2023[63]), and the ability for students (and instructors) to question and appeal automated scores.

Decisions about when AI assistance is acceptable – and how much AI-driven contribution is too much – will differ by course context and disciplinary standards, as well as how far along students are in their development of expertise. In fields like computer science or business, where collaborative problem-solving with tools is central, it may be appropriate to evaluate students in more advanced classes based on how effectively they leverage AI to reach correct solutions. By contrast, in courses emphasising individual mastery of foundational skills, unmediated AI assistance could undermine the competencies being assessed. As such, it may be appropriate to develop guidelines, either at a disciplinary level or in the context of specific standard courses, for which tasks should incorporate AI support, and what types of formative assessment and support are warranted.

Overall, there is considerable potential for the use of generative AI-based Teaching Assistants to support formative assessment, and some possibilities in summative assessment as well, if approached with sufficient caution and human oversight is retained. The design of policy to encourage appropriate use has potential benefits for both students and instructors.

Implications for equity and policies to support equity

The simple existence of AI TAs can play an important role in supporting equity. Many students struggle to get academic help when they need it, not because they do not have the motivation but because structural barriers make access difficult. Some students have jobs or caregiving responsibilities that prevent them from attending office hours. Others study in different time zones and cannot reach instructors or TAs outside of regular class hours. AI TAs can straightforwardly remove some of these barriers by providing immediate responses to course-related questions at any time of day, which makes academic support more accessible to students who might otherwise face difficulties getting assistance when they need it. If designed correctly (and if such design effort is made part of the criteria in requests for proposals and other instruments encouraging the development and use of AI TAs), AI TAs can also be more effective at avoiding unconscious bias, where historically underrepresented students may receive different and lower-quality feedback than other students. When AI TAs are built on large language models that are multilingual, it also becomes possible to provide content in multiple languages, increasing the accessibility of feedback and support for some international students.

Similarly, AI TAs can relieve some equity issues impacting human TAs. Large courses create significant demands on Teaching Assistants, who must divide their time between grading, responding to student questions, and assisting with course administration. Many Teaching Assistants take on these responsibilities while managing their own coursework, research, and professional development. The time constraints they face often affect not only their own studies but also the level of detail they can provide in feedback or the number of students they can support individually. AI TAs can reduce some of these pressures by handling routine inquiries and generating structured feedback on assignments. Therefore, policy to adopt AI TAs where appropriate can have fairly rapid and sizable benefits for equity, for both students and Teaching Assistants.

However, the adoption of AI TAs can also create equity concerns. Holstein and Doroudi's (2021[64]) research suggests that educational technologies often reinforce existing inequities, benefiting students who already possess strong academic skills while leaving others behind. That said, recent studies on large language models indicate a more complex dynamic: LLMs may provide greater relative benefits to less knowledgeable users, a pattern sometimes referred to as the GPS effect (Chiang et al., 2024[34]). Just as GPS systems support those unfamiliar with a route more than experienced drivers, LLMs can scaffold novices more effectively than they assist experts. This suggests that, once access is secured, AI TAs could disproportionately benefit those with less prior knowledge – potentially narrowing achievement gaps rather than widening them. Of course, access itself remains a critical barrier, particularly in terms of digital literacy, language fluency, and reliable connectivity. Moreover, the risk remains that AI systems might encode cultural, linguistic, or epistemic biases that privilege dominant norms. The effort to incorporate specific cultural perspectives and funds of knowledge through personas, discussed above, is our first attempt to address this concern within JeepyTA. It is also important to ensure that AI TAs do not primarily benefit students with higher digital literacy and stronger academic foundations. Ultimately, then, it is important that policy requires evidence that AI TAs are fair for all learners – not just supporting development but validation of fairness across learners as part of evaluation.

Conclusion

This chapter has examined how AI Teaching Assistants, particularly those driven by generative AI, can expand and support the traditional roles played by human TAs and instructors in higher education. We began by considering the foundational role that human TAs play, along with the logistical and pedagogical challenges they often encounter, such as balancing workload with their own academic commitments. Against this backdrop, we reviewed the emergence of AI TAs–tools that are designed to automate routine administrative tasks, deliver real-time student support, and

provide timely formative feedback at scale in ways that complement and extend human capabilities. Rather than viewing AI TAs as replacements for human TAs, we argue for a model of augmentation–one where AI tools provide support and feedback that is otherwise infeasible for human instructors and TAs alone: support and feedback that is immediate, personalised, detailed, and available around the clock.

To ground these ideas, we discussed JeepyTA – a generative AI system developed by the Penn Center for Learning Analytics at University of Pennsylvania. Deployed across multiple graduate-level courses and multiple institutions and integrated into course discussion forums, JeepyTA supports a variety of instructional needs. These include answering logistical queries, providing formative feedback on essays, assisting students with debugging their code, stimulating creative ideation, summarising discussion posts, and suggesting new discussion prompts for deeper engagement. By embedding course materials and rubrics into an LLM-based system, JeepyTA demonstrates how AI TAs can deliver context-aware and curriculum-aligned responses at scale. Throughout the chapter, we also discussed key design and implementation considerations–such as prompt design, the need for human oversight and review, consideration of ethics and bias, and alignment with policy. These factors are essential for ensuring that AI TAs operate responsibly, transparently, and in service of equitable learning outcomes.

Ultimately, the experiences documented here suggest that well-designed AI TAs can ease pressure on human TAs and academic staff, enhance student engagement, and potentially enhance the quality of learning experiences. Importantly, their effectiveness depends not only on the technical sophistication of generative AI, but on careful integration into pedagogical practices. When thoughtfully deployed, AI Teaching Assistants can help institutions create more scalable, responsive, and personalised educational ecosystems–supporting students, TAs, and instructors in new and meaningful ways.

References

Ahmad, S., M. Alam, M. Rahmat, M. Mubarik and S. Hyder (2022), "Academic and Administrative Role of Artificial Intelligence in Education", *Sustainability*, Vol. 14/3, p. 1101, https://doi.org/10.3390/su14031101. [24]

Ahmed, U. (2025), *Feasibility Study of Augmenting Teaching Assistants with AI for CS1 Programming Feedback*. [55]

Aleven, V., B. McLaren, J. Sewall and K. Koedinger (2006), "The Cognitive Tutor Authoring Tools (CTAT): Preliminary Evaluation of Efficiency Gains", in *Lecture Notes in Computer Science, Intelligent Tutoring Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/11774303_7. [22]

Alsafari, B., E. Atwell, A. Walker and M. Callaghan (2024), "Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants", *Natural Language Processing Journal*, Vol. 8, p. 100101, https://doi.org/10.1016/j.nlp.2024.100101. [42]

Baker, R. (2016), "Stupid Tutoring Systems, Intelligent Humans", *International Journal of Artificial Intelligence in Education*, Vol. 26/2, pp. 600-614, https://doi.org/10.1007/s40593-016-0105-0. [23]

Baker, R., M. Hawn and S. Lee (2023), *Algorithmic Bias: The State of the Situation and Policy Recommendations*, OECD Publishing, Paris, https://doi.org/10.1787/c74f03de-en. [63]

Begley, G., B. Berkey, L. Roe and H. Schuldt (2019), "Becoming partners: Faculty come to appreciate undergraduates as teaching partners in a service-learning teaching assistant program", *International Journal for Students as Partners*, Vol. 3/1, pp. 89-105, https://doi.org/10.15173/ijsap.v3i1.3669. [10]

Berta, M. and J. Pembridge (2019), "Exploring Burnout among Graduate Teaching Assistants", *2019 ASEE Annual Conference & Exposition Proceedings*, https://doi.org/10.18260/1-2--32802. [13]

Brooks, H. (2023), *The History of Test Automation*, https://testrigor.com/blog/the-history-of-test-automation/ (accessed on 2/24/2025). [19]

Bulut, O. and M. Beiting-Parrish (2024), "The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges", *Chinese/English Journal of Educational Measurement and Evaluation*, Vol. 5/3, https://doi.org/10.59863/miql7785. [29]

Chiang, C.-H., W.-C. Chen, C.-Y. Kuan, C. Yang and H.-y. Lee (2024), "Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course", *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2489-2513, https://doi.org/10.18653/v1/2024.emnlp-main.146.   [34]

Chi, M. and R. Wylie (2014), "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes", *Educational Psychologist*, Vol. 49/4, pp. 219-243, https://doi.org/10.1080/00461520.2014.965823.   [7]

Dede, C. and J. Richards (eds.) (2018), *Jill Watson: A Virtual Teaching Assistant for Online Education*, Routledge.   [21]

Doshi, A. and O. Hauser (2024), "Generative AI enhances individual creativity but reduces the collective diversity of novel content", *Science Advances*, Vol. 10/28, https://doi.org/10.1126/sciadv.adn5290.   [52]

Essel, H., D. Vlachopoulos, A. Tachie-Menson, E. Johnson and P. Baah (2022), "The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education", *International Journal of Educational Technology in Higher Education*, Vol. 19/1, https://doi.org/10.1186/s41239-022-00362-6.   [27]

Foster, A. and M. Shah (2020), "Principles for Advancing Game-Based Learning in Teacher Education", *Journal of Digital Learning in Teacher Education*, Vol. 36/2, pp. 84-95, https://doi.org/10.1080/21532974.2019.1695553.   [47]

Gao, R., H. Merzdorf, S. Anwar, M. Hipwell and A. Srinivasa (2024), "Automatic assessment of text-based responses in post-secondary education: A systematic review", *Computers and Education: Artificial Intelligence*, Vol. 6, https://doi.org/10.1016/j.caeai.2024.100206.   [30]

Gašević, D. and L. Yan (2026), "Generative AI for human skill development and assessment: implications for existing practices and new horizons", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [1]

Gonzalez, N., L. Moll and C. Amanti (eds.) (2006), *Funds of Knowledge*, Routledge, https://doi.org/10.4324/9781410613462.   [48]

Graham, S. (2023), "Post-Process but Not Post-Writing: Large Language Models and a Future for Composition Pedagogy", *Composition Studies*, Vol. 51(1), pp. 162-168.   [38]

Haaben, F. (2024 (Unpublished Manuscript), *Teaching With AI: Augmenting Professors With LLMs In An Asynchronous Question Center*, https://doi.org/10.14288/1.0443555.   [43]

Hattie, J. and H. Timperley (2007), "The Power of Feedback", *Review of Educational Research*, Vol. 77/1, pp. 81-112, https://doi.org/10.3102/003465430298487.   [6]

Henriksen, D. (2024), "Teaching, Teacher Education, and Practical Wisdom in the Age of Generative AI", S*ociety for Information TechnologyTeacher Education International Conference*, pp. 779-787, https://www.learntechlib.org/primary/p/224040/.   [60]

Holstein, K. (2021), *Equity and Artificial Intelligence in Education: Will "AIEd" Amplify or Alleviate Inequities in Education?*, https://doi.org/10.48550/arXiv.2104.12920.   [64]

IIITD, A., D. Sethi, N. Gupta, S. Sharma, S. Jain, U. Singhal and D. Kumar (2024), T*AMIGO: Empowering Teaching Assistants using LLM-Assisted Viva and Code Assessment in an Advanced Computing Class.*   [33]

Karim, N., A. Maries and C. Singh (2018), "Exploring one aspect of pedagogical content knowledge of teaching assistants using the Conceptual Survey of Electricity and Magnetism", *Physical Review Physics Education Research*, Vol. 14/1, https://doi.org/10.1103/physrevphyseducres.14.010117.   [9]

Karpouzis, K., D. Pantazatos, J. Taouki and K. Meli (2024), "Tailoring Education with GenAI: A New Horizon in Lesson Planning", *2024 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1-10, https://doi.org/10.1109/educon60312.2024.10578690.   [28]

Kereluik, K., P. Mishra, C. Fahnoe and L. Terry (2013), "What Knowledge Is of Most Worth", *Journal of Digital Learning in Teacher Education*, Vol. 29/4, pp. 127-140, https://doi.org/10.1080/21532974.2013.10784716.   [53]

Khine, M., B. Laszlo and E. and Afari (eds.) (forthcoming), *A Step Towards Adaptive Online Learning: Exploring the Role of GPT as Virtual Teaching Assistants in Online Education*, CRC Press.   [16]

Khosravi, H., S. Shabaninejad, A. Bakharia, S. Sadiq, M. Indulska and D. Gašević (2021), "Intelligent Learning Analytics Dashboards: Automated Drill-Down Recommendations to Support Teacher Data Exploration", *Journal of Learning Analytics*, Vol. 8/3, pp. 133-154, https://doi.org/10.18608/jla.2021.7279.   [17]

Koehler, M. (2009), "What is technological pedagogical content knowledge (TPACK)?", *Contemporary Issues in Technology and Teacher Education*, Vol. 9 (1), pp. 60-70.   [58]

Kurdi, G., J. Leo, B. Parsia, U. Sattler and S. Al-Emari (2019), "A Systematic Review of Automatic Question Generation for Educational Purposes", *International Journal of Artificial Intelligence in Education*, Vol. 30/1, pp. 121-204, https://doi.org/10.1007/s40593-019-00186-y. [31]

Ladson-Billings, G. (1995), "Toward a Theory of Culturally Relevant Pedagogy", *American Educational Research Journal*, Vol. 32/3, pp. 465-491, https://doi.org/10.3102/00028312032003465. [49]

Lee, C. (1998), "Culturally Responsive Pedagogy and Performance-Based Assessment", *The Journal of Negro Education*, Vol. 67/3, p. 268, https://doi.org/10.2307/2668195. [61]

Lee, C., J. Myung, J. Han, J. Jin and A. Oh (2023), *Learning from Teaching Assistants to Program with Subgoals: Exploring the Potential for AI Teaching Assistants*. [44]

Lim, L., S. Gentili, A. Pardo, V. Kovanović, A. Whitelock-Wainwright, D. Gašević and S. Dawson (2021), "What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course", *Learning and Instruction*, Vol. 72, https://doi.org/10.1016/j.learninstruc.2019.04.003. [37]

Liu, M. and F. M'Hiri (2024), "Beyond Traditional Teaching: Large Language Models as Simulated Teaching Assistants in Computer Science", *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pp. 743-749, https://doi.org/10.1145/3626252.3630789. [39]

Li, Y. and X. Hu (2026), "Learning with dialogue-based AI tutors: implementing the Socratic method with Generative AI", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [2]

Maiti, P. and A. Goel (2024), *How Do Students Interact with an LLM-powered Virtual Teaching Assistant in Different Educational Settings?*. [40]

Marincovich, M. and J. Prostko (eds.) (1998), *Preparing Graduate Students to Teach: Past, Present, and Future*, Anker Publishing Company. [5]

Marshman, E., R. Sayer, C. Henderson, E. Yerushalmi and C. Singh (2018), "The challenges of changing teaching assistants' grading practices: Requiring students to show evidence of understanding", *Canadian Journal of Physics*, Vol. 96/4, pp. 420-437, https://doi.org/10.1139/cjp-2017-0030. [8]

Miroyan, M., C. Mitra, R. Jain, G. Ranade and N. Norouzi (2025), "Analyzing Pedagogical Quality and Efficiency of LLM Responses with TA Feedback to Live Student Questions", *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pp. 770-776, https://doi.org/10.1145/3641554.3701965. [41]

Mishra, P., M. Warr and R. Islam (2023), "TPACK in the age of ChatGPT and Generative AI", *Journal of Digital Learning in Teacher Education*, Vol. 39/4, pp. 235-251, https://doi.org/10.1080/21532974.2023.2247480. [59]

Nasiar, N. (forthcoming), "Scaffolding Creativity or Homogenizing Outputs? GPT-based Virtual TA's Impact on Student Ideation in Graduate Education. Manuscript submitted to Special Edition on AI and Education.", *Social Innovations Journal*. [54]

Nicklow, J., S. Marikunte and L. Chevalier (2007), "Balancing Pedagogical and Professional Practice Skills in the Training of Graduate Teaching Assistants", *Journal of Professional Issues in Engineering Education and Practice*, Vol. 133/2, pp. 89-93, https://doi.org/10.1061/(asce)1052-3928(2007)133:2(89). [12]

Nye, B., A. Graesser and X. Hu (2014), "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring", *International Journal of Artificial Intelligence in Education*, Vol. 24/4, pp. 427-469, https://doi.org/10.1007/s40593-014-0029-5. [20]

Omi, M. and H. Winant (2018), "Racial Formation in the United States", in *Inequality in the 21st Century*, Routledge, https://doi.org/10.4324/9780429499821-49. [51]

Pardo, A., K. Bartimote, S. Buckingham Shum, S. Dawson, J. Gao, D. Gašević, S. Leichtweis, D. Liu, R. Martinez-Maldonado, N. Mirriahi, A. Moskal, J. Schulte, G. Siemens and L. Vigentini (2018), "OnTask: Delivering Data-Informed, Personalized Learning Support Actions", *Journal of Learning Analytics*, Vol. 5/3, https://doi.org/10.18608/jla.2018.53.15. [25]

Paris, D. (2012), "Culturally Sustaining Pedagogy", *Educational Researcher,* Vol. 41/3, pp. 93-97, https://doi.org/10.3102/0013189x12441244. [50]

Park, C. (2004), "The graduate teaching assistant (GTA): Lessons from North American experience", *Teaching in Higher Education,* Vol. 9/3, pp. 349-361, https://doi.org/10.1080/1356251042000216660. [3]

Polakova, P. and B. Klimova (2024), "Implementation of AI-driven technology into education – a pilot study on the use of chatbots in foreign language learning", *Cogent Education*, Vol. 11/1, https://doi.org/10.1080/2331186x.2024.2355385. [56]

Randall, J. (2023), "It Ain't Near 'Bout Fair: Re-Envisioning the Bias and Sensitivity Review Process from a Justice-Oriented Antiracist Perspective", *Educational Assessment*, Vol. 28/2, pp. 68-82, https://doi.org/10.1080/10627197.2023.2223924.   [62]

Sajja, R., Y. Sermet, M. Cikmaz, D. Cwiertny and I. Demir (2024), "Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education", *Information*, Vol. 15/10, p. 596, https://doi.org/10.3390/info15100596.   [14]

Shaffer, D., P. Nash and A. Ruis (2015), "Technology and the New Professionalization of Teaching", T*eachers College Record: The Voice of Scholarship in Education*, Vol. 117/12, pp. 1-30, https://doi.org/10.1177/016146811511701205.   [57]

Shah, M., M. Pankiewicz, R. Baker, J. Chi, Y. Xin, H. Shah and D. Fonseca (2024), "Students' Use of an LLM-Powered Virtual Teaching Assistant for Recommending Educational Applications of Games", in *Lecture Notes in Computer Science, Serious Games*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-74138-8_2.   [46]

Shermis, M. and J. Burstein (eds.) (2003), *Automated Essay Scoring*, Routledge, https://doi.org/10.4324/9781410606860.   [18]

Sinha, A. (2024), *BoilerTAI: A Platform for Enhancing Instruction Using Generative AI in Educational Forums*.   [45]

Stahl, M., L. Biermann and A. Wachsmuth (2024), *Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation*.   [35]

Teng, D., X. Wang, Y. Xia, Y. Zhang, L. Tang, Q. Chen, R. Zhang, S. Xie and W. Yu (2024), "Investigating the utilization and impact of large language model-based intelligent teaching assistants in flipped classrooms", *Education and Information Technologies*, Vol. 30/8, pp. 10777-10810, https://doi.org/10.1007/s10639-024-13264-z.   [36]

Tomanek, D. (ed.) (2012), "Does Instructor Type Matter? Undergraduate Student Perception of Graduate Teaching Assistants and Professors", *CBE - Life Sciences Education*, Vol. 11/2, pp. 187-199, https://doi.org/10.1187/cbe.11-10-0091.   [11]

Wadams, M. and K. Schick-Makaroff (2022), "Teaching assistant development and contributions in online, MOOC and blended synchronous settings: an integrative review", *Journal of Further and Higher Education*, Vol. 46/8, pp. 1023-1039, https://doi.org/10.1080/0309877x.2022.2038100.   [4]

Wise, A. and Y. Jung (2019), "Teaching with Analytics: Towards a Situated Model of Instructional Decision-Making", *Journal of Learning Analytics*, Vol. 6/2, https://doi.org/10.18608/jla.2019.62.4.   [26]

Yadav, G. and Y. Tseng (2023), *Contextualizing Problems to Student Interests at Scale in Intelligent Tutoring System Using Large Language Models.*   [32]

Yetişensoy, O. and H. Karaduman (2024), "The effect of AI-powered chatbots in social studies education", *Education and Information Technologies*, Vol. 29/13, pp. 17035-17069, https://doi.org/10.1007/s10639-024-12485-6.   [15]

Generative AI tools to support teachers:
A conversation with Dorottya Demszky

This chapter is an interview between Dorottya (Dora) Demszky, Assistant Professor in Education Data Science at Stanford University (United States) and the OECD Secretariat. The conversation discusses research about the emerging evidence about the potential of generative AI tools to support some teacher tasks: lesson planning, professional development based on their actual teaching, real-time support for tutoring, and the provision of feedback to their pupils and students. It concludes with a reflection on the availability of these tools for the teaching profession across the globe.

OECD: What do you think generative AI offers to teachers to support their teaching and the learning of their students, especially when the tools are teacher-facing?

Dora Demszky: My lab, the EduNLP lab[1], primarily focuses on this question: how AI tools, including GenAI, can support teachers in different ways, and of course there is a broader landscape of tools in this area. There are at least 4 areas where GenAI can support: lesson planning, professional development based on their actual teaching, real-time support for tutoring, and the provision of feedback to their pupils and students.

OECD: Great. Let's take those in turn and start with lesson planning and the development of curriculum materials.

Students who are creative in dance or music may not be in science, and vice versa. They must have knowledge and experience in a domain to produce something new and appropriate.

Lesson planning and curriculum material development

Dora Demszky: A main challenge for teachers is the time-consuming and difficult process of designing high-quality lesson plans for students with various needs. Curriculum varies greatly in the United States and this is also true in some other countries. Even when those do not vary, teachers often need to adapt teaching materials to meet students where they are, whether they are below grade level, multilingual newcomers needing language support, or students with special needs requiring visual or other types of tools. Teachers are not necessarily trained for this task.

One major area of work, both in industry and research, is addressing the challenge of curriculum adaptation. There are many possibilities, though some approaches are better than others. It's crucial to consider various

factors, such as maintaining rigour and preserving core components of carefully designed expert curricula, rather than just simplifying content. Our project, ScaffGen, researches how GenAI can support teachers with curriculum adaptation, considering high-quality instructional materials and teacher-specific contexts like students being below the expected proficiency at their grade level. This involves helping teachers adapt and create scaffolds for students that remain aligned with their curriculum.

Specific areas include creating more practice tasks and generating visual aids, like different ways to represent the same problem. We focus on multimodal generation, which GenAI excels at, and currently use LaTeX for diagram generation. We have evaluated scaffolds generated by Large Language Models (LLMs) for high-quality instructional materials against expert-created ones. We found that LLMs are similar and sometimes even preferred by teachers over expert-made ones, showing promise. There are still gaps, especially in visual aid generation. Another upcoming paper is a benchmark with a dataset of thousands of diagrams and LaTeX code from the Illustrative Mathematics curriculum, a leading K-12 math curriculum in the United States. We are releasing this dataset and benchmark studies to understand AI's performance in this area.

OECD: What do we know about the efficacy of AI-generated lesson plans and of your diagram-generation tool?

Dora Demszky: One of my former students built CoTeach.AI, an AI-powered curriculum adaptation tool grounded in the Illustrative Mathematics curriculum. After rolling it out for just a week in a small pilot, the tool has gained significant traction with thousands of regular users. We estimate about 10% of all teachers who use Illustrative Mathematics now use CoTeach. AI, which is substantial. Regarding efficacy, we are currently studying it and planning a pilot focused on our diagram-generation tool. We will test the quality of lesson plans from teachers using it versus those who don't, specifically focusing on the idea of multiple representations. We want to see if the tool's ability to generate diagrams supports students' understanding of connections between different representations (e.g., visualising abstract fractions). The curriculum provides limited representations, and we believe our tool can significantly support teachers in this.

More generally speaking, I haven't seen any efficacy studies for broader lesson planning tools like Magic School or School.ai. Much of it is self-reported usage or

perception. Evaluating efficacy is challenging because it requires rigorous metrics for lesson plan quality and, ideally, measuring student outcomes. Gathering student outcome data is slow, expensive, and logistically difficult, often falling to researchers due to lack of incentives in the EdTech industry. We are working on it, but it's a slow process.

OECD: I don't know any studies on the efficacy of lesson plans on student learning either. Some studies evaluate the generated lesson plan quality through human judgment and the time saved, focusing on productivity rather than whether the lesson led to better instruction quality. It seems your ScaffGen is more granular than full lesson plans.

Dora Demszky: CoTeach can generate full lesson plans, but often it generates activities. My lab, as part of the ScaffGen project, focuses on core R&D that many industry providers lack bandwidth for, such as diagram generation, which requires careful engineering, evaluations, benchmarks, and infrastructure. Many existing tools are essentially LLM wrappers, that is, software layers, or interfaces, built around an LLM): they don't have the capacity to build these challenging but necessary features. We are focused on fundamental technologies and evaluation, though the latter is complex and requires partnerships. We are working on a rubric for lesson plan quality for efficacy studies. It's also ethically challenging to withhold such tools from teachers for a control group. We are interested in gathering evidence despite these open questions.

OECD: You mentioned teachers sometimes preferred LLM feedback over experts' feedback: can you elaborate on that?

Dora Demszky: In a project in 2023, using earlier LLMs than the current models, we evaluated the quality of lesson plans based on predefined dimensions like readiness for classroom use, alignment with lesson objectives, preference, and alignment with student needs. Teachers compared the original curriculum warm-up from Illustrative Mathematics to two different LLM-generated and expert-generated lesson plans. What LLMs and experts produced were much more preferred across all criteria over the original material by a huge margin. On some dimensions, LLMs even outperformed experts. This is promising but needs careful interpretation.

Classroom analytics

OECD: A second application you mentioned belongs to the category of "classroom analytics" applications supporting teacher professional development or real-time classroom orchestration. I have always found this use of AI fascinating and promising. What does GenAI bring to these AI tools?

Dora Demszky: GenAI can support teachers in using pedagogically sound "talk moves" and discourse practices that probe student thinking, instead of just guiding them to a pre-specified solution or drilling. This involves dialogic practices that encourage student expression. GenAI can help analyse classroom discourse and student interactions. This can be done post-session: after a physical or online lesson, a transcript is analysed, and GenAI (or simpler AI models) can provide explicit suggestions on how to improve instructional practice or what talk moves to try next to support active learning.

We have conducted over 4 randomised controlled trials (RCTs) testing how this automated post-session feedback supports instructional improvement. We have a tool called Empowering Teachers. Teachers teach classes, and then they receive a report or feedback focusing on different talk moves, for example, inviting student thinking. The report includes counts and talk time. ChatGPT suggestions are also included in the paper. These talk moves are detected by language model-based classifiers, not GenAI itself.

We found that teachers who received this automated feedback from classifiers used the targeted talk move (e.g., focusing questions, building on or eliciting student ideas) by up to 20% more after only two feedback sessions, compared to a control group who did not receive such feedback. A limitation is the lack of rigorous assessment of student learning outcomes, but we do have access to student engagement metrics like talking more, showing up to classes, and completing assignments. We found students whose teachers received this feedback were more likely to submit assignments and show up to class.

There is room for improvement, but that's promising. GenAI is good at summarising conversations but struggles to accurately identify high-leverage teaching practices, as it requires significant context and understanding of classrooms. Even with careful prompting, it sometimes hallucinates or misclassifies classroom interactions. We see a lot of potential in this area though, especially for novices like volunteer tutors or new teachers who receive limited training. This will offer them professional learning.

I see less industry activity in talk move suggestions, perhaps due to lower profit. More common are fully automated tutoring systems like Khanmigo, though their effectiveness still needs evidence. Our lab focuses on supporting human tutors and teachers so we develop and research these types of teacher-facing tools.

OECD: How much do teachers like these tools? Adoption is usually one of the issues with them.

Dora Demszky: One practical challenge is that some teachers find it hard to act on this feedback. It requires reflection and thus time. While it raises awareness, deep change is often better supported by a human coach. We just published a working paper where instructional coaches helped teachers interpret this feedback, which was very helpful. Coaches were supported in pulling out specific evidence, and teachers felt less judged by the coach because they were looking at a third-party piece of evidence together.

Real time support

OECD: One of your very interesting studies is about providing support to human tutors in real time. Could you tell us about it?

Dora Demszky: Yes, in the real-time suggestions space we have the Tutor Copilot project, a collaboration with SCALE at Stanford. This project partnered with a tutoring provider supporting low-income students in text-based, in-school tutoring. Tutor Copilot allows tutors to activate the tool during the online tutoring sessions when students make math errors and need remediation. It suggests different response strategies and actual editable responses, giving tutors agency while also serving an educative purpose. A randomised controlled trial showed that tutors who had access to Tutor Copilot used better instructional practices, and their students mastered lessons faster. This was particularly helpful for tutors with lower initial quality ratings or less experience (see Figure 10.1).

Figure 10.1. Tutor Copilot: a way to mobilise less qualified tutors effectively, 2024

Results from a randomised controlled trial

Notes: These two figures show the results of using Tutor CoPilot on student learning (upper panel) and on tutor pedagogies (lower panel). The effect on student learning varies based on tutors' initial effectiveness, measured by their quality rating. The results indicate substantial benefits for tutors with lower initial effectiveness. Lower-rated tutors experienced a 9-percentage point increase in students passing their exit ticket (56% to 65% student passing rate from control to treatment). Similar effects were observed with less experienced tutors.

The lower figure shows that some teaching strategies were likely to be used by control tutors (left) vs. treatment tutors (right). Strategies with a z-score below 1 standard deviation are shaded in grey. Control tutors tended to rely on solution-focused, passive strategies, while treatment tutors more frequently used strategies that promote deeper student engagement and comprehension."

Source: Wang et al. (2025[1]), Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise, Retrieved from Annenberg Institute at Brown University, https://doi.org/10.26300/81nh-8262.

OECD: What do you think about the appropriateness of choosing generic LLMs versus more educationally-focused ones? How can you be sure that the use of GenAI is educationally appropriate?

Dora Demszky: That's a very big question, and it's one of the central questions we ask teachers. How do teachers determine appropriateness? Specific models (GPT, Cloud, Gemini) are ever-changing, and we haven't found massive differences. The best model might change next week… The criteria for educational appropriateness vary a lot by context, teacher, and project. It is important to learn about these criteria. We hosted a Practitioner Voices summit for math educators at Stanford this summer, where one of our main goals was to learn about their criteria for evaluating AI tools. Our short report can be found online and we will be releasing a longer paper soon.[2]

OECD: In the case of Tutor Copilot, the system was trained by providing data based on the observation of and work with expert teachers. Do you think GPT-4 at the time would have given similar measurements without this dedicated educational element?

Dora Demszky: No, we explicitly needed the expert teachers' input to improve the model. Without that "expert-informed cognitive task analysis" – where we tell the model how an expert teacher would remediate a student's mistake – it performed significantly worse. We are doing something similar with ScaffGen, giving the model these expert-informed processes. This is related to, but slightly different from, the evaluation criteria for determining if a tool is good, though the two can inform each other.

OECD: Could these tools supplement and augment teachers to improve educational quality, especially in countries or contexts with teacher shortages or teachers with a lack of expertise?

Dora Demszky: I want to problematise the premise that we don't have human teachers available. It's risky to accept that technology should (or could) replace the human teacher role, as this could worsen inequities in access to human teachers, not just in low-income countries but within the United States and OECD countries too. If situations genuinely lack a human teacher, we must think carefully about what roles these

tools can fulfil. The relationship-building part cannot be replaced by technology, though other aspects could potentially be, which remains to be tested.

OECD: I was not thinking of replacing teachers, but more, like, if you have inexperienced or low-quality teachers or people with little subject and pedagogical knowledge, could these tools help them improve their performance? In many countries, there would be just too many teachers to train, so being able to enrol the next person to tutor or teach could help. If the humans don't really know yet what they're doing, could tools like the Tutor Copilot help?

Dora Demszky: Our central focus for these technologies isn't time-saving, but rather the educative element – supporting teachers' professional learning. All teachers have room to grow. Different versions of the tools could be tailored to the user's experience level; for example, a novice teacher might be overwhelmed with too many decisions or information before they gain more training. We have tested some tools with complete novices. In the "Code in Place" global programming course run by Stanford University, we implemented the teacher feedback tool with thousands of volunteer section leaders, most of whom had zero teaching experience. This feedback tool helped them, so this is a significant user base we are targeting. But we would need to do pre-work to ensure these technologies translate to different languages and local needs if we were to use them in the contexts you mentioned.

OECD: What do you think GenAI can never do as well as a human being, if anything, especially regarding the human dimension in education?

Dora Demszky: Motivation and relationship building are key elements that GenAI may never do as well as humans. While more research is needed, experts in education agree, and it intuitively makes sense. An AI won't be seen as a role model. Students might share things with AI they wouldn't with a human because they're less afraid of vulnerability. However, a human is better able to support emotional well-being and create accountability. With AI, there's no accountability. A student might not care what they do because the AI won't get hurt. Social-emotional skills for example are learned better with a human teacher and human peers. Learning involves much more than just knowledge or information gathering.

Feedback on student work

OECD: Let us move to the final area that you mentioned initially: feedback to students. We know this is essential for learning, for teachers as we have already mentioned, but also for students. What could GenAI offer on that?

Dora Demszky: A significant area of research and development for both industry and academia is teachers giving feedback to students on their work. Teachers often lack time (imagine they have 150 students) but also training to give high-quality feedback. As we aim to go beyond just productivity and time-saving, focusing on improving the quality of teaching and of feedback is essential. Some tools exist, like Brisk, supporting GenAI-driven feedback, especially in writing. We are working on rigorously validated tools that also support professional learning around feedback provision.

OECD: Are you talking about formative assessment, where feedback is given on students' written assignments? Or is it linked to the applications of real-time feedback and dialogic practice you told us about? How do these two link, and what do we know about the efficacy of tools being developed?

Dora Demszky: Our work focuses on formative feedback that teachers can give on student assignments, but with a strong emphasis on revision. One goal of feedback is to help students improve and revise their work, and students are less likely to read feedback if they don't have a chance to revise. We focus on areas with room for student improvement and lower stakes. The

efficacy of these new feedback tools remains to be seen, as they are very new, but conceptually the design seems sound.

Teachers often accept GenAI suggestions without editing. That's a problem. We explicitly design our tools to support teachers in creating feedback, not to replace their feedback, because research indicates students are less likely to act on feedback perceived as coming from AI rather than their teacher. It's crucial for students to feel the feedback is from their teacher. We are developing a benchmark for feedback quality, a set of measures for assessing feedback from teachers or AI tools, which we hope industry will adopt.

We have a working paper that compares expert-written feedback to LLM-generated feedback. While LLMs are not bad, they significantly lag behind experts in key areas. For example, LLMs are much less dialogic, tending to give specific rewrite suggestions ("this was not right, here's how to rephrase") rather than engaging with holistic arguments or probing student thinking to encourage revision. Also, LLM comments can be disjoint, unlike a teacher's coherent feedback where comments build on one another. We are actively working on improving and evaluating the use of GenAI tools.

The two strands of projects – teacher-facing feedback on talk (which is maths/STEM-focused) and student-facing feedback on writing – are not directly linked currently. However, we envision integrating them. For example, teachers could receive a post-lesson report summarising student assignments and class discussions, offering feedback suggestions for assignments, and guiding future lesson planning. This could be a complementary system down the line.

Real-life implementation

OECD: So, are all the tools you've mentioned useable in real life in-person classroom or instruction settings, except Tutor Copilot, which is for virtual platforms? For example, can you think of uses of Tutor Copilot in an in-person setting?

Dora Demszky: We need to be careful not to make real-time suggestion tools distracting in virtual or physical face-to-face contexts, or to take away educator agency. One idea we're exploring is surfacing feedback during high-leverage moments in a real-time classroom, like when students are working on problems and there's a pause, rather than giving suggestions constantly. Identifying these non-distracting periods could be very useful.

Doing this in a virtual context is straightforward; you can speculate when these moments might occur and surface feedback. In a physical classroom, it's harder due to challenges in accurately capturing student voices, surfacing real-time feedback to teachers (e.g., via iPad), and instrumentation. We need to talk to teachers about this. One question for participants at our Practitioner Voices summit was how these tools could support teachers in a physical classroom, whether by analysing group work or teacher discourse. They might help us envision practical implementation. There might be variation, with some teachers preferring post-teaching feedback and others appreciating real-time tools. Our longer report will report what we learned from teachers.

Notes

1. https://edunlp.stanford.edu/

2. https://hai.stanford.edu/news/how-math-teachers-are-making-decisions-about-using-ai

References for further reading

Demszky, D., H.C. Hill, E.S. Taylor, A. Kupor, D. V. Dennison and C. Piech (2025), "Does Increased Agency Improve the Effectiveness of Self-Directed Professional Learning for Educators?", *EdWorkingPaper No. 25-1162*, Annenberg Institute for School Reform at Brown University, https://doi.org/10.26300/04kc-7085. [7]

Demszky, D., J. Liu, H.C. Hill, D. Jurafsky and C. Piech (2024), "Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course", E*ducational Evaluation and Policy Analysis*, Vol. 46(3), pp. 483-505, https://doi.org/10.3102/01623737231169270. [8]

Demszky, D. and J. Liu (2023), *M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1: 1 Instruction and Student Outcomes*, https://doi.org/10.26300/s8xh-zp45. [9]

Demszky, D., J. Liu, H.C. Hill, S. Sanghi and A. Chung (2025), "Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement", *Computers & Education*, Vol. 227, https://doi.org/10.1016/j.compedu.2024.105183. [4]

Demszky, D., R. Wang, S. Geraghty, and C. Yu (2024), "Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform", *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 632-644, https://doi.org/10.1145/3636555.3636924. [6]

Handa, K., M. Clapper, J. Boyle, R. Wang, D. Yang, D.S. Yeager, and D. Demszky (2023), "Mistakes Help Us Grow": *Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms,* https://doi.org/10.18653/v1/2023.emnlp-main.549. [10]

Mah, C., M. Tan, L. Phalen, A. Sparks, and D. Demszky (2025), "From Sentence-Corrections to Deeper Dialogue: Qualitative Insights from LLM and Teacher Feedback on Student Writing", *EdWorkingPaper: 25-1193*, Retrieved from Annenberg Institute at Brown University, https://doi.org/10.26300/p397-2p46. [5]

Malamut, J., D. Demszky, C. Bywater, M. Reinhart, H.C. Hill (2025), *Facilitating Evidence-Based Instructional Coaching With Automated Feedback on Teacher Discourse,* Retrieved from Annenberg Institute at Brown University, https://doi.org/10.26300/xx9z-8f27. [14]

Malik, R., D. Abdi, R. Wang, and D. Demszky (2025), "Scaffolding middle school mathematics curricula with large language models", *British Journal of Educational Technology*, Vol. 56/3, pp. 999-1027, https://doi.org/10.1111/bjet.13571. [2]

Malik, R., D. Abdi, R. Wang, and D. Demszky (2024), "Scaling High-Leverage Curriculum Scaffolding in Middle-School Mathematics", *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pp. 476-480, https://doi.org/10.1145/3657604.3664698. [3]

Malik, R., R. L. Hao, R. Kacholia, and D. Demszky (2025), "MathemaTikZ: A Dataset and Benchmark for Mathematical Diagram Generation", P*roceedings of the Twelfth ACM Conference on Learning @ Scale*, pp. 95-104, https://doi.org/10.1145/3698205.3729558. [13]

Wang, R. and D. Demszky (2023), *Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance for Scoring and Providing Actionable Insights on Classroom Instruction*, https://doi.org/10.18653/v1/2023.bea-1.53. [11]

Wang, R., A.T. Ribeiro, C.D. Robinson, S. Loeb, and D. Demszky (2025), *Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise*, Retrieved from Annenberg Institute at Brown University, https://doi.org/10.26300/81nh-8262. [1]

Wang, R., Q. Zhang, C. Robinson, S. Loeb, and D. Demszky (2023), *Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes*, https://doi.org/10.48550/arXiv.2310.10648. [12]

Part 3

Improving system and institutional management

11 AI in institutional workflows:

Learning from higher education to unlock new affordances for education systems and institutions

Zachary A. Pardos*, Conrad Borchers**
*University of California, Berkeley, United States
**Carnegie Mellon University, United States

This chapter provides an overview of how artificial intelligence is being leveraged at the institutional level to support core academic processes in higher education, including course articulation, credit transfer, academic advising, admissions, and curricular infrastructure. It explores the potential of AI, ranging from embeddings and classifiers to predictive and generative models, to enhance decision-making, streamline complex workflows, and surface novel opportunities to strengthen institutional practices. The chapter emphasises the macro-level affordances of AI systems, particularly those designed for administrative use, and highlights the importance of iterative collaborative development across research and operational units for successful implementation. Finally, it offers considerations for policymakers and system leaders seeking to responsibly integrate AI into the fabric of higher education institutions as well as in lifelong learning and school systems, where these techniques are also applicable.

Introduction

This chapter examines the emerging role of generative AI (GenAI), some of the techniques on which it builds and its AI predecessors, in back-end functions of higher education, including course articulation, student transfer, advising, admissions, and content infrastructure. Unlike instructional uses of AI, which often focus on the learner as the end user, the systems discussed here are typically administrator- or staff-facing or are embedded into educational platforms installed at an institutional level to produce insights, reduce task complexity, and support academic pathway navigation (Kizilcec et al., 2023[1]). In most cases, the AI models behind the tools utilise data collected at the macro- or meso-level as opposed to the micro-level (e.g. learning process data: Fischer et al., 2020[2]).

Drawing on recent research, case studies, and early-stage prototypes, this chapter identifies how AI can:

- Support credit mobility and transfer prediction across institutional boundaries
- Support academic advising, such as with personalised course and major recommendations and curricular analytics
- Diagnose novel opportunities to enhance admissions and resource allocation, and
- Structure the classification, tagging, and reuse of learning content and curricular components

While not all of these tools involve generative AI as means to create content or directly interface with end-users, many depend on machine learning, natural language processing, and representation learning (e.g. embeddings), technology at the heart of generative AI, to support institutional decision-making. The chapter foregrounds macro-level, institutional infrastructure as the critical site of innovation for unlocking more evidence-based, personalised, data-informed, and ultimately student-serving higher education ecosystems. While the chapter focuses on research carried out at the higher education level, many of the covered possibilities are relevant to the secondary school sector, at the system rather than institutional level, as well as to support lifelong learning.

Emerging opportunities

Easing learning mobility through AI equivalency models

As students traverse academic pathways, their ability to have learning acknowledged when moving between segments or systems of education can be the difference in their ultimate academic success. In the United States, when students move from a 2-year community college to a 4-year university, for example, agreements called course articulations dictate how much credit will come with them and which requirements it will satisfy. Similarly, prior learning gained in industry from a professional certificate that is then attempted to be counted as equivalent to institutional course credit is referred to as *Credit for Prior Learning* (CPL) (Lakin, Seymour and Crandall, 2015[3]). In other countries the same issue can arise when individuals want to change study paths, to transition from a 2-year study programme to a bachelor's degree, change higher education institution outside traditional study paths, etc. This can also happen for the international recognition of foreign degrees in the frame of international student mobility - or just professional mobility. Demonstrating mastery of a skill from one taxonomy and then seeking acknowledgement of mastery in a similar skill from another taxonomy requires mapping, or cross-walk between taxonomies (Li et al., 2021[4]).

These variations on credit and learning acknowledgement scenarios are critical to student success in higher education; however, they have historically been constructed and maintained by hand, often with missing or inequitably distributed pathways for mobility that favour credit from institutions with higher socioeconomic standing (Pizarro Milian and Aurini, 2025[5]; Goulart and Pardos, 2025[6]). Generative AI and the natural language processing technology behind it could and is increasingly being used to address these deficits in ways that have the potential for equitable scaling.

Representing student pathways with AI

One promising direction to better map different types and levels of educational programmes is to identify closeness across courses with AI techniques. This involves representing course content as AI vector embeddings (see Box 11.1), enabling semantic similarity comparisons across thousands of courses. This representation can be informed both by natural language signals, such as a title and course catalogue description, but also by historic enrolment data, that is, the actual choices courses that individual students made within their higher education programmes. Using the latter, Pardos and Nam (2020[7]) visualised the semantic topology of courses offered at a large public university (Figure 11.1) and queried the underlying course vector representation to reveal differences between courses. For example, when asked what the difference was between the Econometrics and Advanced Econometrics courses, the model correctly responded with "Linear Algebra." It perhaps would not be surprising for contemporary LLMs to be able to answer this question, given their access to troves of data; however, the model from this example used only course enrolment histories, showing the effectiveness of these models with a better defined amount of data. This effectiveness of course vectors was later demonstrated in successfully predicting student course workload perceptions (Borchers and Pardos, 2023[8]) and also performed strongly on prerequisite prediction (Recall@10 = 0.70) and average-enrolment prediction (RMSE = 42.48) as described in Jiang and Pardos (2020[9]).

Box 11.1. AI vector embeddings

An embedding can be thought of as the organisational structure that AI uses to map out what it knows about a domain. Neural networks, the type of AI behind ChatGPT and self-driving cars, "think" in terms of vectors in this embedding. They take an input (e.g. an image) and convert it into a vector that can then be manipulated through a series of matrix operations (e.g. an image being scaled or rotated as a result of multiplications). Modern AI is usually trained to predict the next word in a sentence, or more generally the next item in a sequence, or to fill in the blanks in past sequences of data. For AI to do this efficiently, it has to pick up on patterns and similarities that allow it to generalise what it has

observed in the past (i.e. training data) to new scenarios. It does this by organising the embedding "space" in ways that recognise these patterns, for example by placing synonymous words close together in the space or by giving pairs of words with the same relationship to one another (e.g. Paris -> France, Seoul -> South Korea) similar relative distances to one another in the vector space (e.g. a capitol's vector in the space might be approximately two units down and one to the left of its country vector in the space). This spatial representation is how neural networks encode what they have learned (i.e. generalised) from the data. When they overgeneralise, this can manifest in hallucinations (e.g. when the expected capitol vector is not at the expected offsets from the country and the AI subsequently gives the wrong answer).

A vector space is commonly hundreds to thousands of dimensions in length. As of this writing, OpenAI offers 1 536 length vectors for its small model and 3 072 for its large. Utilising vectors (i.e. generating and downloading them) in higher education workflows, instead of querying an LLM, is a way of utilising the power of modern AI without having to rely on real-time querying and incurring ongoing API costs. This is a practical approach when a natural language interaction is not the goal. Free vector embeddings are available from the extended research community (e.g. http://www.sbert.net).

Figure 11.1. Projections of courses at a large public university
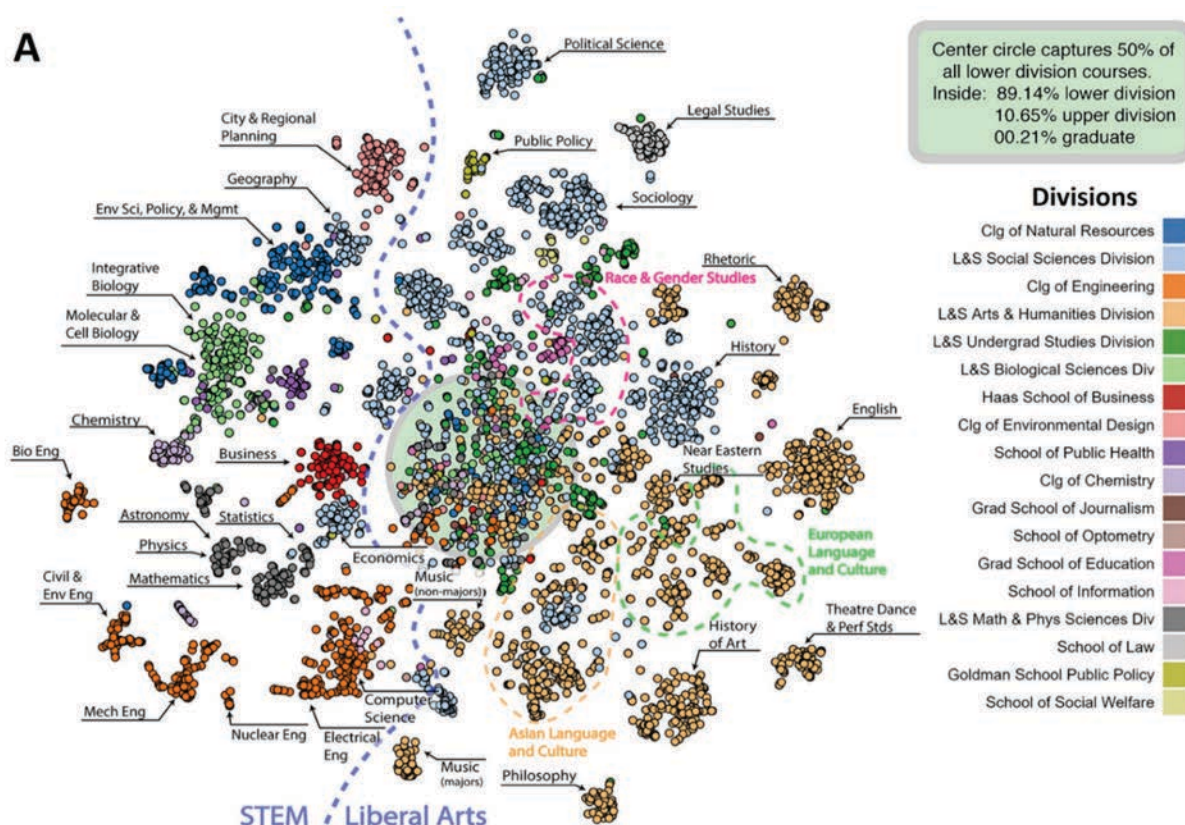
Note: The projections are learned from a neural network representation model (i.e. course vector embedding) based only on historic course enrolment sequences.

Source: Pardos and Nam, (2020[7]), "A university map of course knowledge", *PLOS ONE*, Vol. 15/9, https://doi.org/10.1371/journal.pone.0233207.

Historic enrolment data within institutions can also be leveraged by AI to learn and provide course recommendation pathways using the same type of neural networks which power generative AI (Shao, Guo and Pardos, 2021[10]). Much like a Generative model could complete your sentence, this similar model applied to course enrolments can complete a student's course sequence to include necessary requirements and nurturing budding personal interests to satisfy electives.

The projections were produced by reducing course vectors to 2-D using t-SNE. The space may suggest to a Dean or other administrator, where a department may have a concentration of topical strength and in what areas "neighbouring" departments may be collaborated with to fill in gaps of a major or work together on a major. If another institution's course vectors were to be overlaid on to this one, it can suggest where the institutions complement one another, where they are aligned, and where expected alignment could be improved.

AI-assisted programme alignment between institutions

Beyond single institution course planning and recommendation, AI can also support the development and maintenance of cross-institutional course equivalency models - an important enabler for student transfer between degree programs in higher education and a facilitator for lifelong learning and the recognition of prior learning. For example, in the United States, starting at local, more affordable community college and then transferring to a bachelor's programme has been the greatest source of upward social and economic mobility (Johnson, 2020[11]). Similarly, in the European Union, the European Credit Transfer and Accumulation System (ECTS) is designed to promote student mobility by standardising how learning achievements are measured and recognised across institutions (Grosges and Barchiesi, 2007[12]). In practice, however, the reality of exchange programs and institutional transfers often involves negotiations for the accreditation of specific courses toward degree requirements. In both cases, course equivalency agreements between institutions are required to allow for transfer to work as intended (Ignash and Townsend, 2000[13]). Across many higher education systems, articulation and credit transfer remain time-consuming, manual processes. Faculty or articulation officers typically review syllabi and catalogue information to determine equivalency across institutions. AI, particularly natural language models and course embeddings, has begun to offer data-driven alternatives and support structures.

The same embedding models that underpin generative AI-used to represent meaning – rich relationships between words, images, and multimodal content – can also represent relationships between courses at different institutions. Pardos, Chau and Zhao (2019[14]) demonstrate that machine translation techniques, built on these embeddings, can "translate" between the course vector spaces of different colleges. These vectors, learned from students' historical course or programme-level enrolment patterns and course catalogue descriptions, capture latent curricular structures, enabling the prediction of equivalences and surfacing gaps in transfer agreements. These gaps correspond to equivalences that could have been offered given course contents and pathways but that had not been identified or considered yet. In their proof-of-concept, the approach successfully matched courses between a two-year and a four-year US institution and validated 65 pre-established articulations. This methodology is being piloted with 59 US higher education institutions and four systems of higher education to explore its feasibility and utility in practice[1].

Methodologically, the educational data mining community has explored additional neural course representations outside the transfer context. Khan and Polyzou (2024[15]) evaluated session-based methods such as CourseBEACON and CourseDREAM (neural architectures that recommend well-suited course bundles based on enrolment sessions) and showed improved performance of these methods over traditional factorisation or association models. These session-based models recommend full next-semester sets of courses by modelling (1) which courses pair well together and (2) semester-to-semester orderings using RNN/LSTM encoders (CourseBEACON uses an explicit co-occurrence matrix; CourseDREAM learns latent basket vectors). They improve accuracy over popularity and sequential baselines (CourseDREAM achieves the best Recall@k on test of about 0.30). Similarly, Kim et al. (2025[16]) demonstrate that deep embeddings of course descriptions, coupled with traditional classifiers, can automate equivalency judgments with near-perfect performance (as measured by F1 scores). Both works demonstrate how representations of courses can be used for various institutional tasks guiding student pathways and transfer mobility.

While AI-assistive equivalency models can dramatically speed up articulation, adoption depends on trust-especially among domain experts that ultimately hold the keys to unlock credit approvals. Xu et al. (2023[17]) studied algorithm aversion in higher education administrators tasked with course credit decisions. Using a 2×2 experiment with an AI-based matching platform. One factor was whether low-confidence or outlier AI recommendations were inserted into the results or not. The other factor was whether the interface prompted users to flag inappropriate AI recommendations or not. They found that not including outlier recommendations improved acceptance and productivity; however, asking users to flag recommendations reduced administrators' acceptance of the suggestions unless outlier recommendations were turned on. While the literature suggested user flagging

as a means to increase adoption, it may have led to a negative mindset in this case, unless users were given clear results worth flagging. The takeaway is perhaps that without the right implementation, even accurate AI recommendations may risk being undervalued or outright rejected in a socio-technical system.

These findings show that embedding-based methods, central to modern generative AI, are not only useful for producing language or images, but that they can also map complex academic structures across institutions. When combined with careful human-AI collaboration design, they can accelerate equivalency mapping, reveal hidden curricular alignments, and reduce administrative burden while keeping human experts in control.

Tagging educational content for reuse and curriculum alignment using AI

While many AI applications focus on student-facing outcomes, others operate behind the scenes to power the discoverability, classification, and reuse of educational content. A growing area of impact is the back-end organisation of large-scale resource libraries through tagging, aligning, and curating learning materials to match institutional or state-wide taxonomies or support transition and alignment between them. The significance of using AI for annotating and grouping educational content is that content standards are often changing over time (e.g. the U.S. Common Core or the Finnish National Core Curriculum for Basic Education). Recategorising content and aligning it with new standards is expensive, but could be made substantially more cost-effective and efficient using AI. For example, this can work for Open Educational Resources (OER). While many countries and international organisations like the OECD and UNESCO have supported the development of Open Educational Resources, the COVID-19 pandemic highlighted that identifying those aligned with country or jurisdictions' curricula was not trivial, because of a lack of domestic or international taxonomy (e.g. Vincent-Lancrin, Cobo Romaní and Reimers, 2022[18]). The strength of generative AI techniques at classifying and mapping text (with embeddings) may help to solve or at least mitigate that problem.

Figure 11.2. Embedding-based models for mapping problems, skills, and curricula
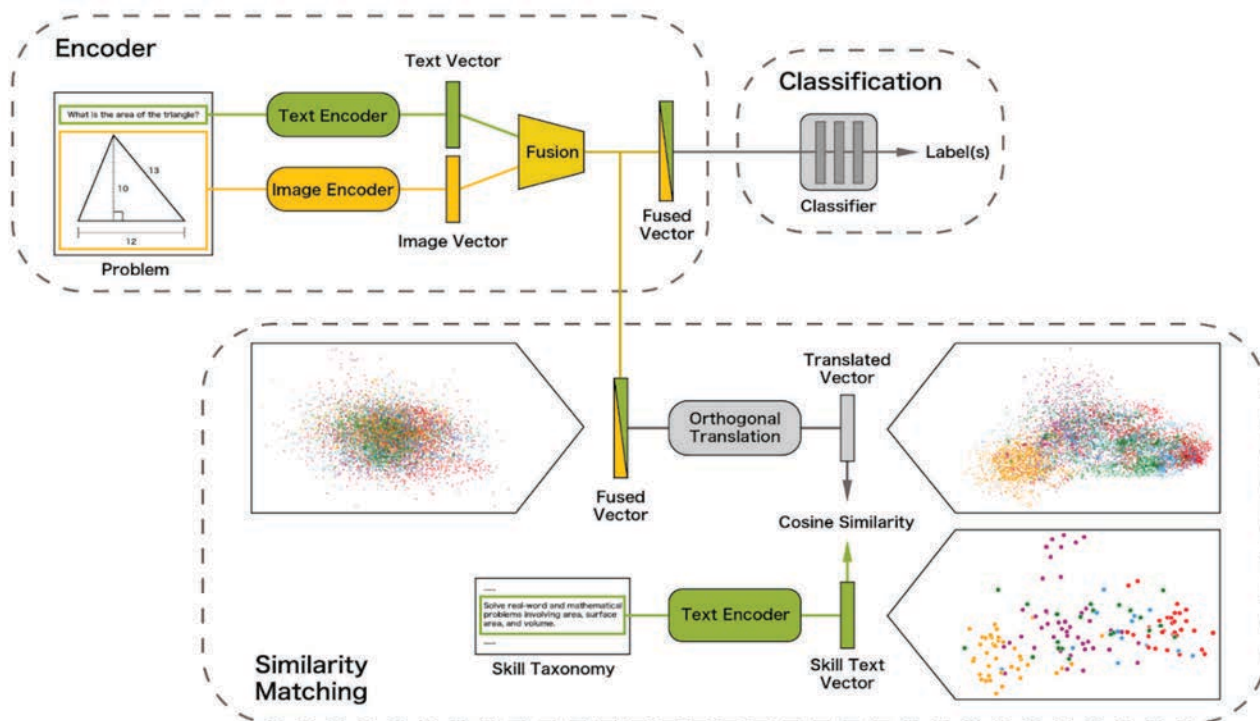
Note: Embedding spaces are shown as point clouds of problems or skills coloured by labels. Labels for a curriculum could be Pythagorean Theorem. The similarity matching phase provides information about different taxonomies and reconciles them based on their similarity of the skills descriptions with the problem text and information from any available figures.

Source: Adapted from Liu, Bhandari and Pardos (2025[19]), "Leveraging LLM respondents for item evaluation: A psychometric analysis", *British Journal of Educational Technology*, Vol. 56/3, pp. 1028-1052, https://doi.org/10.1111/bjet.13570

Another example is the conversion of a mastery profile from the proprietary skill taxonomy of an intelligent tutoring system into the US state common core standards, a set of agreed educational standards in English language arts and mathematics across a large number of States in the United States. Intelligent tutoring systems supporting the acquisition of content and procedural knowledge in language arts, maths, science, etc., have their own "knowledge maps" that are often not specific to any particular curriculum or standards. Allowing these translations is essential to help teachers and educators using particular systems to align them with local curricula. These infrastructure-oriented uses of AI may be invisible to learners, but they are foundational for enabling efficient resource retrieval, supporting instructional planning, and ensuring alignment with evolving curricular goals (Figure 11.2).

Recent advances apply techniques central to generative AI, particularly embedding models, to create rich vector representations of learning resources that capture semantic relationships between content items. These embeddings support clustering, that is, the grouping of content according to common characteristics, similarity search, and cross-walking between taxonomies, while classification algorithms, often fine-tuned on top of embeddings, map resources to categories in both established frameworks and newly defined skill taxonomies. For example, when a new mathematical curriculum or taxonomy, like the US Common Core Standards, are introduced, these methods can aid in re-mapping the estimated millions of existing open educational resources to the new taxonomy. Such methods have been deployed to support initiatives like common course numbering, enhance tutoring systems' ability to link resources to specific knowledge components, and keep course catalogues aligned with rapidly changing programme requirements.

Research suggests that these AI-assisted systems can approach or even match non-expert human tagging performance with relatively small, labelled datasets, and in some cases rival expert performance at scale. For example, Li et al. (2024[20]) found that their approach combining embedding and classification could achieve non-expert accuracy with as few as 100 labelled examples, and near-expert accuracy with 5 000. Importantly, these models incorporated multimodal features from text, images, and videos-mirroring the multi-input capabilities of contemporary generative AI systems-and were publicly released for use with both the US Common Core and novel taxonomies.

Ren et al. (2024[21]) extended this line of work to study human-AI collaboration in taxonomy alignment. Compared to humans working alone, AI suggestions reduced tagging time by roughly 50% ($p \ll 0.01$) but led to a modest decline in recall, that is, the identification of all relevant resources of a specific category (- 7.7%, $p = 0.267$), and substantial decline in accuracy, that is, the overall correctness of tagging suggestions (-35%, $p = 0.117$). Notably, the AI-alone condition performed worst, while the human-alone condition performed best for accuracy-placing the collaborative condition in between. These findings highlight a trade-off between efficiency, here speed of tagging, and precision, and suggest that while AI can accelerate large-scale taxonomy updates, quality assurance remains essential – as of now, performed by humans.

As educational taxonomies continue to evolve, whether through new competency frameworks, new curricula or standards, or institutional redesigns, embedding-powered alignment tools offer a scalable way to re-tag resources, identify content gaps, and maintain interoperability across systems. In doing so, they extend the same representational methods powering generative AI into the critical, though less visible, infrastructure that underpins curriculum management.

Personalisation at institutional scale

Personalised and AI-augmented advising

Generative AI in higher education can involve helping personalise the guidance students receive, not just from a recommender system, but from human academic advisors. In Lekan and Pardos (2025[22]) an advisor-facing, GPT-driven model was tested whereby first year college students (n = 33) were asked questions about course preferences and career goals, typical of a human advising session. These responses were fed to a GPT model that, instead of giving advice directly to the students, gave major recommendations and justifications to an advisor (n = 25). The study found that academic advisors rated the suggestions of the GPT model favourably and exactly agreed with the model's major recommendation 33% of the time. In this case, participating advisors were positive on this type of human-AI collaboration, as providing assistance and leaving them as the point of contact to students, rather than supplanting them.

Research collaborations with registrars and admissions offices have begun to explore how advisors can use analytics to better support student course selection. For example, nascent work (Borchers, (n.d.)[23]) with undergraduate advising integrates Big Five personality traits, such as conscientiousness and neuroticism, with multi-semester course

enrolment data and found that students high in conscientiousness or self-efficacy tend to perform well even under heavy workloads, while those lower on these traits are more likely to struggle. This suggests that advisors could move beyond general heuristics (e.g. "don't overload") and instead offer more individualised recommendations based on a student's likely capacity to manage challenging course schedules.

The same form of advising could also be used in high school to help students choose their higher education study programmes, or later on in their life, to support their choices of lifelong learning options. While generative AI is also explored as means to provide or facilitate other advising and counselling services to college students (e.g. mental health), these applications are not without controversy and have shown growing pains in their current stage of development (Moore et al., 2025[24]).

Enhancing production and evaluation of assessments with GenAI

Assessment represents one of the most resource-intensive components of higher education. Generating high-quality items for standardised tests in some subjects requires significant faculty time, while evaluating and calibrating those items demands large respondent pools and psychometric expertise. Recent advances in generative AI offer institutions the opportunity to refresh assessment practices by accelerating both the production and the evaluation of large-scale item pools. Importantly, these processes often occur in the institutional "backend"-funded, managed, and maintained by campuses or system-level services-in addition to being driven by faculty or students directly. Developing assessment items is also relevant at the school level in countries with national assessments, whether they are developed by public evaluation agencies or by private companies.

LLMs offer novel perspectives on automating the creation of multiple-choice and short-answer items, particularly when anchored in existing curricular material. Studies comparing LLM-generated questions to textbook-sourced questions find comparable psychometric properties. For example, Bhandari et al. (2024[25]) report that ChatGPT-generated algebra items demonstrated difficulty and discrimination parameters statistically indistinguishable from traditional textbook items when evaluated with item response theory. Notably, the LLM-generated items exhibited slightly stronger differentiation between high- and low-ability respondents, suggesting that GenAI can produce assessment content of similar or even superior quality under controlled conditions. This holds particular promise for large lecture courses and general education programs where instructor time is scarce and item demand is high (e.g. test banks must be regularly refreshed to ensure continued assessment validity). To retain the instructor's agency over course assessments while decreasing the time it takes them to create an assessment is not just an algorithmic matter but a human-computer interaction design issue. New human-computer interaction research, such as work on the PromptHive tool, provide examples of placing subject matter experts in the driver's seat of generative AI to integrate their expertise into the workflow of assessment creation (Reza et al., 2025[26]). An instructor, for example, provides her existing assessments as a style reference as well as the new learning objectives she wants additional assessments to cover. PromptHive creates a pool of assessment items covering the learning objectives and allows the instructor and TAs to instruct PromptHive on the types of hints that should be produced to scaffold learning of the related content. The instructional team can then preview the generated hints and assessments on a subset of items or all items. The limitation here is that generative AI still hallucinates in most topic areas. If hallucination rates are not evaluated to be 0% in the topic area, this necessitates the instructional staff to check every problem and hint being produced before it is seen by a student.

Relatedly, generative models can also address known limitations in traditional test banks. One persistent challenge is the overexposure of items, in cases where repeated use narrows the effective variance of assessments and introduces unwanted correlations between items. For example, many high-stakes, summative tests are strictly proctored with test items that are not released to the public; however, over time, test takers may socialise the contents of the test to future test takers, or test prep companies, leading to overexposure of the items if they are not changed frequently. By creating novel assessments and even well-crafted multiple-choice distractors, large language models can diversify item pools, reducing the risk of distributional shifts when the same questions are repeatedly deployed. Yet producing new distractors at scale raises the question of quality control. Poorly written distractors-those that are implausible, misleading, or inadvertently cue the correct answer-can reduce both the fairness and the psychometric value of multiple-choice questions. Here, automated evaluation methods are beginning to complement generative approaches. Moore et al. (2023[27]) provide evidence that such methods can systematically detect flaws in student- and AI-generated multiple-choice questions. The authors evaluated undergraduate students in introductory courses who were prompted to generate multiple-choice questions on recently learned material. Comparing a rule-based system to GPT-4 on 200 student-generated questions across four domains, they found that the rule-based approach identified 91% of item-writing flaws flagged by human annotators, compared to 79% for GPT-4. Many of these flaws involve distractor design-such as

implausibility, redundancy, or overlap with the correct answer-precisely the weaknesses that can undermine the benefits of automated generation at scale. Hence, human expertise and quality control will remain important pillars as LLMs are used for content generation at larger scale.

Beyond production, GenAI is emerging as a tool for evaluation. Item calibration - the process of estimating psychometric properties such as difficulty and discrimination - typically requires thousands of student responses. Liu et al. (2025[19]) demonstrate that multi-agent AI models bringing together ensembles of LLMs can serve as "synthetic respondents," producing response distributions with psychometric properties closely aligned to those of college students. While a single LLM was not measured to exhibit abilities similar enough to the target human population, ensembles of different LLMs expand variance, yielding item parameter estimates highly correlated (> 0.8) with human-calibrated values. Augmentation strategies, such as adding LLM responses to even a small set of human respondent data further improves alignment with exclusively human responses. These findings suggest a new institutional workflow: LLM-based calibration can complement limited student response data, reducing costs and accelerating item validation cycles. While human responses remain essential for final benchmarking, AI-assisted evaluation can substantially shorten development timelines for new assessments.

Beyond opportunities and feasibility, a practical open challenge is instituting policies for the use of generative AI in assessment. Drawing from emerging frameworks proposed by Corbin et al. (2025[28]), open issues include:

- where to set meaningful limits on AI assistance for different outcome types;
- what disclosure, attribution, and provenance practices are sufficient (e.g. prompts, drafts, and model/version logs);
- how to handle discipline-specific variation without sacrificing consistency; and
- how to mitigate workload burdens for staff while maintaining validity and fairness.

A challenge in these policies is that AI models and their capabilities will continue to change. Therefore, policies will need to be adaptive rather than static and focused on guiding principles and review mechanisms rather than fixed prohibitions. These considerations are especially important as AI is deployed in assessing high-stakes scenarios for learners, such as for university admissions. For example, von Davier and Burstein (2024[29]) discuss several practices toward human involvement in AI decisions to ensure ethical, accountable, and valid use. These include ongoing human oversight of automated scoring and item generation, systematic review of algorithmic outputs for fairness and bias, engagement of diverse stakeholder groups in test development and validation, and transparent communication of AI roles and limitations to test-takers and institutions. There is also a tension between the continued adoption of rule-based approaches (e.g. college degree audits) and AI evaluation approaches, with hybrid approaches being a fruitful area for future exploration.

Box 11.2. Emerging Initiatives of AI Systems Improving Domains Impacted by Higher Education Administration

AI-powered systems are increasingly being integrated into academic advising and course planning infrastructures, often spearheaded by researchers in collaboration with administrative units. While many of these tools remain student-facing, their design and deployment reflect broader shifts toward data-informed institutional coordination and administrative personalisation.

1. Administrative collaboration and oversight. Course planning platforms such as AskOski at the University of California, Berkeley and ATLAS at the University of Michigan exemplify tools developed in partnership with faculty and administrative stakeholders. These systems aim to guide students through course selection while integrating data from past students' course selection pathways.

2. Faculty and advisor-facing systems. Some platforms are oriented primarily toward professional staff and faculty advisors, supporting them in tailoring study guidance recommendations. For example, ongoing research explores how articulation officers interact with algorithmic guidance to make course articulation decisions (see atain.org).

3. AI in admissions and curricular governance. Beyond advising, machine learning and analytics increasingly support institutional decision-making in admissions and curriculum planning. Administrative priorities should be the starting point to help co-develop or curate algorithmic systems.

4. Towards analytics-informed curricular innovation. Institutions may adopt systems that combine machine learning-driven insights with human advisor expertise to rethink credit hours, assimilate curricular content to new content standards, and enhance curriculum-level analytics, including of course difficulty.

These examples demonstrate how research-administrator collaborations for course advising, transfer, and curricular planning platforms are becoming testbeds for the future of responsible AI integration in institutional workflows. Their success hinges on administrative collaboration, methodological rigour, and thoughtful engagement with human expertise.

AI for curriculum analytics

In the last 10 years, the field of learning analytics has increasingly expanded from student-facing dashboards and systems to analytics that improve programme evaluation, curriculum design, and course delivery in higher education (Greer et al., 2016[30]). Although a recent review of the literature concluded that there is a lack of curriculum analytics studies investigating how these AI systems influence higher education stakeholders (De Silva et al., 2024[31]), we summarise case studies that offer clear perspectives into how curriculum can be designed using machine learning models and AI trained on enrolment, course, and other institutional data. As generative AI becomes increasingly capable in explaining learning analytics to stakeholders (Yan et al., 2025[32]), innovation informed by AI will increasingly shape institutional workflow practices.

For instance, recent work has demonstrated how curriculum analytics can be enhanced with statistical and psychometric techniques to identify inequities in course difficulty and monitor changes over time. Baucks et al. (2024[33]) introduced Differential Course Functioning (DCF), an Item Response Theory (IRT)-based method that controls for overall student performance while detecting systematic differences in course-specific success rates between student groups. Applied to data from over 20 000 undergraduates, the Differential Course Functioning method revealed patterns linked to disciplinary alignment and preparedness, guiding targeted interventions for students taking courses outside their major and transfer students. In a complementary study, Baucks et al. (2024[34]) applied IRT to quantify temporal shifts in course difficulty, finding a marked downward trend during the COVID-19 pandemic and proposing IRT-adjusted pass rates to mitigate the confounding effects of fluctuating cohort performance. Both approaches provide actionable evidence for policymakers, accreditation bodies, and student advisors aiming to improve fairness and consistency in academic programs.

Analytics have also been used to address mismatches between credit hours and actual student workload, offering an actionable basis for curriculum analytics. Credit hours, while central to degree requirements and course planning, explained only 6% of the variance in how students perceive their course workload in Pardos et al. (2023[35]), whereas learning management system (LMS) features based on forum, assignment, and submission activity explained six times more variance (36%) in measures of time load, mental effort, and psychological stress. LMS indicators such as number of assignments and late-semester course drop ratios as well as historical course GPAs provided a more accurate reflection of the student experience, giving institutions greater confidence in these measures as a basis for action. Building on this, Borchers and Pardos (2023[35]) developed Course Load Analytics (CLA), a predictive model that integrates LMS and enrolment features to estimate perceived workload at the course and semester level. Applied across an entire university catalogue across a full undergraduate degree duration, CLA revealed that first-semester students-particularly in STEM fields-often carry some of the heaviest predicted workloads despite low credit-hour counts (Figure 11.3), a hidden load linked to higher attrition. Such meso-level insights position CLA as a practical tool within curriculum analytics, enabling institutions to redesign programme structures, adjust course sequencing, and align workload expectations with student capacity. Because CLA's modelling approach generalises to new courses and contexts, institutions can deploy it broadly to monitor and balance workloads, improving retention and the overall first-year experience.

Figure 11.3. Average semester-level credit hours (left) and predicted semester workload (right) for STEM and non-STEM at a large public university in the United States
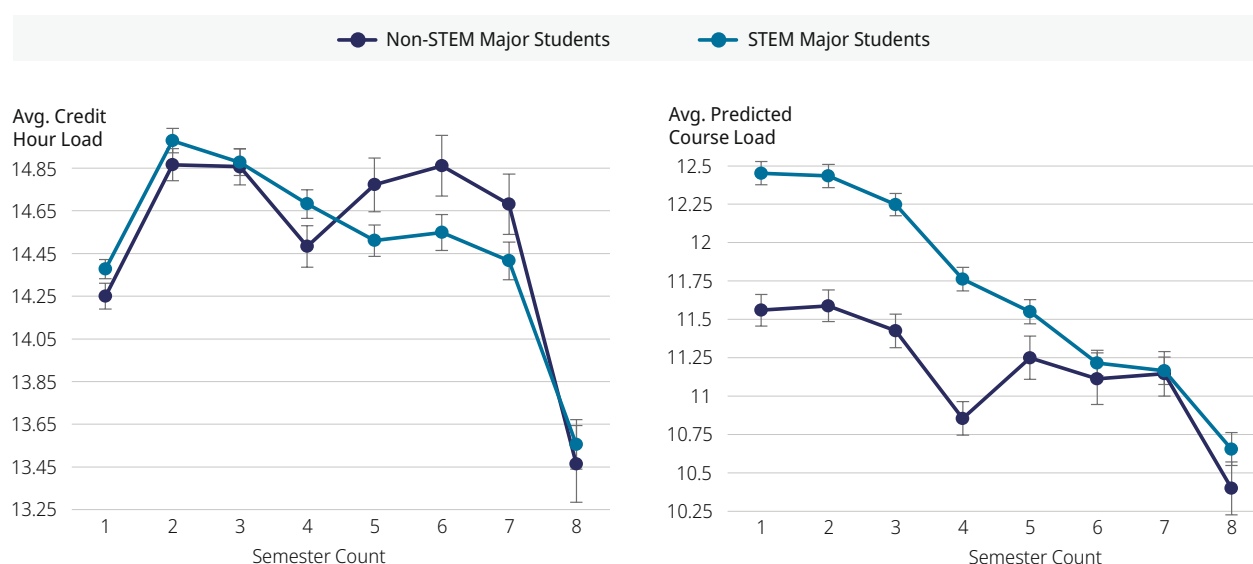
Note: While first-year workload is nominally among the lowest based on credit hours, machine learning predictions based on institutional data approximating student workload perceptions highlight that it is among the highest, revealing a discrepancy enabling curricular redesign. The left panel shows the lack of difference in hours between STEM and non-STEM majors while the right panel shows the big difference in workload during the first 5 semesters.

Source: Borchers and Pardos (2023[35]), "Credit hours is not enough: Explaining undergraduate perceptions of course workload using LMS records", *The Internet and Higher Education*, Vol. 56, https://doi.org/10.1016/j.iheduc.2022.100882.

Looking ahead, curriculum analytics research continues to span a wide range of curriculum-related areas, including programme structures, course sequencing, competency attainment, workload measurement, and curriculum-employability alignment. Studies examine curriculum components from multiple angles, such as mapping prerequisite networks, identifying instructional bottlenecks, tracking competency coverage, modelling student progression pathways (De Silva et al., 2024[31]), and analysing elective course selection strategies (Srivastava et al., 2024[36]).

Recent work has also expanded curriculum analytics beyond student modelling and workload analysis to include the automation of academic record processing through direct application of generative AI in institutional workflows. For instance, Bhaskaran and Pardos (2025[37]) conducted a comparative study of Optical Character Recognition (OCR) and vision-language model pipelines for transcript evaluation, a critical task in credit transfer and course articulation. They showed that combining OCR with semantic reasoning through multimodal models such as GPT-4o1 and Claude 3.7 achieved extraction accuracies above 90 percent, reducing the manual effort needed to align courses and grades across institutions, and thus facilitate transfer and degree/credit recognition across institutions domestically (and potentially internationally).

By converting unstructured transcript data into structured formats suitable for downstream analysis, these methods help connect administrative processes with analytical insights in curriculum design and enable scalable pipelines for workload modelling and course equivalency mapping. In the future, Curriculum Analytics research is likely to close the gap between research applications and practical impact gained through sustainable deployment of tools, as identified by De Silva et al. (2024[31]), by improving efficiency in syllabus, learning objective, and content generation using large language models (Sridhar et al., 2023[38]) and by supporting the wider deployment of university-level initiatives such as course load analytics.

While currently being developed in higher education, where course diversity is much greater than at the school level, those techniques could also be used for middle and upper secondary education in education systems where students can choose different tracks, majors or options to get their high school diploma. This may either highlight the relative difficulty and workload of choosing different study paths or change their perception or design and provide more equal opportunities to all students to study in higher education, regardless of their subject preferences.

Box 11.3. Administrative goals and outcomes of AI-enabled institutional systems

AI systems deployed in institutional workflows aim to advance a range of administrative and student success outcomes. While specific goals may vary by domain (e.g. advising, transfer, content infrastructure), a shared emphasis lies in improving efficiency, personalisation, and educational equity. Below are key outcome categories and associated metrics that institutions may target:

Student progress and completion

Time-to-degree: Reduction in semesters or credits attempted to graduate. Curriculum analytics can help identify student mismanagement during academic planning (e.g. workload overload) and help learners arrive at more balanced course sets (Pardos, Borchers and Yu, 2023[35]).

Credit applicability: Increased proportion of transferred or enrolled credits that count toward degree requirements. AI can help match more course equivalencies that help students graduate (Pardos, Chau and Zhao, 2019[14]).

GPA and academic performance: Stability or improvement in course and cumulative GPA under AI-supported planning systems. Curriculum analytics can help identify which parts of a course are challenging or lack instructional effectiveness, guiding more effective re-design.

Retention and persistence: Improved term-to-term and year-to-year retention, particularly among at-risk or transfer students. Curriculum analytics can identify which courses are particularly challenging for transfer students (Baucks et al., 2024[33]) and help allocate institutional resources.

Transfer and articulation outcomes

Articulation coverage: Expansion in the number and accuracy of course equivalencies across institutions (Goulart and Pardos, 2025[39]).

Credit mobility: Decrease in articulation loss, especially for community college students, through better support of equivalency officers through AI-based articulation recommendations (Xu, Pardos and Pai, 2023[17]).

Degree completion for transfer students: Increase in successful 4-year degree attainment among students transferring from 2-year institutions.

Time and administrative effort saved: Reduction in manual reviews required by articulation officers and advisors or total review time (Xu, Pardos and Pai, 2023[17]).

Advising and personalisation

Workload fit: Better alignment between student capacity and course intensity, potentially reducing dropout or course failure, which has been shown to be associated with student workload (Borchers and Pardos, 2025[40]).

Well-being indicators: Improved student reports of stress, burnout, or overload through course workload analytics when supported by personalised advising tools (Pardos, Borchers and Yu, 2023[35]; Borchers and Pardos, 2025[40]).

Advisor efficiency: Enhanced capacity of advisors to manage caseloads through intelligent recommendations (e.g. regarding workload or course sequences: Khan and Polyzou, 2024[15]) and predictive alerts.

Curriculum and learning infrastructure

OER discoverability and reuse: Increased alignment between institutional curricula and open educational resources (Li, Pardos and Ren, 2024[20]). Improved alignment models could significantly help reduce human authoring and search time.

Content production efficiency: Reduction in faculty time spent authoring assessments or tagging materials, through generative tools.

Curriculum analytics: Enhanced ability to detect gaps, redundancies, or misalignments in curriculum via structured metadata and content classification.

Rationale for adoption and future direction

We see three practical reasons for why AI will be increasingly adopted by relevant stakeholders in institutional educational workflows.

First, content generation can yield strong increases in authoring efficiency that come with substantial economic cost reduction. For instance, Reza et al. (2025[26]) found that PromptHive-an open-source collaborative prompt-authoring interface for the OATutor adaptive tutoring system-enabled subject-matter experts to produce AI-generated maths hints of comparable instructional quality to exclusively human-authored hints with no AI support. The tool also reduced perceived cognitive load by half, shortened authoring time by more than twenty-fold, and was found to be substantially more usable than the legacy authoring interface. In a controlled study with over 350 learners, the AI-assisted hints achieved student learning gains statistically indistinguishable from those of expert-written materials, demonstrating that human-centred prompt-engineering workflows can preserve expert control and quality standards while dramatically increasing the scalability of educational content creation. These improvements in content generation can be useful to both (a) higher education course instructors seeking to revise practice problems and assessment and (b) primary and secondary education instructors and content vendors seeking to revise curricular sequences.

Second, as many higher education institutions compete for the most capable and promising students, pressure to adopt AI that enables better course offerings and equivalences may increase. As transfer pathways improve, students will become more likely to select institutions that recognise a greater share of their prior learning and minimise credit loss, directly affecting enrolment and completion outcomes. Studies have found that seamless credit transfer is a strong predictor of degree completion among community college transfer students, and that improved articulation coverage can reduce time-to-degree and attrition (Monaghan and Attewell, 2015[41]; Hodara et al., 2017[42]). This is also a key development to make lifelong learning a reality.

Third, AI-augmented advising can directly improve other outcomes that education institutions care about-retention, on-time graduation, and advisor capacity-thereby improving cost effectiveness and institutional appeal. Advisor-facing recommender and triage tools can help tailor course plans, while course workload analytics enable programs to audit sequences, rebalance hidden load, and reduce course withdrawal and failure where overload is detected (Pardos, Borchers and Yu, 2023[35]; Borchers and Pardos, 2025[40]). In parallel, machine-learning approaches to curriculum analytics can surface when nominally identical offerings fluctuate in difficulty over time and across groups, informing targeted redesign and quality assurance (Baucks, Schmucker and Wiskott, 2024[34]; Baucks et al., 2024[33]).

Across domains as varied as credit transfer, advising, admissions, and curriculum management, artificial intelligence is becoming part of the institutional infrastructure of higher education and beyond. These applications signal a shift in how institutions manage complexity, not by automating human decisions, but by introducing new forms of prediction, representation, and adaptive support. Properly designed, such tools augment institutional and education system judgment, surfacing course equivalencies that might otherwise be missed, helping advisors tailor guidance to individual students, and enabling administrators to detect inequities or inefficiencies across programmes.

With this integration comes a need for governance and privacy frameworks that can sustain trust. New privacy-preserving and open-source large language models, such as the Swiss-developed LLaMA-Open series, illustrate how innovation can proceed while respecting data sovereignty and transparency. Beyond regulatory frameworks, research can contribute frameworks for AI-supported analytics, as demonstrated in seminal models for responsible data use (Drachsler and Greller, 2016[43]), offering templates for institutions developing AI policies today. Rapidly evolving technologies merit adaptive rather than static governance, that is, principles and review mechanisms that evolve alongside technology.

What distinguishes this emerging wave of institutional AI is less the technology itself than the collaborative ecosystems that enable it. Many of the most promising examples discussed in this chapter arose from partnerships among researchers, administrators, and platform developers. Such collaborations are critical for aligning technical design with institutional and end-user values and for empirically studying the consequences of AI adoption on efficiency, equity, and educational quality.

Although the focus of this chapter has been on higher education, many of the same institutional affordances extend naturally to lower levels of education as well as adult learning, where advising, assessment, and curriculum alignment face similar pressures. Generative models can assist teachers by creating or reviewing assessment items,

tagging resources across standards, and generating formative feedback, thereby reducing workload and enhancing standardisation. It is likely that some institutional AI will merge toward a connected, data-informed educational ecosystem that links primary and secondary education and postsecondary systems, improving personalisation, equity, and mobility across educational pathways.

Note

1.   https://atain.org/

References

Baucks, F., R. Schmucker and L. Wiskott (2024), "Gaining Insights into Group-Level Course Difficulty via Differential Course Functioning", *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pp. 165-176, https://doi.org/10.1145/3657604.3662028.   [33]

Baucks, F., R. Schmucker and L. Wiskott (2024), "Gaining Insights into Course Difficulty Variations Using Item Response Theory", *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 450-461, https://doi.org/10.1145/3636555.3636902.   [34]

Bhandari, S., Y. Liu, Y. Kwak, Y. and Z. Pardos (2024), "Evaluating the psychometric properties of ChatGPT-generated questions", *Computers and Education: Artificial Intelligence*, Vol. 7, p. 100284, https://doi.org/10.1016/j.caeai.2024.100284.   [25]

Bhaskaran, M.A and Z. Pardos (2025), *Automating Academic Transcript Evaluation: A Comparative Study of OCR Techniques for Course and Grade Evaluation*, https://doi.org/10.1145/3698205.3733958   [37]

Borchers, C. (n.d.), *Performance under high course workload: A Big Five personality trait account in higher education* [Manuscript under review].   [23]

Borchers, C. and Y. Pardos (2025), "Workload overload? Late enrolment leads to course dropout", *Journal of Educational Data Mining*, Vol. 17(1), pp. 126-156, https://doi.org/10.5281/zenodo.14907388.   [40]

Borchers, C. and Z. Pardos (2023), "Insights into undergraduate pathways using course load analytics", *LAK23: 13th International Learning Analytics and Knowledge Conference*, pp. 219-229, https://doi.org/10.1145/3576050.3576081.   [8]

Corbin, T., P. Dawson, K. Nicola-Richmond and H. Partridge (2025), "'Where's the line? It's an absurd line': towards a framework for acceptable uses of AI in assessment", *Assessment & Evaluation in Higher Education*, Vol. 50/5, pp. 705-717, https://doi.org/10.1080/02602938.2025.2456207.   [28]

De Silva, L., M. Rodríguez-Triana, I.-A. Chounta and G. Pishtari (2024), "Curriculum analytics in higher education institutions: a systematic literature review", *Journal of Computing in Higher Education*, Vol. 37/3, pp. 898-944, https://doi.org/10.1007/s12528-024-09410-8.   [31]

Drachsler, H. and W. Greller (2016), "Privacy and analytics", *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, pp. 89-98, https://doi.org/10.1145/2883851.2883893.   [43]

Fischer, C., Z. Pardos, R. Baker, J. Williams, P. Smyth, R. Yu, . . . M. Warschauer (2020), "Mining Big Data in Education: Affordances and Challenges", *Review of Research in Education*, Vol. 44/1, pp. 130-160, https://doi.org/10.3102/0091732x20903304.   [2]

Goulart, S. and Z. Pardos (2025), *Measures of Articulation Coverage and Credit Granting, Center for Open Science*, https://doi.org/10.31219/osf.io/rp67h_v2.   [39]

Goulart, S. and Z. Pardos (2025), *The Impact of Institutional Resources on Articulation Agreement Coverage*, Center for Open Science, https://doi.org/10.35542/osf.io/y7avd_v1.   [6]

Greer, J., M. Molinaro, X. Ochoa and T. McKay (2016), "Learning analytics for curriculum and program quality improvement (PCLA 2016)", *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, pp. 494-495, https://doi.org/10.1145/2883851.2883899.   [30]

Grosges, T. and D. Barchiesi (2007), "European Credit Transfer and Accumulation System: An Alternative Way to Calculate the ECTS Grades", *Higher Education in Europe*, Vol. 32/2-3, pp. 213-227, https://doi.org/10.1080/03797720701840807.   [12]

Hodara, M., M. Martinez-Wenzl, D. Stevens, D. and C. Mazzeo (2017), "Exploring Credit Mobility and Major-Specific Pathways: A Policy Analysis and Student Perspective on Community College to University Transfer", *Community College Review*, Vol. 45/4, pp. 331-349, https://doi.org/10.1177/0091552117724197.   [42]

Ignash, J. and B. Townsend (2000), "Evaluating State-Level Articulation Agreements According to Good Practice", *Community College Review*, Vol. 28/3, pp. 1-21, https://doi.org/10.1177/009155210002800301.   [13]

Johnson, H. (2020), *Increasing Community College Transfers: Progress and Barriers*, https://www.ppic.org/publication/increasing-community-college-transfers-progress-and-barriers/ (accessed on 3 December 2025).   [11]

Khan, M. and A. Polyzou (2024), "Estimate Undergraduate Student Enrollment in Courses by Re-purposing Recommendation Tools", *The International FLAIRS Conference Proceedings*, Vol. 37, https://doi.org/10.32473/flairs.37.1.135584.   [15]

Kim, M., S. Raghuraman, A. Puder, A. and C.H. Yang (2025), *Predicting Course Transferability Using Deep Embeddings and Traditional Classifiers*, https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.poster-demo-papers.282/index.html (accessed on 3 December 2025).   [16]

Kizilcec, R., R. Baker, E. Bruch, K. Cortes, L. Hamilton, D. Lang,… M. Stevens (2023), "From pipelines to pathways in the study of academic progress", *Science*, Vol. 380/6643, pp. 344-347, https://doi.org/10.1126/science.adg5406.   [1]

Lakin, M., D. Seymour and C. Crandall (2015), *Credit for Prior Learning: Charting Institutional Practice for Sustainability*, http://www.acenet.edu/news-room/Documents/Credit-for-Prior-Learning-Charting-Institutional-Practice-for-Sustainability.pdf.   [3]

Lekan, K. and Z. Pardos (2025), "AI-Augmented Advising", *Journal of Learning Analytics*, Vol. 12/1, pp. 110-128, https://doi.org/10.18608/jla.2025.8593.   [22]

Liu, Y., S. Bhandari and Z. Pardos (2025), "Leveraging LLM respondents for item evaluation: A psychometric analysis", *British Journal of Educational Technology*, Vol. 56/3, pp. 1028-1052, https://doi.org/10.1111/bjet.13570.   [19]

Li, Z., Z. Pardos and C. Ren (2024), "Aligning open educational resources to new taxonomies: How AI technologies can help and in which scenarios", *Computers and Education*, Vol. 216, https://doi.org/10.1016/j.compedu.2024.105027.   [20]

Li, Z., C. Ren, X. Li, and Z. Pardos (2021), "Learning Skill Equivalencies Across Platform Taxonomies", *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 354-363, https://doi.org/10.1145/3448139.3448173.   [4]

Monaghan, D. and P. Attewell (2015), "The Community College Route to the Bachelor's Degree", *Educational Evaluation and Policy Analysis*, Vol. 37/1, pp. 70-91, https://doi.org/10.3102/0162373714521865.   [41]

Moore, J., D. Grabb, W. Agnew, K. Klyman, S. Chancellor, D. Ong and N. Haber (2025), "Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.", *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 599-627, https://doi.org/10.1145/3715275.3732039.   [24]

Moore, S., H. Nguyen, T. Chen, T. and J. Stamper (2023), "Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods", in *Lecture Notes in Computer Science, Responsive and Sustainable Educational Futures*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-42682-7_16.   [27]

Pardos, Z.A. and A.J. Hun Nam (2020), "A university map of course knowledge", *PLOS ONE*, Vol. 15/9, https://doi.org/10.1371/journal.pone.0233207.   [7]

Pardos, Z., C. Borchers and R. Yu (2023), "Credit hours is not enough: Explaining undergraduate perceptions of course workload using LMS records", *The Internet and Higher Education*, Vol. 56, https://doi.org/10.1016/j.iheduc.2022.100882.   [35]

Pardos, Z., H. Chau and H. Zhao (2019), "Data-Assistive Course-to-Course Articulation Using Machine Translation", *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, pp. 1-10, https://doi.org/10.1145/3330430.3333622.   [14]

Pizarro Milian, R. and J. Aurini (2025), "Barriers to Articulation: Resource Dependencies, Networks and Status Dynamics", *Higher Education Policy*, https://doi.org/10.1057/s41307-025-00423-1.   [5]

Ren, C. and Z. Li (2024), *Human-AI Collaboration Increases Skill Tagging Speed but Degrades Accuracy*, http://arXiv preprint arXiv:2403.02259.   [21]

Reza, M. et al. (2025), "PromptHive: Bringing Subject Matter Experts Back to the Forefront with Collaborative Prompt Engineering for Educational Content Creation", *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1-22, https://doi.org/10.1145/3706598.3714051.   [26]

Shao, E., S. Guo and Z. Pardos (2021), "Degree Planning with PLAN-BERT: Multi-Semester Recommendation Using Future Courses of Interest", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35/17, pp. 14920-14929, https://doi.org/10.1609/aaai.v35i17.17751.   [10]

Sridhar, P., A. Doyle, A. Agarwal, C. Bogart, J. Savelka and M. Sakr (2023), Harnessing LLMs in curricular design: Using GPT-4 to support authoring of learning objectives, http://CEUR-WS.org. https://ceur-ws.org/Vol-3487.   [38]

Srivastava, N., S. Nawaz, Y.-S.Tsai and D. Gašević (2024), "Curriculum Analytics of Course Choices:", *Journal of Learning Analytics,* Vol. 11/1, pp. 116-131, https://doi.org/10.18608/jla.2024.8095.   [36]

Vincent-Lancrin, S., C. Cobo Romaní and F. Reimers (eds.) (2022), *How Learning Continued during the COVID-19 Pandemic: Global Lessons from Initiatives to Support Learners and Teachers*, OECD Publishing, Paris, https://doi.org/10.1787/bbeca162-en.   [18]

von Davier, A. and J. Burstein (2024), "AI in the Assessment Ecosystem: A Human–Centered AI Perspective", in *Intelligent Systems Reference Library, Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-71232-6_6.   [29]

Whitehill, I. (ed.) (2020), *Evaluating Sources of Course Information and Models of Representation on a Variety of Institutional Prediction Tasks*, in A. Rafferty & J. R. Whitehill (Eds.).   [9]

Xu, L., Z. Pardos and A. Pai (2023), "Convincing the Expert: Reducing Algorithm Aversion in Administrative Higher Education Decision-making", *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pp. 215-225, https://doi.org/10.1145/3573051.3593378.   [17]

Yan, L., R. Martinez-Maldonado, Y. Jin, V. Echeverria, M. Milesi, J. Fan,... D. Gašević (2025), "The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics", *Computers & Education*, Vol. 234, p. 105322, https://doi.org/10.1016/j.compedu.2025.105322.   [32]

12

Generative AI for standardised assessments:

A conversation with Alina von Davier

This chapter is an interview between Alina von Davier (Duolingo and Edastratech, United States) and the OECD Secretariat. The conversation is about new possibilities offered by generative AI (GenAI) to develop and implement standardised and high-stakes assessments. After showing how GenAI can enhance the productivity of item development, the discussion dwells on how some innovations are made possible thanks to GenAI, taking the assessment of foreign languages as a case in point. The processes have to remain tightly controlled by humans and usually involve different types of artificial intelligence.
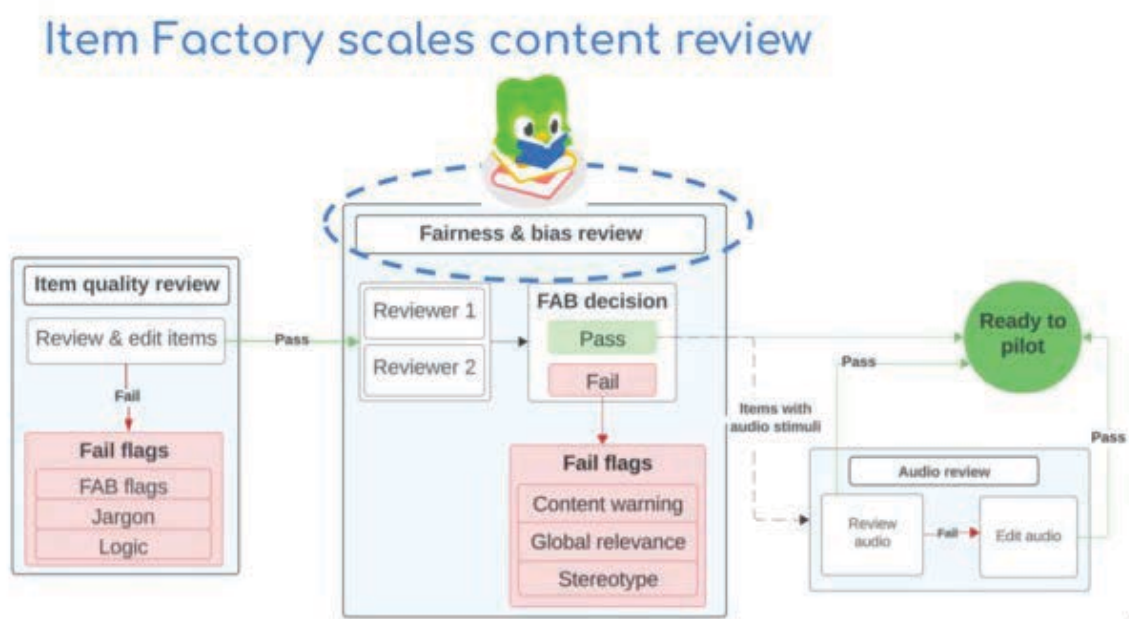
Enhancing the productivity of item design

OECD: Many AI tools work quite effectively for assessment, sometimes even better than generative AI. However, we want to explore what new possibilities generative AI can offer. We are interested in two main areas: how generative AI can help perform traditional assessment tasks more effectively – for instance, item generation for standardised assessments – and how it can enable different and better assessments . Perhaps we can start with the first aspect?

Alina von Davier: For assessment, at Duolingo we utilise AI end-to-end, but not in isolation: it just contributes to our processes. For example, we use generative AI to generate items at scale after human experts have designed item prototypes. The content experts collaborate with the AI engineers and psychometricians during the design phase in order to ensure that the design of the new item type is viable in operational settings. Once the initial item design is complete and we are comfortable with it, the AI

engineers and scientists define the desired display of the items for the online delivery and create scripts that will generate multiple items on a large scale. After that, the items are reviewed by human experts for quality, fairness, and appropriateness for a global administration. This has become a mainstream process. We created the item factory, a full system where humans and machines collaborate. Let me emphasise that there is a substantial amount of work involved at the outset. It demands a very high level of expertise and considerable effort when you set up a GenAI system for a particular item type for the first time. That's where the main work lies: in setting things up and determining what works and what doesn't. However, once that setup is complete, compared to human development the efficacy increases tenfold. It's incredible in terms of speed and cost. We remain very conservative and continue to use humans to review every single item, but we plan to explore how to make this even more efficient.

Figure 12.1. A process for human raters to review assessment items generated with GenAI

Improving the assessment of writing and speaking skills

OECD: Beyond enhancing the productivity of conventional assessment design, how can we use GenAI to innovate the ways we assess people's knowledge and skills?

Alina von Davier: In our Duolingo English Test, we have, for example, two other applications of GenAI: one for a writing task and one for a speaking task. In April 2024, we launched a new writing task, which works as follows. We provide a prompt, such as "Please write about topic X – you have y minutes", and after they complete the assignment, AI intervenes in real time, analysing the text that has been written so far and comparing it to a set of themes we created for that specific item. The AI then acts as a peer or a professor, suggesting to continue the writing by covering new sub-topics, for example asking "Can you also write about this?". This type of interactive capability for such item was not possible before GenAI: it allows to more closely resemble a real-life task, thereby offering greater authenticity – a problem that most assessments have.

More recently, in July 2025, we launched an interactive, adaptive speaking task, during which a test taker converses with an AI agent. The generative aspect primarily involves generating the agent's utterances. While it's not a live agent and is extremely constrained, it allows us to create interactivity, as was the case for the

writing item. These two examples use generative AI to assess differently than before. It would not be possible without the technology.

The writing task involves only one intervention from the AI. You receive a prompt, you write, and then the AI comes in and asks you to write more about a specific topic. In contrast, the speaking task is a conversation, involving multiple interactions. Managing these multiple interactions is what makes it challenging. It's difficult because the AI needs to be embedded to "understand" what the person says. When test-takers are non-native speakers of the language that is being tested, meaning they have all types of accents and abilities, a lot of work goes into ensuring the AI can understand each person's speech, evaluate it, and then select the appropriate response to that person. So, it's actually much more difficult to implement than the writing task. To my knowledge, this is for the first time where a high-stakes assessment with million of test takers includes such an interactive and adaptive speaking task.

OECD: What is the purpose of these two tasks? What do you want to assess (or to achieve)?

Alina von Davier: As I mentioned before, these two tasks are examples of something that could not be achieved previously. We are trying to accomplish two

main objectives with this approach: authenticity and support to test-takers. Before, you could have just said, "Write for 10 minutes," rather than evaluating what they wrote in the middle of the assignment, and then encourage the test takers to write more. The key difference here is that in real life, people often have their writing reviewed and receive suggestions on how to proceed further.

Take the writing task. First, we believe this makes the task more akin to real-life situations. For instance, at college/university, someone reviews your writing, provides feedback, and asks you to expand on it. We do the same. This is the interactivity aspect. But second, we also aim to assist test-takers. When we provide an initial prompt for the writing task, it often has multiple potential writing directions. We want to encourage test-takers to cover other aspects they haven't yet addressed, helping them become better writers and giving them an opportunity to demonstrate their ability to write about different topics. At the end, what we assess is the quality of their writing in English as a foreign language.

Regarding the speaking task, there is currently no other high-stakes test that uses purely technology-enabled interactive speaking. Previously, tests might have offered a prompt, you listened and responded, then listened to something else and responded, but it was neither interactive nor adaptive. Our speaking task is both adaptive and interactive. For example, if a test-taker's English proficiency level is not very high, the AI agent will adjust and engage in a simpler conversation. This was simply not possible before. The only other English proficiency test that features a real, back-and-forth interview is the IELTS English test, but they conduct it with humans. As a test-taker, you have to schedule an appointment, travel to a test centre, and speak with a human. We are trying to maintain that conversation but replace the logistics that comes with traveling to a centre. It's an extremely expensive and difficult process for test takers to travel to a centre and take a test delivered by human interviewers. Our test is continuous – it can be taken anytime, anywhere, hence a technology-based solution makes more sense for this delivery model. Furthermore, humans have their own issues, such as the halo effect, for example: if a test-taker responds well on one question, the examiner may transfer this positive impression to evaluating the following questions.

OECD: What is the level of efficacy of these tasks? I assume you've tested how they perform for those taking the test: does it work well? How does it compare to human raters? And finally, how do you combine traditional AI with generative AI in this highly constrained AI scenario, given your requirements like adaptivity and potentially broadening the topics of the conversation? How does it work?

Alina von Davier: After the setup, the generation of these tasks at scale is extremely efficient. Moreover, the quality is outstanding and experts cannot tell them apart from those that are generated by humans.

To your second question: the work we do is not solely with generative AI; other types of AI models and psychometric models are also invoked. GenAI and other AI programmes work together. We write models and scripts, and we have many scripts that call upon AI and GenAI for different applications. People have to realise that GenAI is just one type, and we use many other types as well.

Let's consider the specific speaking task I described above. Parts of it can only be accomplished with generative AI, such as understanding what the person says and evaluating it quickly in real time. Then, as part of the rest of the scripts, having the agent selecting the correct answer and delivering the right spoken response to the test-taker is not solely a GenAI process. It includes psychometrics as well. So, there is a component that only generative AI can handle, but other parts that are done by other types of AI and psychometrics, hence computational psychometrics. For the writing task, at the scale at which we operate, the real-time reading of the text can only be done by generative AI. However, everything I've described is embedded within other scripts and programmes. GenAI is not used as a standalone process where we simply give the GenAI a prompt and say, "Do this". We have a script that designs the prompt and then feeds it to the GenAI. It's quite an elaborate process. That's why I mentioned that "setting it up" requires a great deal of expertise and time.

OECD: So, what is the next step after the writing and speaking tasks? Is it leading to another task or a score? Do you score your test-takers with AI?

Alina von Davier: Yes, scoring is done with AI and psychometrics, but using machine learning, not GenAI. For scoring the speaking task, there is one component that relies on generative AI: the evaluation of pronunciation. We employ multiple forms of AI, not just large language models (LLMs). For instance, we use an automatic speech recognition (ASR) system for sound processing, text-to-speech, and speech-to-text. So, yes, we already use AI for scoring, and also even for proctoring – that's what I meant by end-to-end. However, we also use psychometric models to obtain the final score and evaluate the reliability and validity of the scores.

Our next significant task involving GenAI will be the provision of feedback, starting with writing. This is almost ready, but it's primarily for the practice part of the test. Another desired improvement would be to relax some of the constraints on the AI speaking agent. We need to be confident that we maintain comparability though, which

is extremely important in standardised testing. If you allow an LLM to operate independently, you risk losing comparability. One time you get one response, and the next time a different response. While that's acceptable in some contexts, it's not suitable for high-stakes tests. That's why we have these constraints. We can, of course, adjust or relax them to make the tests more authentic, but we still need to maintain the accuracy and comparability required for a quality exam.

OECD: Thank you so much for sharing how AI and GenAI are involved in current assessments.

Alina von Davier: My pleasure. If you go to the Duolingo English Test website, there's something called a practice hub. Just take a look to see what these items are like. It's all free and open, so anyone interested to know more can just check it out.

High-stakes assessment and next steps

OECD: Precisely, for high-stakes tests, there has been intense discussions about the grading of exams and papers. It seems that "good old-fashioned AI" performs better than generative AI in terms of accuracy and consistency. People seem to suggest that older machine learning tends to be more accurate but less flexible. It requires a lot of time and money to train, and if you change the task, you often have to restart and redo it. Whereas with generative AI, you don't lose as much time because it adapts much better to new contexts, but you might not achieve the same level of accuracy. This means that, if it's not a high-stakes scenario, it's perfectly acceptable, but if it is high-stakes, you need to think carefully about your approach. Do you agree with this assessment? Do you think this is likely to change? What is your experience?

Alina von Davier: First, let me say that the most important is to know how to use GenAI properly. Simply using a prompt one dreamed up one morning is not going to lead to good quality assessment. It doesn't work like that. One has to think carefully and plan. People hope that if they just say, "Do this for me," it will do it perfectly, but that's not the case. We need to be very cautious when people make overly positive or negative claims: take it with a grain of salt because most people may not have built sufficient experience with GenAI. That's my main observation now. Many think it's easy and conversational, but it isn't that simple for high-quality exams. One may obtain a full range of output quality from generative AI if one prompts it properly or if one builds it correctly. When I teach the use of generative AI for item generation, I advise people to "divide and conquer". By that, I mean, don't try to get everything done with one prompt. Let's say one needs an assessment passage followed by questions for 8th graders. Don't put it all into one prompt; it won't be very good. I suggest for people to divide the task: first generate the passage, review it to ensure its quality,

and when one is satisfied, develop the questions. That approach works so much better, but it is not yet widespread. Many people try to put everything into one prompt, don't experiment further, and then claim it's not working well. So, I would say, be careful and ask more questions with both extreme claims, positive or negative.

OECD: It's an interesting observation because there is a new research literature on how many prompts are needed to achieve comparable quality to older machine learning types of scoring.

Alina von Davier: For scoring, we use our own machine learning models, sometimes with some generative AI components. We use our own models, but our biggest concern – as big as accuracy – is comparability. If generative AI scores the same essay one way at one time and differently at another, it affects comparability and replicability, which are crucial for a reliable assessment .

GenAI can sometimes be accurate, but again, it depends on how one uses it. For instance, if one is exploring using generative AI for scoring essays, and if one provides it with a very good rubric (the same one one would give to humans) and a few more examples than what one would usually do, I believe it can do quite a good job. However, this depends on the task's complexity and the purpose of the scores. It varies. It's true that all these applications are still task-specific for both generation and scoring. While some parts can be reused, generally, the main model is task-specific, and one needs to test it again to see if it works for other tasks.

We also incorporate numerous checks afterwards, with multiple filters to ensure nothing is released if it's not good enough. We have automatic tools for monitoring the quality of our assessments. The first tool we built is called AQuAA, which stands for Analytics for Quality Assurance in Assessment. This tool incorporates some machine learning models and a lot of psychometrics. It functions as an alert system, continuously analysing data as it comes in. If anything unusual occurs, we receive an

alert that prompts us to investigate what is going on. We have another, newer system called AQuAP, which stands for Analytics for Quality Assurance for the Pools (item pools). It also operates in the background, monitoring the items to ensure they don't suddenly become more difficult or exhibit other unusual behaviours within the pool. This is also an automatic tool that heavily utilises machine learning. We are also developing one called AQuATT, for test-takers, which will focus on the test-taker level.

OECD: Thank you so much for sharing how AI and GenAI are involved in current assessments.

Alina von Davier: My pleasure. If you go to the Duolingo English Test website, there's something called a practice hub. Just take a look to see what these items are like. It's all free and open, so anyone interested to know more can just check it out.

References for further reading

Burstein, J., G. T. LaFlair, K. Yancey, A. von Davier, and R. Dotan (2024), "Responsible AI for Test Equity and Quality: The Duolingo English Test as a Case Study".   [4]

Hao, J., A. von Davier, V. Yaneva, S. Lottridge, M. von Davier, D. J. Harris (2024), "Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI", *Educational Measurement: Issues and Practice*, Vol. 43/2, pp. 16-29, https://doi.org/10.1111/emip.12602.   [2]

von Davier, A. (2024), *How Will AI Change Adaptive Testing?*   [5]

von Davier, A. and J. Burstein (2024), "AI in the Assessment Ecosystem: A Human–Centered AI Perspective", in *Intelligent Systems Reference Library, Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-71232-6_6.   [1]

von Davier, A., A. Runge, Y. Park, Y. Attali, J. Church, and G. LaFlair (2024), "The item factory", *Machine Learning, Natural Language Processing, and Psychometrics*.   [3]

13 Generative AI and the transformation of scientific research

Dominique Guellec* and Stéphan Vincent-Lancrin**

*Observatory of Science and Technology, France
**OECD

This chapter provides an overview of the various uses of AI in scientific research, and highlights the extent to which those uses are relevant to education research. Specifically, it examines the use of generative AI in facilitating language-related tasks, programming and data, modelling complex phenomena, managing knowledge (in search engines, literature reviews, research reports and summaries) and hypothesis generation, as well as the use of AI research assistants and robots. The chapter examines the possible effects of AI on the creativity of research and the reliability of science – both highly debated issues in the scientific community – which have direct connections with education. It concludes by exploring some of the potential consequences of the use of generative AI for education research.

Introduction

Scientific advancement over time has been characterised by the use of increasing amounts of data for solving issues of increasing complexity. Moreover, scientists have been using growingly powerful instruments that supplement the human brain and senses, like the microscope, medical imaging devices or particle accelerators. Writing, printing and now the computer also allow scientists to store, analyse and communicate scientific information. Artificial intelligence (AI), whose latest avatar is generative AI (GenAI), is another step in the historical trend of science instrumentation, which allows the processing of enormous quantities of complex data in new ways.

Scientific research has been at the forefront of AI adoption and the arrival of GenAI has further accelerated its use, making it more accessible to all scientists. Recent surveys suggest that more than half of scientists now use GenAI tools. GenAI can analyse and generate both structured and unstructured data, including text, tables, statistics, images, videos, diagrams, graphs, chemical and mathematical formulas, DNA sequences and other biological data. It can generate new data with specific properties by learning and re-combining patterns based on its training data; for example, responses to questions, synthetic datasets, predictions such as simulations or weather forecasts, and even simulated agents. GenAI helps researchers accelerate existing research tasks (such as writing text and statistical processing), improve the quality of others' tasks (like editing and producing figures), and perform tasks that were previously out of reach (such as analysing extremely large datasets). In helping facilitate such tasks, GenAI is transforming scientific research.

This chapter will first present an overview of the various uses of AI in science, and highlight the extent to which those uses are relevant to education research. As education research draws on many disciplines (Vincent-Lancrin and Jacotin, 2023[1]), knowing how GenAI is used in a variety of disciplines will help governmental research funders envisage upcoming

changes in education research. The chapter will then examine the possible effects of AI on the creativity of research and the reliability of science, both highly debated issues in the scientific community, and which have direct connections with education. The conclusion will explore some of the possible consequences of the use of GenAI for education research.

The use of GenAI in scientific research

GenAI is used in research for four distinct but related purposes: manipulating language (including scientific language); managing knowledge; generating knowledge; and managing the entire chain of operations in a research project. Accordingly, GenAI models as used in research can be divided into four categories: 1) general purpose models, notably large language models (LLMs) – like GPT, Gemini and others – used by many researchers for generating text, images or computer code; 2) specialised models dedicated to managing various language-related tasks such as literature reviews, refereeing, and generating hypotheses or suggestions for experiments ("ideation"); 3) specialised models used to tackle highly complex scientific problems that involve vast quantities of data or complicated mechanisms (for example, the 3D shape of proteins); 4) research assistants and "robot labs", which autonomously manage entire sequences of operations in a research project, from initial analysis of data to experimentation. We will examine these four categories of models in turn, after presenting the evidence concerning the diffusion of GenAI among researchers. A clear trend of increasing cognitive power and agency of GenAI over time will then appear.

To what extent do researchers use GenAI?

The use of GenAI in research is advancing rapidly, though statistics remain limited. Broader trends in AI adoption can help indicate future GenAI patterns.

AI adoption has been progressing fast. Duede et al. (2024[2]) trace the share of AI-engaged publications (1985-2022), which rose from ~2% in 2015 to 8% in 2022 across all scientific fields (Figure 13.1). Evans et al. (2024[3]), analysing 100 million papers (1980-2024), identify over 1 million AI-assisted papers (1.57% overall), showing pervasive adoption across biology, medicine, chemistry, physics, materials science and geology.

Figure 13.1. Change in AI engagement across all scientific fields
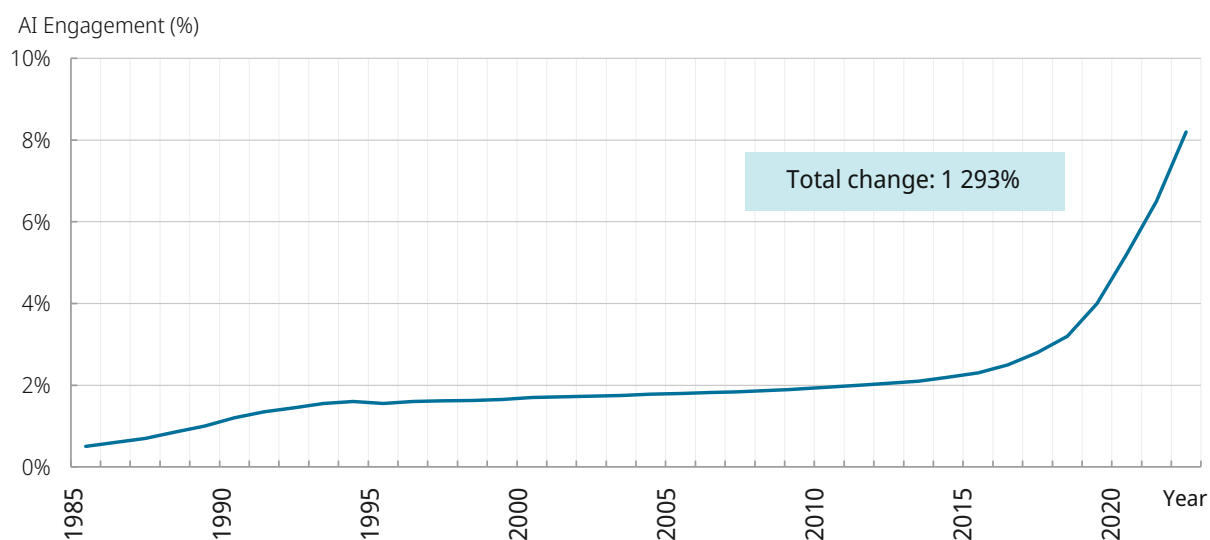
Source: Duede, E., W. Dolan, A. Bauer, I. Foster and K. Lakhani, (2024[2]), Oil & Water? Diffusion of AI Within and Across Scientific Fields.

For GenAI specifically, Liang et al. (2024[4]) studied nearly 1 million papers (2020-2024) and found steady growth in "LLM-modified" papers, with the sharpest rise after ChatGPT's release. Uptake is strongest in Computer Science (17.5%) and weakest in Mathematics (6.3%) (Figure 13.2). The analysis does not include social science, which is one of the main contributors of education research, although natural sciences represent a significant minority of the education research production (Vincent-Lancrin and Jacotin, 2023[1]).

Figure 13.2. Proportion of "LLM-modified papers" by discipline
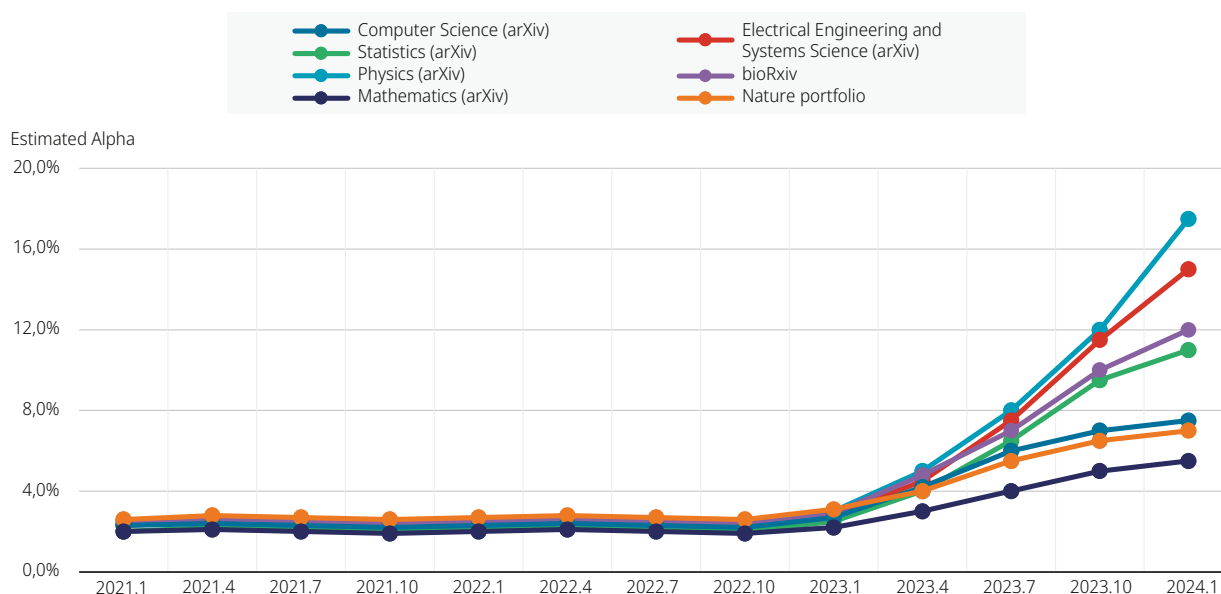
Source: Liang et al. (2024[4]), Can large language models provide useful feedback on research papers? A large-scale empirical analysis, https://arxiv.org/abs/2310.01783.

A survey (conducted in March 2025, with 5 000 researchers) estimates that over half of researchers already use AI for manuscript preparation and error detection (Naddaf, 2025[6]). About one-third use or plan to use GenAI for data collection/processing, while its use for complex tasks (journal choice, citations) remains less common. More than half of respondents believe that AI outperforms humans in tasks like literature review, summarising, plagiarism checks and citation management, anticipating mainstream adoption within two years (Figure 13.3). Early-career researchers show higher enthusiasm than senior colleagues, although many remain cautious about AI's role in higher-level tasks. Despite the differences across fields, one can assume that the uptake for these mainly language-related tasks is similar for education researchers.

Language-related tasks

A first category of uses relates to language: translating, editing, writing and summarising papers. These tasks often involve general-purpose LLMs (ChatGPT, Claude, Gemini, etc.), which are the most accessible, but not exclusively, since some science-specialised tools also have such capabilities. AI can help adapt papers to meet journal submission guidelines, draft abstracts, write peer reviews, and assist in drafting grant proposals (Heidt, 2025[7]). Although researchers were already using some AI writing assistants, the release of LLMs brought a substantial change in the extent of such use (Lenharo, 2024[8]).

The AI model generates text in response to a query (a "prompt") given by the researcher. Here is an example of a prompt: "I'm writing a paper on [topic] for a leading [discipline] academic journal. What I tried to say in the following section is [specific point]. Please rephrase it for clarity, coherence, and conciseness, ensuring each paragraph flows into the next. Remove jargon. Use a professional tone." (Gruda, 2024[9])

Machine-assisted editing is especially useful for non-native English speakers, due to its potential to improve flow, grammar and tone. In a poll by the European Research Council (ERC), 75% of more than 1 000 ERC grant recipients felt that generative AI would reduce language barriers in research by 2030 (Prillaman, 2024[10]).

A *Nature* poll (Kwon, 2025[11]) surveyed more than 5 000 researchers worldwide (with China underrepresented) in March 2025. More than 90% of respondents said they believe it is acceptable to use generative AI to edit or translate one's research paper. When it comes to generating text with AI – for instance, writing all or part of a paper – a majority (65%) think it is ethically acceptable, but about one-third are against it. The most popular use was editing a research paper, but only around 28% said they had done this. That number dropped to about 8% for writing a first draft, making summaries of other articles for use in one's own paper, translating a paper, and supporting peer review. While 42% of PhD students report using AI for editing purposes, the percentage drops to 22% for senior researchers.

Figure 13.3. Uses of AI by researchers, 2025

Which, if any, of these represent use cases or solutions that are similar to anything you are already doing and/or have already tried with AI in the past?

Use case	Percentage
Writing assistance	57%
Detect errors/bias in own writing	47%
Data collection and processing	35%
Check own work for unintended plagiarism	35%
Creating plain language summaries of article findigs	32%
Reviewing large amounts of published and preprint studies	30%
Populating citations	28%
Automated processing of unstructured data	18%
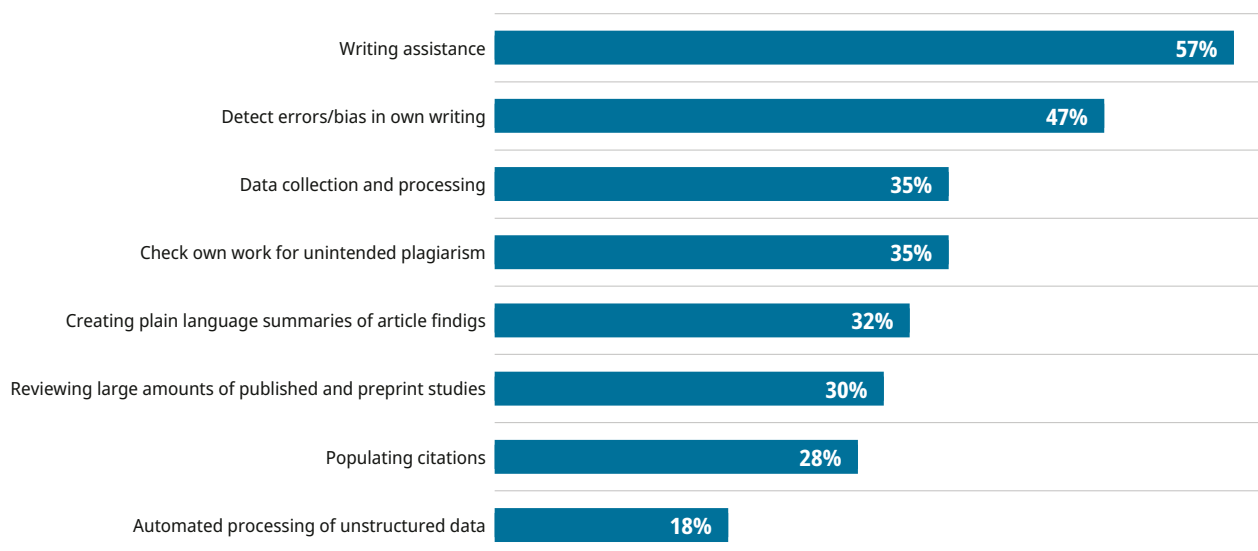
Note: Respondents were asked the following question: "Which, if any, of these represent use cases or solutions that are similar to anything you are already doing and/or have already tried with AI in the past?"

Source: Naddaf, M. (2025[6]), "How Are Researchers Using AI: Survey Reveals Pros and Cons to Science", Nature, 04 Feb 2025, https://doi.org/10.1038/d41586-025-00343-5.

Work by Kobak et al. (2025[12]) finds that one in seven biomedical abstracts published in 2024 (among 1.5 million papers indexed in PubMed) was written with AI assistance. They detect such abstracts by identifying "excess words," i.e. words whose frequency has surged since the rise of LLMs but that have no functional role (e.g. "delve," "unparalleled"; in total, there are 454 excess words).

Support from AI in writing can boost researchers' productivity for certain non-core tasks, like polishing style or handling administrative duties, freeing up time for more conceptual work (Gruda, 2024[9]).

Specialised models (e.g. the Black Spatula Project and YesNoError) are used to spot errors in research papers, including factual mistakes, calculation errors, methodological flaws and referencing issues.

> *The systems first extract information, including tables and images, from the papers. They then craft a prompt, which tells a 'reasoning' model — a specialised type of LLM — what it is looking at and what kinds of errors to hunt for. The model might analyse a paper multiple times, either scanning for different types of errors each time or cross-checking results. However, the rate of false positives — instances in which the AI claims an error where there is none — is a major hurdle (10% on average according to some tests; for example, the model may state that a figure referred to in the text does not appear in the paper when it actually does)* (Gibney, 2025[13]).

Programming and data analysis

As science is getting ever more quantitative, programming and analysing data are the main tasks of many researchers, especially PhDs, across all disciplines (including, increasingly, the humanities). This is certainly the case in education research, where the share of quantitative research has increased over the past decades, even though it remains a minority of education research (Vincent-Lancrin and Jacotin, 2023[1]). These tasks require specific skills in complex techniques and can consume a lot of time (e.g. for "debugging", i.e. tracking mistakes in computer code), while exposing the researcher to significant risks of errors. Special tools, based notably on GenAI, have been developed to alleviate these burdens.

Code editors are tools that aim to make it easier for researchers to use coding to organise data, create analytical sequences, generate descriptive statistics or visualisation. Such tools are now widespread, having overtaken GitHub and Stack (a community site) for troubleshooting. These tools allow researchers to save a lot of time, generate higher

quality outputs and allocate time to more substantive matters. Rather than spending hours waiting for answers from correspondents, users can simply highlight a section of code and ask a GenAI chatbot to fix it (Heidt, 2025[7]).

There are also more sophisticated AI models that can do extensive analysis of large tables of numbers and generate output like predictions (imputing), error detection, etc., hence avoiding the need for the researcher to do the programming by themself. For instance, TabPFN is a "tabular machine learning" model that infers outcomes from tables of any sort of data. It can take a user's dataset and immediately make inferences about new data points (McElfresh, 2025[14]).

GenAI models have also a strong capability of processing "unstructured data", like texts and images, that can thereby be quantified and subjected to powerful statistical treatment: this is clearly of specific interest for humanities and educational research.

Modelling complex phenomena

While LLMs process words, models based on the same techniques (notably so-called "transformer" architecture) can be trained with other types of data: chemical formulae, mathematical concepts, astronomical pictures, DNA, etc. These models can also mix different types of data as inputs ("multi-modal"), or they can take one type of data as input and generate a different type of data as output. This diversification in data types allows for the application of GenAI models to a broad variety of problems, in a broad variety of disciplines and contexts.

These models are commonly applied to so-called "closed-world problems", where the fundamental laws are known, but drawing out predictions is computationally difficult, because the parameters and variables are too numerous, or the relations are complex and non-linear. Examples abound in biochemistry, material sciences or weather forecasting. This allows for the combination of fundamental, established knowledge with an algorithm's superior capacity to find meaningful correlations in data. These models are statistical in nature and are thus trained on vast amounts of data, which restricts the cases where they may apply (not all domains offer sufficient amounts of data).

These models allow researchers to gain time and reduce the cost of research. "Our goal," says a biologist, "is to create computational tools so that cell biology goes from being 90% experimental and 10% computational to the other way around." This comment was made in regard to a project using AI to create a "virtual cell" (Callaway, 2025[15]).

While education systems generate large amounts of data, privacy and ethical concerns has made their widespread use for education research complex. Still, some of these techniques could increasingly apply to the analysis of large data sets and are already in use to some extent. For example, Pardos and Borcher (2026[16]) use these AI analysis and visualisation tools to show the similarity between higher education courses based on student enrolment history. Moreover, education research increasingly builds on neuroscience, cognitive science and one can imagine that learning science will benefit from advances in the study of the brain from a chemistry or biological perspective as well. For example, advanced AI techniques may help to better understand the clinical and socio-genetic dimensions of learning and education performance (Isungset et al., 2022[17]; Morris et al., 2022[18]).

Chemistry and biology

Chemistry and biology are leading disciplines where GenAI has been applied, due notably to the availability of a lot of data and to the combinatorial nature of mechanisms at work. The task of most models is to relate some property (therapeutic or physical) to the composition of a compound. Hence, a model can either predict the properties of given compounds, or, alternatively, it can predict the composition of compounds that display given properties. Some models can also do retrosynthesis, i.e. predict the sequence or network of chemical reactions that allow for the production of a particular compound with given ingredients (reactants).

These properties and reactions obey known physical and biological laws, but the number of components, the non-linearity of many mechanisms involved and the sensitivity of aggregate properties to minor modifications make it difficult, or even impossible, to analytically solve most cases. In the case of biology, the compounds are extremely complex (proteins can be made of several thousand molecules). Most models mix data analysis with knowledge of the basic rules of the domain, so that the generated items comply with the known laws of the domain.

There is a strong analogy between chemistry or biology on the one hand, and language on the other hand, as both are compositional: they are made of elementary components (words or molecules) that combine to produce emergent properties (meaning or physical characteristics). Hence, the techniques used to train LLMs have been directly transferred to these domains. Certain researchers even directly use LLMs to conduct chemical analysis, although the training base of LLMs in the domain is not as large as specialised models. They note: "Our results show that LLMs can accurately reason about chemical entities in both local and global terms, analysing single reactions but also whole synthetic routes, and that such capabilities can be exploited through search algorithms for solving chemical problems in more flexible terms." (Bran, 2025[19]).

Some models are made of several interconnected modules handling different types of data, and leverage the synergies between these types of data. Certain models relate natural language with chemical or biological data, which lets users query the model about the chemical composition that some particular property would possess, explained in natural language.

Box 13.1 provides examples of specialised scientific AI models.

Box 13.1. Examples of specialised models

ESM3

ESM3 is a generative model of the three fundamental properties of proteins: sequence, structure and function, through language modelling. ESM3 can be prompted with instructions from each of these sorts of inputs. This allows prompts to be specified at multiple levels of abstraction, from atomic level structure to high-level keywords describing the function and fold topology (Hayes, 2025[20]).

AlphaFold

AlphaFold is a family of models produced by Google DeepMind, aimed at predicting the 3D shape of proteins, and for which its main authors were awarded the Nobel Prize in chemistry in 2024.

AlphaFold has had a broad and deep impact on structural biology (Saplakoglu, 2024[21]). All researchers who use proteins now use AlphaFold2 or an alternative, so that they can reorient their time and efforts to other tasks. As "simple proteins" have been solved, researchers are turning to new matters: complex proteins (e.g. those whose shape is context-dependent or those that are composites) and RNA, giving rise to AlphaFold3 and other models. The impact of AlphaFold on downstream research (as proteins' shape is used as an input to other research) has been enormous: reviving structure-based drugs discovery; accelerating the creation of hypotheses in biology; and creating new avenues to understand complex interactions occurring within cells.

AlphaFold can still present risks for researchers. Errors occur, mainly in complex cases, sometimes leading to incorrect assumptions if unchecked.

Mathematics

DeepMind's AlphaProof and AlphaGeometry 2 (2024[22]) solved four of six International Mathematical Olympiad problems, reaching silver-medallist level. These systems combine formal language training with neuro-symbolic reasoning, reducing "hallucinations" and enabling rigorous verification (Castelvecchi, 2024[23]). Progress is impressive, though still short of tackling research-level mathematics.

Astrophysics

Krenn et al. (2025[24]) developed Urania, an AI algorithm for designing gravitational wave detectors. The system produced numerous innovative designs, some improving sensitivity more than tenfold; such designs reimagine known techniques, opening new possibilities for astrophysics.

AI simulating human agents

LLMs can be used to simulate human participants in empirical studies, for example, producing synthetic interviews, interactions between actors or specific behaviour in particular situations. There is active research on simulating human behaviour, with models being especially trained on psychological material (e.g. Binz, 2025[25]) and integrating knowledge from the cognitive sciences. AI models trained on human behavioural data can serve as test benches for simulating human decisions regarding various contexts including educational ones. They could play a similar role as organoïds (self-assembled constructs that mimic certain properties of in vivo organs) play in medical research. Such models accelerate studies and reduce their cost.

There are still limitations to the potential of this approach though, as the capacity of LLMs to simulate the diversity of human behaviours in complex situations is still limited. For instance, in a study about factory working conditions, a

worker on the floor and a manager would likely have different responses about a variety of aspects related to the work and workplace. However, an LLM participant's generated responses might combine these two perspectives into one answer, conflating attitudes in ways that are not reflective of reality (Kapania et al., 2025[26]).

Should this use of AI become fruitful, it could have a strong impact on education research, notably for the generation of survey answers, which, according to survey implementers, have become increasingly difficult to collect. For example, in the production of standardised assessments, (Liu, 2025[27]) show that multi-agent AI models bringing together ensembles of LLMs that can serve as "synthetic respondents" producing response distributions with psychometric properties closely aligned to those of college students. Pardos and Borchers (2026[16]) argue that LLM-based calibration can complement limited student response data, reducing costs and accelerating item validation cycles. While human responses remain essential (not least because they are used to generate simulated ones), AI-generated responses could augment them and, as is the case of answers to test items, expand their variance while remaining aligned with them. While it will take time to assess when simulated answers add value without distorting human responses, this is a line of AI impact that will be particularly useful for education research if it becomes successful.

Managing scientific knowledge

"We are dwarfs standing on the shoulders of giants," Bernard de Chartres famously said, characterising the cumulative dynamics of knowledge: new discoveries are primarily elaborations and combinations of past ones. Access to and mastery of existing knowledge is key for researchers to build on that knowledge and make new discoveries. Hypothesis generation is key to the research process and making new discoveries; it is closely tied with the existing knowledge on which it relies, but it also involves distinct mechanisms that will be examined in the next section.

With the mounting number of scientific publications (articles, databanks, images, computer programmes, etc.), it has become increasingly difficult for researchers to keep pace with advances in their own field, despite increasing specialisation within scientific domains. Hence a new challenge for researchers: improving on knowledge with which one lacks familiarity. AI has given rise to tools that can support researchers in these tasks, such as generalist LLMs or specialised models (e.g. Elicit, Consensus, Clarivate, PaperQA2, BioloGPT) (You, 2024[28]). These tools can conduct knowledge management operations like searches in the literature, summaries and literature reviews, which we will examine below. These models work as well for education research although they are not fine-tuned for this domain.

Search engines

The researcher enters a particular research question in the model (for example, "is virus X responsible for disease Y?"), and the model responds with a list of publications relating to the query, and for each publication, a summary of its results relating to the question. Some models offer a synthetic (consensus) view of the literature, with lists of publications agreeing or disagreeing with the consensus and the corresponding arguments. Some tools can generate a graphic picture of the concerned research landscape, with citations-based relations between publications (who cites whom, who is co-cited with whom, etc.) (Kudiabor, 2024[29]).

Compared to LLMs, specialised models aim to offer superior reliability, as they make use only of scientific publications, avoiding blogs and other sources of lesser scientific reputation. Certain tools can also offer other products beyond the aforementioned search results, like literature graphs (extracting the main concepts or results of a domain and relating them to each other in a knowledge graph). Certain platforms have a "Chat with PDF" function, which allows the user to upload a paper and ask questions about its content (Heidt, 2025[7]).

Producing summaries

AI models can produce summaries of publications on request, which gives researchers a rapid overview of a set of papers of interest, gaining time in reading and allowing focus on the most relevant papers. However, the quality of such summaries can be low at times. Peters and Chin-Yee (2025[30]) compared the AI summaries to human summaries that some journals provide for 4 900 examples in medicine and science overall. They found that all AI models tend to over-generalise the results presented in the papers, as they often omit important details that restrict the domain of validity of the results and leave out relevant nuances. They might, for example, just state that a drug is effective for treating a certain condition, without specifying in which dose or for which group of patients. This reflects the difficulty for AI models to fully recognise the importance of "details" that intelligent human readers find significant. The same issue would apply to education where results may be more or less relevant depending on country, socio-economic background, sex, etc.

Literature reviews

GenAI models can provide structured summaries of the literature relating to particular questions (Skarlinski et al., 2024[31]). These reviews are useful to researchers for getting a broad view of a question while saving time and making sure that they do not miss the most important relevant publications.

Some models offer "systematic" reviews, which include granular information on the methods and results of each paper in a standardised manner; this is necessary if a researcher wishes to reproduce an experiment or perform a meta-analysis. Some researchers are sceptical about the quality of AI-generated systematic reviews (Pearson, 2024[32]), as AI models tend to skip specific but important information like the precise dose of a drug, as mentioned above. This would be the same for the precise pedagogical context of, say, studies about the impact of project-based learning or lecturing or the use of technology in education.

More generally, literature reviews conducted with AI have certain limitations. First, many models can access only the abstracts of all publications and the full text of open access publications. Access to a significant part of the scientific literature is restricted (although publishers' tools can, of course, access their own publications), so that many important research findings, and most notably methods, are often skipped in AI-generated reviews.

Second, GenAI literature review tools sometimes struggle to identify the most relevant papers in a field, and to identify topical versus outdated literature, and can first list literature with ideas that used to dominate a domain but have now become outdated.

Such systems could still be used to update human-authored literature reviews rather than generate new ones. Reviews cannot feasibly be updated by humans very frequently and AI could provide for this, even though authoritative reviews may still need human involvement.

Research reports

Most recent models, like OpenAI Deep Research or Gemini Deep Research, can provide "research reports", which go beyond a literature review as they provide broader background and context and identify pending questions (Heidt, 2025[7]). The user can enter a query, together with their own data (articles, etc.), and the model returns a full report, including text, figures and corresponding bibliographic references. These models mimic how a person would approach a research question. This is especially interesting when exploring a domain with which the user is not familiar: it helps access general knowledge in a clear language (Jones, 2025[33]).

One specialised model, PaperQA2, writes Wikipedia-style summaries of scientific topics that are cited and significantly more accurate than existing, human-written Wikipedia articles. It can identify contradictions within the scientific literature, a task that is challenging for humans (Skarlinsky et al., 2024[31]).

Certain models generate draft research reports in which they identify gaps in knowledge, getting close to suggesting further possible research topics. The quality of these reports is debated (Jones, 2025[33]): they often include incorrect (or invented) citations, they fail to distinguish authoritative information from simple suppositions, and they do not convey uncertainty accurately.

Hypothesis generation and ideation

Hypothesis generation is a defining activity for a researcher. It consists of generating ideas from the literature or from data which are altogether novel, plausible and testable. Whereas a literature review is about what is known, hypothesis generation is about jumping into the unknown: identifying possible responses to questions that are not answered by the literature, while keeping consistent with established knowledge. This has been, until recently, a preserve of humans. Now AI can also do this. It usually involves three steps: generation of the hypothesis, evaluation/validation (or rejection), and improvement/refinement.

In the case of education research, these techniques could help to explain learning trajectories or some puzzling aspects of education outcomes. AI techniques could combine multiple and remote sources of information to generate original hypotheses. For example, one could imagine AI systems generating hypotheses based on international or national datasets to make hypotheses on the factors explaining the increase or decrease of student outcomes, exploiting the big size of these data sets. But it could also connect these results to other possible sources and point education researchers to possible explanations that are not immediately visible in their data sources (e.g. learning outcomes might increase due to the recent availability or social services that lead to less absenteeism of students or better mental health as their parents get better support).

Extracting hypotheses from research literature

Most models draw hypotheses from the literature, but some models can also do it directly from the data they are offered to analyse. As compared to humans, AI models have the advantage of a broader knowledge of the literature: not only in the concerned discipline, but possibly in others, accessing more diverse sources (assuming that the model has been trained or can access this knowledge, which may be problematic, as illustrated by the limits of automatic literature reviews). However, AI models are confronted with specific difficulties for automatically drawing hypotheses

from the literature: 1) the source texts might not make clear what the problems and corresponding hypotheses are; 2) the link between a problem and hypothesis as stated in the literature might not be straightforward; 3) the novelty or feasibility of an hypothesis can be difficult to evaluate, but they need to be measured and ranked; 4) initially designed hypotheses usually need to be improved so as to strengthen their novelty or feasibility, requiring further operations after the extraction from the literature.

One simple way for a researcher to interact with a model is through brainstorming, with a prompt like "give me ten ideas of mechanisms that could explain how A influences B". The researcher can also challenge the model, by submitting their own hypothesis and asking the LLM for counterarguments or alternative hypotheses. This simple procedure allows for initial suggestions, but for working out structured and plausible hypotheses, a more articulated approach is required, using specialised models.

One approach is to insert highly structured data into a prompt so as to strictly constrain the model's response. For instance, a researcher who investigates how microplastics are transported through soil and into groundwater could use a visualisation tool, Research Rabbit. The tool takes a single "seed paper" and generate an interconnected web of research linked by topic, author, methodology or other key features. By inserting its results into an LLM, "it's possible to query the body of work for hidden links or new ideas" (Heidt, 2025[7]).

Regarding the difficulty for LLMs to link problems and hypotheses from reading a corpus, one solution is to fine-tune an existing LLM so that it can better identify problems and hypotheses in papers. O'Neil et al. (2025[34]) have assembled a database of 5 500 scientific hypotheses (HypoGen), with which they trained an existing LLM. These data are structured in a way that makes it clear what the problem is, what the hypothesis is, and what the chain of reasoning from the problem to the hypothesis is.

Extracting hypotheses from data with AI models is made difficult by the lack of explainability of AI. For instance, the correlation between certain events might be difficult to attribute to particular features of the concerned events. The model will observe that phenomenon A is linked with B, but it could not say whether this is due to feature C or D of these phenomena. The model can see patterns in the data that are not visible to humans, and it might be difficult for the model to translate them in words, as hypotheses that can be understood by humans. Ludwig and Mullainathan (2023[35]) propose a procedure that enables the expression of correlations found by AI in the data in words, so that they can be explicated to humans and tested. For this purpose, they use counterfactuals: they generate synthetic data that exaggerate the correlations found in the initial data, to a point where the concerned pattern becomes visible to humans and can be interpreted.

Improving AI-generated hypotheses

Inferring hypotheses from the literature might not be enough, or fully satisfactory, as "raw" hypotheses might be insufficiently novel or articulate, too similar to their sources, weakly plausible (not fully coherent with the evidence), or difficult to test experimentally. Thus, a process of refinement of the ideas extracted from the literature is warranted. This is one of the most difficult challenges for AI in science, as it requires both imagination and reasoning capacities: the capacity to improve an idea while keeping its core; to infer logically; to assess the proximity of an idea with the "real world", etc. A lot of developments are occurring in this domain. The main techniques include: multi-step reasoning ("chain of thought", requesting the machine to make explicit the steps it is following in its reasoning); reinforcement learning (training the model so that it strengthens its successful features and weakens others); evolutionary computation and multi-agents systems (see below). Models developed since 2024 include one or more of these techniques.

Evolutionary computation is a technique inspired by the mutations and natural selection in Darwinian evolution. It begins with a review of the literature, from which it extracts an initial list of hypotheses. It applies small, random changes to an algorithm and selects the ones that improve the model's efficiency. To do so, the model conducts its own "experiments" by running the algorithms and measuring how well they perform. Afterward, the model produces and evaluates a paper. After "augmenting the literature" this way, the algorithm can then start the cycle again, now building on its own results (Castelvecchi, 2024[23]).

Agentic AI is being applied also to science. An agent is an autonomous system that can pilot various tools towards a given objective. Multi-agent systems are comprised of several agents with specific objectives and specialised skills; each one pilots an AI model (e.g. an LLM) and interacts closely with others under the supervision of a "lead agent", who acts like the conductor of an orchestra. Some models also integrate reasoning capacities like the aforementioned "chain of thought". A multi-agent model aims to function like a group of researchers. Some researchers are specialised in a particular discipline; some researchers play a particular role, of making proposals, or challenging others' proposals or combining them; at each step they are assigned specific tasks by the lead, and they work with their respective tools to implement them; they meet and confront their respective findings, with open discussions whose conclusions are included in a report shared with human researchers. The whole process starts with a prompt that includes a description

of the problem and contextual information, which is submitted to the model. The lead and the agents then design a research plan, possibly with a sequence of sub-questions, a set of parallel tasks, a list of skills required from agents, etc. Then an iterative process can take place, in which each agent accomplishes the tasks it has been charged with, and reports to the lead and to other agents; the lead synthesises the findings at each stage and monitors the advancement of the whole process. At the end, the model can draft a research report (Biever, 2025[36]).

In education, multi-agent models with GenAI agents are for example used to develop assessment items. They may hold promise for educational research, which is often multi- or inter-disciplinary or address broad-based socio-technical issues (such as the adoption and use of AI in education). Education researchers may use such models to ensure different types of expertise, information sources and constraints are brought together, for example to generate new ideas on children's school rhythm, which involves expertise in children's biological and psychological development and needs, learning science and pedagogy, parental work schedules, etc. (Figure 13.4). One can also imagine some fruitful uses to generate ideas or improve usual hypotheses for education policy research, with some of the AI agents playing the role of different education stakeholders and providing ideas on addressing some education policy issues such as the provision of equal opportunities. The multi-agent models could then make new suggestions of educational interventions or of policies in this simulated environment.

Figure 13.4. A possible interdisciplinary Human-AI collaborative educational research model
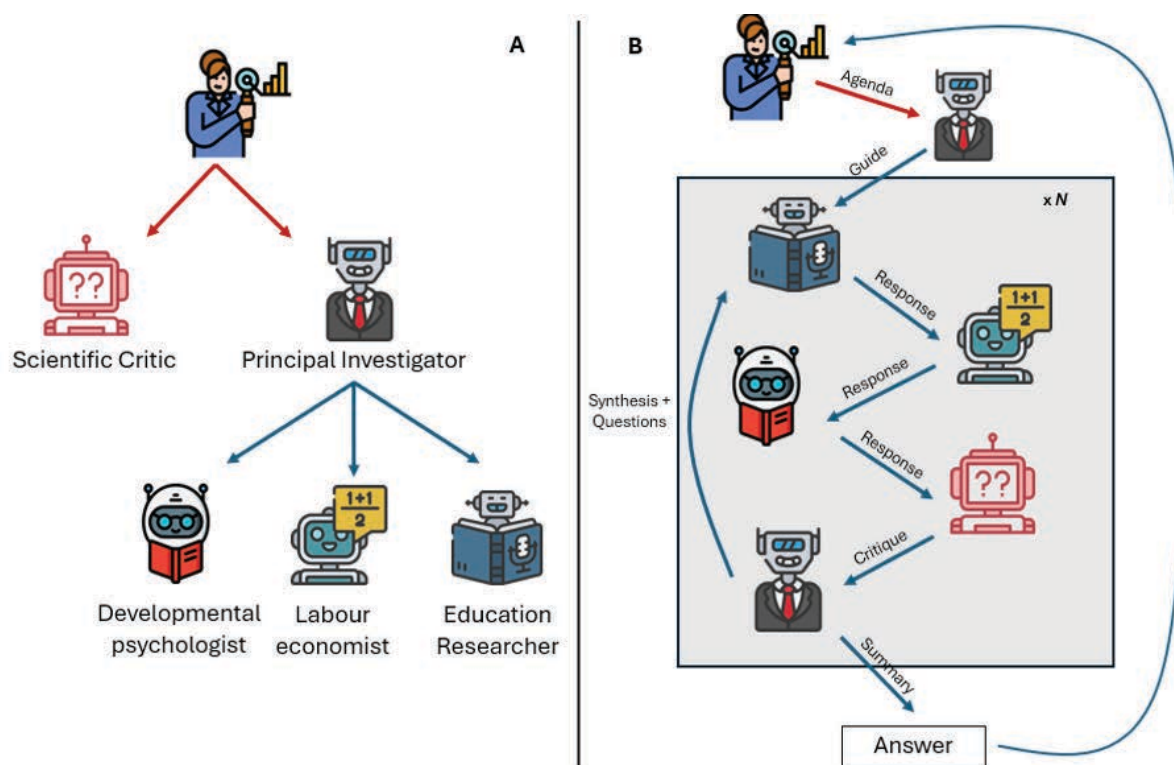
Note: This imaginary case applies ideas of Box 13.2 to education research. In panel A, the human researcher provides a brief, for example on students' school rhythm, and requests a GenAI agent (principal investigator) to set us an interdisciplinary research team, and a Critic to review what the interdisciplinary produces. In that case, the AI agent chooses the expertise of the other AI scientists (developmental psychology, education research, labour economics). They may generate research literature reviews, retrieve the old literature and generate new hypotheses. Panel B illustrates how the human researcher can make progress by iteratively asking specific questions for the AI agents to discuss collectively: once the human sets the agenda of their meetings, like in the Virtual Lab, the GenAI scientific agents share their expertise, which is critiqued by the Critic, allowing for iterative refinement before an answer is provided to the human researcher – who can then ask follow up research questions.

Source: Authors' elaboration.

Box 13.2 presents some examples of (specialised) scientific multi-GenAI agent models.

Box 13.2. Examples of multi-agent models based on GenAI

The AI co-scientist

When provided with a research goal in natural language, the system searches and analyses relevant literature, summarising and integrating existing work to generate new research hypotheses and experimental protocols for later validation. The co-scientist supports its recommendations by citing key sources and explaining the logic behind its suggestions (Gottweis, 2025[37]).

Scimon

Given a background problem description, models first dynamically retrieve inspirations from past literature in the form of related problems and their solutions, along with contexts from a scientific knowledge graph. These retrieved inspirations serve to ground the generated ideas in existing literature. Generated ideas at this stage are not necessarily new. Models have the the capability to iteratively boost the novelty of their generated ideas. Given an idea generated by the LLM model compares the idea with existing research in the literature; if it finds strongly overlapping research, the model is tasked with updating its idea to be more novel relative to prior work (much like a good researcher would do). This model also introduces an in-context contrastive model which encourages novelty with respect to background context (Wang et al., 2024[38]).

Virtual Lab

Swanson et al. (2025[39]) developed the Virtual Lab, a multi-agent model that tasks different AI agents to accomplish different tasks under the supervision of human researchers to perform inter-disciplinary science and investigate broad, complex research questions First, the human researcher defines two core AI agents: the Principal Investigator (PI) and the Scientific Critic. Based on a short project description, the PI creates additional AI scientist agents, assigning them specific expertise, goals and roles. The human-AI collaboration then happens through two types of workflow: team meetings and "bilateral" meetings. The authors describe the team meeting as follows: first, the human researcher sets an agenda for the meeting, then the PI opens with initial thoughts and guiding questions, after which each scientist agent contributes in turn. The Scientific Critic reviews their input, and the PI synthesises the discussion, posing follow-up questions. After several discussion rounds, the PI generates a final summary and conclusion for the human researcher. In the case of "bilateral" meetings, a designated scientist agent responds to the agenda set by the human researchers, receives critiques from the Scientific Critic, and iteratively improves its answer. After multiple rounds, the AI scientist delivers a refined final response.

DeepMind's AlphaEvolve

AlphaEvolve is based on the (Google) LLM. Each task begins with the user providing a question, evaluation criteria, and an initial solution. The LLM then generates hundreds or even thousands of possible variations. An evaluator algorithm scores these alternatives against the defined metrics, and the best-performing ones guide the LLM to propose new ideas. Through this iterative process, the system gradually develops a population of increasingly effective algorithms (Gibney, 2025[13]).

How effective are these models?

How effective are GenAI models in generating, refining and evaluating scientific hypotheses? The evidence is still scarce, as testing such models is complex and costly. The most effective tests are implemented by performing actual research and examining the model's achievement. Diagnoses are diverse due to the diversity of models, the diversity of research questions and the diversity of testing methods.

On broad research questions, the models seem able to provide useful suggestions that point to potentially fruitful research directions. This is due to their very good access to the literature and their ability to process it in a highly structured way. According to Anthropic (2025[40]), their "internal evaluations show that multi-agent research systems excel especially for breadth-first queries that involve pursuing multiple independent directions simultaneously."

When it comes to more specific research questions, the evidence is mixed. There have been certain impressive achievements, with models able to identify and describe hypotheses that were then successfully tested by researchers (see Box 13.3), but also less successful cases. Successful cases come with significant involvement of humans in the process.

Certain studies also point to the tendency of some models to propose solutions which are plausible but not really novel, including some that have been already explored and abandoned in the past. Wang et al. (2024[38]) conducted extensive evaluation experiments using human annotators with domain expertise to assess the proposals of a multi-agent model called Scimon (Box 13.2). They found that "ideas still fall far behind scientific papers in terms of novelty, depth and utility – raising fundamental challenges toward building models that generate scientific ideas."

Regarding the AI Co-Scientist (Box 13.2), specialised in research on machine learning: "the authors admit that the papers that AI Scientist has produced contained only incremental developments. Some other people were scathing in their comments on social media. 'As an editor of a journal, I would likely desk-reject them. As a reviewer, I would reject them,' said one commenter on the online forum Hacker News." (Castelvecchi, 2024[23])

In mathematics, there is also contradictory evidence. Some tests point to the fact that certain of the claimed achievements of some models in solving Olympiad silver or gold medal level problems were due to "data leakage" (Olympiad is a global competition in maths): solutions had been published before and were accessed by the models (Petrov et al., 2025[41]). On the other hand, very rigorous testing by professional mathematicians showed that o4.mini, an OpenAI reasoning model, can solve most of the PhD-level problems they submitted, demonstrating extremely powerful reasoning capacities. However, there is consensus that current models are not up to the level of mathematical research, although they are getting closer (Chiou, 2025[42]).

Two further caveats need to be mentioned. First, not all negative results of testing are necessarily published, and limitations of the models could be underestimated; second, it is difficult to estimate the quantity of human involvement in the models' work, but it could be sometimes important, and the role of the models could then be over-estimated. However, these models are still in a very early stage of development and much progress is to be expected in the near future.

Box 13.3. Cases of GenAI discoveries

Co-scientist

The AI Co-scientist was tested on a question that took years to resolve: how cf-PICIs (a defence mechanism against phages) spread across bacterial species. It generated and tested several hypotheses. Its top-ranked hypothesis matched the experimentally confirmed mechanism. Other hypotheses opened new research avenues, showing AI's potential as a creative engine for scientific discovery. Generated hypotheses were novel, logically coherent and experimentally testable, highlighting the system's ability to accelerate discovery and build transdisciplinary bridges (Penadés et al., 2025[43]).

The Virtual Lab

The Virtual Lab tackled the design of nanobodies binding to new SARS-CoV-2 variants. It combined several models – ESM, AlphaFold-Multimer, and Rosetta25 – to mutate nanobodies targeting the spike protein. Of 92 designed nanobodies, over 90% were expressed and soluble, with two showing unique binding to recent variants JN.1 and KP.3. This demonstrates the potential of AI-human collaboration to deliver complex, validated scientific results (Swanson et al., 2025[39]).

AstroAgents: AI to track extra-terrestrial life

AstroAgents is a multi-agent system with eight AI agents generating hypotheses in astrobiology. Using LLMs (Claude Sonnet 3.5, Gemini 2.0 Flash), it analysed meteorite and soil mass-spectrometry data, producing 101 hypotheses from Gemini and 48 from Claude. Gemini's ideas were more novel but error-prone; Claude's clearer but less original. Overall, the system produced plausible, pattern-spotting insights beyond human capability, useful for analysing samples such as those planned for return from Mars (Biever, 2025[36]).

Box 13.4. Examples of AI research assistants

ResearchAgent

The system autonomously identifies new research problems, proposes methods and designs experiments, refining them through feedback from LLM-based reviewing agents. Beginning with a core scientific paper, ResearchAgent expands its scope by linking related publications through an academic graph and incorporating concepts extracted from a knowledge base built across many papers. Emulating peer review, it uses multiple LLM Reviewing Agents that provide critiques and iterative revisions. These agents are aligned with human preferences, with their evaluation criteria derived from real human judgments through LLM prompting.

AgentRxiv

The system is built around specialised agents that collaborate across different phases of research. During the experimentation phase, PhD and Postdoc agents design the research plan, while machine learning and software engineer agents handle data preparation and coding tasks. An automated module then generates, tests and iteratively improves machine learning code. It also includes an LLM-driven repair function to correct errors during execution. In the report writing phase, Professor and PhD agents compile the results into a structured report. This process mirrors peer review, with iterative revisions and checkpoints that can be autonomous or human-guided.

Source: Baek et al. (2025[46]), *ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models*.

Automating science: Research assistants and robot labs

The AI systems examined above keep to the cognitive tasks of research: analysing and generating information. Other systems go one step further and aim at performing the whole range of tasks of a research assistant, notably the design of experiments (research assistants), or even the realisation of experiments (robot scientists).

AI research assistants

There has been a surge in the supply of AI research assistants since 2024 (see Box 13.4). They can be compared to the teaching assistants described by (Baker, 2026[44]), even though their functionalities and workings are different. The characteristics shared by AI research assistants are the following: 1) they perform all tasks expected from a research assistant: reviewing literature, generating hypotheses, designing experiments, drafting articles; 2) they are technically similar to the hypothesis generation models examined above (multi-agents, etc.); 3) they are very interactive, as their functioning involves frequent and important exchanges with a supervising human, who remains in close control of the research process; they are just "assistants" after all.

According to So (2025[45]), AI research assistants offer numerous benefits, including accelerated research timelines, 24/7 availability, personalised support, enhanced objectivity and improved accessibility for non-native English speakers. AI research assistants are evolving to support various collaboration models, from passive assistants to full research partners. Despite their impressive capabilities, AI research assistants face significant challenges, including generating inaccurate information, limitations in critical analysis, and ethical concerns around plagiarism and attribution.

Robot labs

The "robot scientist" marks the automation of the last step in the research cycle, the performance of experiments. The robot scientist works by connecting laboratory equipment to an AI system: the AI designs the experiments and controls the equipment so that it performs these experiments.

Why automate experiments? Here is one example. According to pioneer Ross King (2024[47]):

> *Studying eukaryotic system biology is a complicated task as even simple eukaryotic cells such as yeast have thousands of genes, proteins, and other small molecules, that interact in a complex spatial and temporal manner. The high complexity of the models means that their development and evaluation require the execution of millions of experiments based on a hypothesis. Only AI systems with automated labs have the capacity to plan, conduct, and monitor such a high number of experiments.*

In this case, the robot allows researchers to perform experiments that would be beyond human capabilities. An additional advantage of robot labs is that their experiments generate large quantities of high-quality, controlled data that can be used to train AI models.

One example in chemistry is CRESt (Ren et al., 2023[48]): users exchange with CRESt like with a colleague, in natural language. CRESt helps to craft and run experiments by retrieving and analysing data, turning equipment on and off, powering robotic arms, documenting findings and alerting scientists when something requiring their attention arises. CRESt-assisted researchers identified candidate alloys for fuel cells.

While important for many scientific fields, the automation of laboratory work seems less relevant for education research, where most of the experiments involve humans in a controlled or real-life environment. This may however support work about chemistry, biology or neuroscience that will help understand how human learns or where some impediments can come from.

Synthesis

Table 13.1 presents a summary of the possible roles of GenAI at the different steps of the research process, as well as their current achievements and limitations.

Table 13.1. Phases of the research process and capacities of available AI tools

Research task	GenAI achievements	GenAI limitations
Identifying a research problem	Identifying anomalies in the data; gaps in the literature.	The proposals are sometimes obvious or outdated (already solved).
Reviewing the literature	Literature reviews can be done by many AI models; they usually allow for a fairly broad and accurate view of a domain.	Hallucinations (referring to non-existent sources); difficulties differentiating between up-to-date or reliable results and outdated or dubious ones.
Formulating and evolving research hypotheses	Various approaches, from a brainstorming with a researcher to multi-agent and reasoning models simulating a team of researchers discussing and checking the proposed ideas.	Originality and feasibility of the proposed hypotheses are still limited but are improving with the latest models.
Curating, analysing and generating data	Foundation models can conduct highly sophisticated data analysis and can generate data.	Models are highly specialised and need training with large quantities of data; results are often difficult to interpret.
Designing and implementing experiments	Recent models have the ability to design experiments, and AI-piloted robots can implement them.	Still in early stages.
Inferring research conclusions and drafting a paper	Models can draw conclusions from results of experiments and analysis, relate results to the literature and draft a paper.	Despite successes in specific cases, these papers are often of limited quality, novelty and reliability.

Effects and challenges of GenAI in scientific research

GenAI displays a growing range of capacities increasingly used by scientists. It is therefore affecting core features of the scientific endeavour: creativity, reliability, publishing and academia-industry relations.

Impact on creativity

In scientific research, as in any other field, creativity means producing something novel and relevant. Empirical studies and researchers' own experiences illustrate that GenAI can influence scientific creativity either directly, through its own discoveries, or indirectly, by augmenting scientists' creative capacities. Moreover, it can also affect collective creativity.

Offloading scientific creativity to GenAI

Creativity in scientific research requires mastery of existing knowledge and the cognitive skills to move beyond it (Simon, 2001[49]). Current GenAI models excel at storing and combining vast amounts of information; they can identify patterns and generate plausible re-combinations (interpolations) of existing ideas using statistical techniques. However, according to most tests, they struggle to extrapolate new concepts beyond their training data. Even their capacity to combine "ideas" might concern words rather than ideas, due to their lack of genuine understanding (the "stochastic parrot" metaphor). The capacity of AI to reason can also be questioned: many tests show a limited capacity of AI to perform logical inference (Shojaee et al., 2025[50]), which is a necessary operation for articulating scientific ideas. As a result, GenAI's capacity for conducting radical scientific breakthroughs remains limited, while its capacity for incremental novelty requires the significant involvement of humans. In testimonies of researchers collected by Castelvecchi (2024[23]), one argues that current LLMs "cannot formulate novel and useful scientific directions beyond superficial combinations of buzzwords."

Yet, there are striking cases of GenAI models demonstrating true creativity. For instance, the model ESM3 has designed fluorescent proteins radically different from naturally occurring ones, variants that nature might have taken hundreds of millions of years to evolve (Hayes, 2025[20]). In an experiment conducted in the discipline of natural language processing, Claude 3.5 Sonnet outperformed human experts in generating interesting research questions, although its suggestions were often less feasible in practice (Chenglei Si, 2024[51]). In another example, an AI algorithm designed unconventional configurations for space wave detectors, some of which could dramatically increase the sensitivity of these instruments (Krenn, Drori and Adhikari, 2025[24]). Such cases demonstrate that GenAI can produce surprising and valuable outputs, although they remain rare and highly domain-specific.

One factor that can make GenAI creative is the immensity of the search space it navigates for potential solutions. AI models can identify and examine huge numbers of ideas, situated in extremely remote parts of the search space. Scientific discovery often involves chance, finding unexpected solutions by exploring many possibilities. Because AI models can test and combine so many ideas, they can stumble upon configurations that humans might never have considered yet. This combinatorial advantage is particularly effective in fields like chemistry, materials science and biology, where novelty often comes from exploring novel combinations. The fluorescent proteins generated by ESM3 illustrate how "interpolation" within an immense search space, that humans know only small parts of, can still yield significantly novel outcomes.

Several additional aspects are worth noting:

- Curiosity: Human creativity is driven by curiosity. Ongoing efforts aim to endow AI models with a form of "artificial curiosity" to encourage questioning established ideas, notably the ones it has found itself.

- Anomaly detection: Many discoveries begin with spotting inconsistencies between observations and expectations (predictions by a theoretical model). GenAI can excel at detecting simple anomalies in vast datasets, but lacks the deep reasoning and mere "representation of the world" required for complex theoretical contradictions.

- Hallucinations: Also known as "confabulations", these are (often plausible though wrong) statements or references invented by the AI model, often resulting from the accidental merger of two or more source pieces of information. They are not an example of true creativity as they lack relevance; they are disconnected from the original question and do not carry meaning.

Human-AI collaborative creativity ("augmented creativity")

Although GenAI cannot yet rival human creativity, it offers complementary strengths that can enhance human creative work. A well-designed division of labour between humans and AI might then result in greater creativity of a human-AI hybrid system. Researchers are experimenting with human-AI collaborations in various ways:

- Brainstorming: AI can stimulate new ideas by challenging human assumptions, despite its tendency to mirror users' inputs uncritically. Anderson, Shah and Kreminski (2024[52]) found that people who used an LLM generated more creative stories than those working alone.

- Productivity: Freeing a researcher's time (by taking over simpler tasks) so that they can focus on higher-level thinking.

- Balancing exploration and exploitation: GenAI's strength in re-combining known ideas (exploitation) may allow humans to concentrate on genuine leaps into the unknown (exploration) (Gans, 2025[53]).

- Search and screening (evolutionary computation): GenAI can scan vast combinatorial spaces and present promising options for human evaluation. For example, Si et al. (2024[54]) showed that LLMs can generate research plans, but human judgment remains essential to assess feasibility.

- Flexible collaboration modes: Tools like Agent Laboratory (an AI-based research assistant) offer both autonomous and co-pilot modes, where researchers can choose when to intervene and guide AI-generated work (Schmidgall et al., 2025[55]).

These mixed-initiative approaches underline a key insight that current GenAI is best seen as an augmentation tool that complements, rather than replaces, human scientific creativity.

Additionally, some researchers worry that reliance on AI might erode human creativity, as it assumes tasks that support human's creativity: if "writing is thinking" and if humans leave the writing to machines, then they will do less thinking as well. A recent study (Lee et al., 2025[56]) suggests that GenAI can reduce the perceived effort of critical thinking, fostering over-reliance on AI and diminishing independent involvement in problem-solving. However, these findings are based on studies conducted in laboratory conditions, related to one-off tasks and literary creation rather than real-world research contexts, so more robust evidence is needed.

Collective creativity and the risk of homogenisation

Even if GenAI enhances the output of individual researchers who use it, its widespread use could have a negative effect on overall scientific production; in particular, it could reduce the diversity in scientific inquiry. A large-scale study found that scientists who adopt AI tools publish significantly more papers and receive more citations than others, but AI also narrows the range of topics being explored collectively, as researchers using AI tend to concentrate on similar topics (Evans et al., 2024[3]). Rather than inspiring bold ventures into new fields, AI appears to accelerate work in established, data-rich areas, raising concerns about a "homogenisation effect."

Similar findings emerge from creative writing experiments. For instance, Doshi and Hauser (2024[57]) found that while generative AI ideas made stories more creative and engaging - especially for less creative writers - they also made the stories more similar to each other. Anderson (2024[52]) similarly observed that users produced less semantically distinct ideas with ChatGPT than with other collaborative support tools, and felt less responsible for the ideas they generated.

In sum, GenAI is unlikely to replace human creativity in science any time soon. The time has not come for an AI system to be awarded a Nobel prize (Kitano, 2016[58]), except to make a point about the creation potential of AI (as was the case for AlphaFold). GenAI might have the potential to reshape and boost human creativity in profound ways though. It can extend the combinatorial power of researchers, accelerate repetitive tasks, and foster serendipitous ("by chance") discoveries. At the same time, risks remain: over-reliance on AI could dampen human critical thinking, and collective creativity might suffer if the research community focuses too narrowly on questions where GenAI works best.

Impact on reliability and truthfulness

Veracity is another pillar of science. A scientific statement is accepted as such only after it has been validated according to certain socially vetted procedures and rules among the scientific community. GenAI has been challenged on this front, with a number of cases where it generated clearly inaccurate results.

Technical limitations

A principal problem is that AI models have a notion of "veracity" that is restricted to their training set, which is significantly narrower than that of the real world, especially for LLMs that are essentially trained on texts whose own veracity is often dubious.

A second problem is "hallucinations", or confabulations, which consist of confusing and conflating two or more pieces of information from the training set, resulting in incorrect, usually plausible though often nonsensical statements. This sometimes happens with references to the literature, where certain articles are simply invented by the AI (attributing one title to one author, as the two do exist but are not connected and the reference itself does not exist). Confabulation is generally attributed to technical factors: 1) compression of the data, a process that sometimes generates accidental

mixture of data in the decompression stage; notably, the machine will mix the relevant story with a similar but irrelevant one that it has memorised, and will issue a response based partly on the irrelevant story; 2) lack of "metacognition", a set of reasoning procedures that operate in the human brain and cross-check thoughts before they are expressed, making incoherent ideas less likely. It is also favoured by the sycophantic bias of most LLMs, which are trained to answer queries even in cases when they only find responses with low probability.

Reproducibility

Reproducibility is a pillar of scientific operations. To be accepted by the scientific community, results must (usually) be verifiable, and reproducible by others. One condition for reproducibility is full disclosure of the methods and data that led to the conclusion, meaning transparency and accessibility. From this perspective, GenAI models do not meet scientific criteria. First, the most popular models of GenAI are "black boxes", as neither their weights (the parameters that define a neural network) nor their training data are publicised. Thus, disentangling the contribution of the data and the contribution of various components of the model is difficult in any scientific result coming from such a model. This comes from the very nature of neural networks: knowledge is distributed, hence difficult to localise. As GenAI models have a random component at their core, some results might not be robust. In addition, access to the training data can be limited due to the proprietary nature of many GenAI models: one example is the "AI Structural Biology Consortium", a follow-up to AlphaFold-3, an ongoing project which makes use of data owned by pharmaceutical companies, which is secret and will remain secret (Callaway, 2025[15]). Currently, solutions for access include open weights (e.g. Llama) and open source (including access to training data). The importance of openness was demonstrated by AlphaFold2, as the disclosure of its code and data triggered a series of initiatives refining the tool (Saplakoglu, 2024[21]). Openness is essential to the cumulative progress at the core of science.

Two further issues might make AI-based research less reproducible. First is the random aspect of the working of models, which is that a model can generate different results for the same prompt for no substantive reason; second is model drift: models are regularly updated, and asking the same question to the same model at two different moments in time might bring different responses due to a change in the model's parameters.

But AI might also bring transparency to scientific research: AI systems record all that they do and can report all of their operations and corresponding outputs in more precise ways than humans. This makes it possible to reach higher traceability in research activities, and thus easier reproducibility.

This transparency might also help identify and publicise "negative research results" (e.g. drugs that did not pass clinical trials). Such results are important for follow-up research, indicating avenues to avoid and saving time and resources, but are currently seldom publicised by researchers as they are under-rewarded on the market (no one got a Nobel Prize for a negative result yet). AI itself would also greatly benefit from the availability of negative results, as they would significantly enrich the data sets used for training.

Impact on publishing and integrity

Some scientists support the view that GenAI will boost certain dimensions of their productivity. In responding to a Nature poll of 1 600 scientists, a majority noted that AI provides faster ways to process data, that it speeds up computations that were not previously feasible, and that it saves scientists time and money. An ERC survey had 85% of respondents who thought that generative AI could take on repetitive or labour-intensive tasks, such as literature reviews. 38% felt that generative AI would promote productivity in science, such as by helping researchers to write papers at a faster pace (Prillaman, 2024[10]).

To date, there has not been a direct study of the impact of GenAI on the productivity of researchers, but such studies have been performed in other professions performing some similar tasks as researchers. Noy (2023[59]) examined the productivity effects of ChatGPT on mid-level professional writing tasks. "In a preregistered online experiment, we assigned occupation-specific, incentivised writing tasks to 453 college-educated professionals and randomly exposed half of them to ChatGPT. Our results show that ChatGPT substantially raised productivity: The average time taken decreased by 40% and output quality rose by 18%. Inequality between workers decreased, and concern and excitement about AI temporarily rose. Workers exposed to ChatGPT during the experiment were 2 times as likely to report using it in their real job 2 weeks after the experiment and 1.6 times as likely 2 months after the experiment."

Overall, it can be expected that GenAI will: 1) allow scientists to do things that are not feasible otherwise (read a vast and variegated selection of literature, study complex systems, process big or unstructured data, generate data, etc.), and 2) save researchers time by taking over certain tasks (Schmidgall et al., 2025[55]). This double process of complementing and substituting humans should have a positive effect on productivity. But it might generate costs as well (financial cost, more data-related tasks, non-explainability, fake results, weakening of research on AI-poor issues, reduced diversity, etc.), so that the net effect could be lower than the raw effect.

Figure 13.5. Annual number of scientific publications
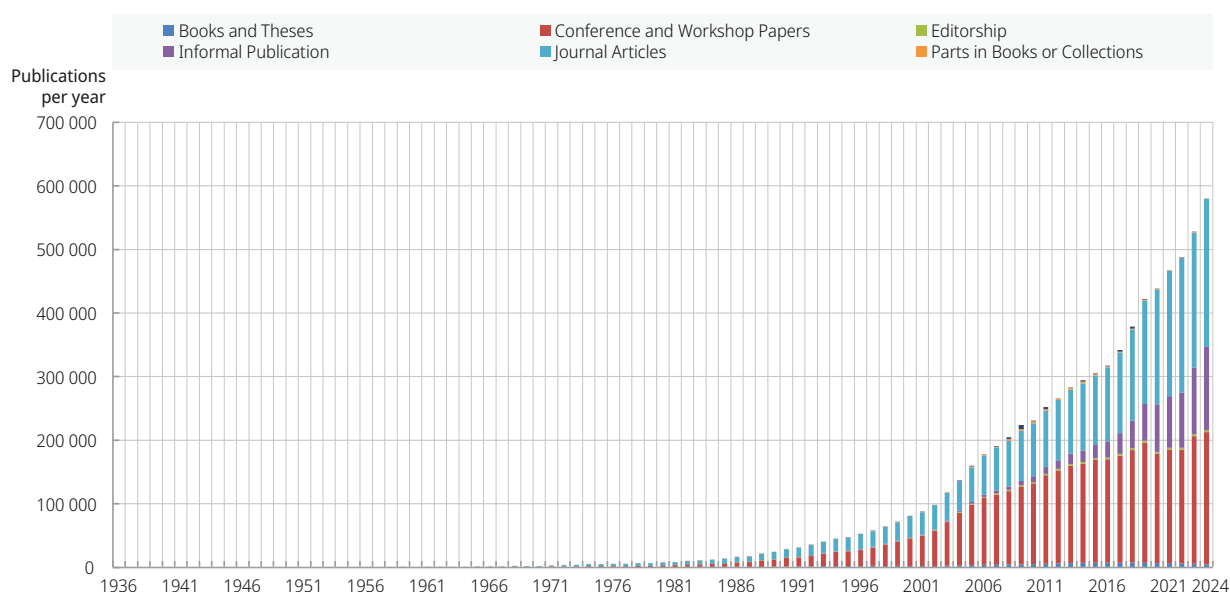
Source: Retrieved from DBLP Computer Science Bibliography at https://dblp.org/statistics/publicationsperyear.html

GenAI has made it much easier (less time- and effort-consuming) to draft papers (relative to other research tasks, except, perhaps, research grant applications). One would therefore expect that for a given level of research input, the number of papers would increase. That's exactly what is observed (Figure 13.5), and it raises an acute question to the scientific system regarding how to maintain the quality of published science in this abundance of new papers.

In this context, GenAI risks flooding the market with papers, creating more clutter than knowledge. "Paper mills" exploit low costs to mass-produce possibly poor research work. Liang et al. (2024[4]) finds that higher AI use in papers correlates with more frequent pre-prints, crowded fields, and shorter papers, i.e. quantity without guaranteed knowledge gain (although there might be some).

Initial reactions by publishers were restrictive: the International Conference on Machine Learning (ICML) (2023[60]) and Science banned AI-generated text. As use spread, publishers adopted nuanced policies. Most now distinguish between: 1) AI for editing (no disclosure needed); and 2) AI for content generation (requires disclosure). The International Association of Scientific, Technical and Medical Publishers (STM) (2023[61]) endorsed this; the scientific publisher Wiley (2025[5]) asks authors to specify usage; Science sometimes requires prompts; Nature asks for documentation but not for copy-editing (Kwon, 2025[11]).

Paper growth outpaces reviewer supply, stressing the capacity for peer review (Bergstrom and Bak-Coleman, 2025[62]). One possible response is automated reviewing. GenAI is being tested to assist or replace referees. Gruda (2024[9]) proposed structured prompts; Liang et al. (2024[4]) showed strong overlap between AI and human feedback, with many researchers finding it useful. Yet scepticism is strong: over 60% of scientists oppose full automation (Kwon, 2025[11]). Elsevier and the American Association for the Advancement of Science (AAAS) ban AI reviewing; Wiley and Springer Nature permit limited, disclosed use, forbidding confidential uploads.

Authorship is key to the traceability of scientific discoveries and to human researchers' careers. Currently, AI is not recognised as a co-author, though a fully AI-written paper ("The AI Scientist-v2") was accepted in 2025 (Sakana.ai, 2025[63]). Recognition of AI authorship would raise legal questions, since AI lacks rights and responsibilities.

The number of cases questioning research integrity has been rising, and many of them recently have involved GenAI. GenAI can create convincing simulated and synthetic data, useful for augmentation as mentioned above, but vulnerable to misuse. Simulated data can and have already fuelled dubious research, as shown in health datasets where post-2021 papers surged with questionable results (Suchak et al., 2025[64]).

Impact on business-academia connections

Developing core GenAI models requires massive resources beyond those available in academia or most governments, leading to concentration in business, which controls talent, data and computing power (Ahmed, Wahed and Thompson, 2023[65]). This strengthens trends already visible in non-generative AI.

This situation has several consequences:

- Large funding flows to basic research, boosting fields like cognitive sciences and mathematics.
- Much academic AI research depends on industry financing, giving business strong influence over agendas; business-related priorities such as advertising may displace societal goals like health research or research in education.
- Academic standards of openness clash with corporate secrecy, so widely-used models may lack reproducibility or transparency.

A new type of science?

GenAI is generating a transformation of science that non-generative AI, and even data-based research prior to the rise of AI initiated. Although most observers agree on this diagnosis, they emphasise different aspects of this transformation and have different views of the future. One view is that science as generated by GenAI will be too complicated for humans to understand; another view (which is not exclusive) is that science will need less and less theory and become closer to data.

Towards non-human science?

GenAI is one step on the way to the automation of science. If science continues its advance on this path, and current technical change makes this scenario plausible, it might become fully automated at some point: machines would not only perform research tasks that are beyond human capabilities, but also design these tasks and take the decision to execute them, after having worked out the underlying hypotheses and before drawing an interpretation of the results. This is the "Nobel prize awarded to an AI" scenario (Kitano, 2016[58]). One possible correlate of this scenario is the emergence of a new science that would be beyond the understanding of humans: using models, concepts and techniques that are too big or too complex for humans to grasp. This sort of science would be more powerful than the current one, freed from the limits of human cognition; it would allow the invention of more sophisticated technologies. At the same time, it would largely escape the grasp and control of humans, and it would not fulfil one of the traditional missions of science: to deepen human understanding and conceptualisation of nature. It would be a science by and for machines. Initial steps have already been taken toward this scenario, with publication formats for scientific endeavours that are machine readable, so that the limits of human communication could be overtaken (Stocker et al., 2025[66]).

One possible solution to this problem is the development of AI models in charge of translating the findings of AI scientific models into human language.

Towards data-driven science?

Theory can be viewed as a sort of compression of data, which allows for the preservation of information (predictive power) while skipping less relevant details. Theory is needed for humans, as the brain has limited computing capacities. It works by abstracting vast amounts of information in concepts and their linkages. It is used to predict features of future situations, on the basis of experience with past situations. From this perspective, with AI, and notably GenAI, this capacity is being expanded and therefore the need for compression, for theory, is getting weaker. In 2008, Chris Anderson, in a famous piece in the journal Wired, envisaged "the end of theory" (Anderson, 2008[67]). In the words of Kristin Persson, a professor in materials science:

> "we are entering a new era in science — the fourth paradigm. The first paradigm was empirical science based on experiments; the second was model-based science that develops equations to explain experimental observations; and the third created simulations based on those equations. The fourth paradigm will be science driven by big data and AI, you now have enough data that you can train machine-learning algorithms" (Persson, 2025[68]).

There are certain signs of this "atheoretical science" in publications. Duede (2024[2]) finds that AI-engaged papers in any discipline are becoming more semantically similar to the AI-engaged papers of Computer Science and less similar to other research published in their discipline. There would thus be a sort of convergence of scientific fields towards a common, data- and AI-driven framework. In the field of computational linguistics, for instance, LLMs have taken over what used to be structured, theory-based analysis of language.

Such an evolution might raise a risk of focusing research on questions that are data-rich, i.e. more operational questions, at the expense of more "blue sky", i.e. conceptual and foundational ones (Evans et al., 2024[3]). It could be associated with a decrease in research productivity (Box 13.5), understood as the number or variety of ideas in the field rather than the number of papers produced per researcher. It may also just be a phase in science that will lead to new theories and research questions.

Box 13.5. The impact of AI on research productivity

Certain observers have detected a possible decline in the productivity of research (OECD, 2023[69]), manifested by a slow-down in the number of significant discoveries and impactful inventions (e.g. the "Eroom's law", the reverse of Moore's law that points to a doubling of the effectiveness of computer chips every 18 months). Factors often mentioned as affecting this possible decline include notably 1) the increasing quantity of knowledge to manage in order to design new knowledge, which creates a "burden" for researchers; 2) the increasing complexity of phenomena to be analysed by scientists (the simplest ones having been grasped already); 3) a growing administrative burden as researchers have to do increasing amounts of paperwork; 4) distorted incentives that orient researchers towards research topics that are "low risk, low reward", or exploitation at the expense of exploration. AI, notably GenAI, could alleviate the first two factors, as it can manage great quantities of knowledge and it can model extremely complex processes. It could also alleviate the third one, as it can help in paperwork (submitting grant applications, etc.). As for the fourth factor, its impact is uncertain, as explained above.

Conclusion

The diffusion of GenAI in scientific research has been extremely rapid. GenAI has gained a core role in all text-related tasks like writing, editing and translating; it is making breakthroughs in analytical work with specialised foundation models for complex phenomena; it is progressing in its ideation capabilities, for suggesting hypotheses or experiment design; and it is becoming a credible, if fallible, research assistant. Its impact on the functioning, the productivity and the epistemological dimensions of research are not fully clear yet, not least because the technology is in rapid evolution. It is, however, already clear that scientific research will be transformed in the years to come, especially as new generations of researchers with increasing grasp of the technology will take over (see Table 13.2).

Policies for encouraging GenAI, notably in the domain of research, have been examined extensively already (OECD, 2023[70]; OECD, 2023[69]); they research funding, investment in computing, infrastructure and appropriate regulatory frameworks. One important aspect is training: research in most, if not all, domains does or would benefit from AI, and researchers with the required skills and competencies are in short supply. Most scientists able to master AI techniques are junior researchers, fresh from specialised curricula, while senior researchers struggle with new techniques. This also risks creating a rift between AI experts and domain experts, resulting, on the one hand, in publications intensive in AI but poor in domain relevance, and, on the other hand, in publications strong in domain relevance but under-using AI. Hence, there is a need to train more scientists in AI curricula and to encourage multi-disciplinary teams, but also to involve more senior researchers in AI-related life-long learning. This is a new responsibility for higher education.

This evolution will also gradually affect education research. One can assume that many of these trends are already or will soon be visible in education research. It is likely that education researchers already use GenAI tools for language-related tasks, such as writing and editing, as well as for programming and data analysis. This is the same for GenAI tools used or designed to manage scientific knowledge: they can also be used to search the education research literature, to produce literature reviews or even research reports. While most specialised scientific GenAI tools were developed for the natural sciences, similar ones will most likely become available for human and social sciences (beyond the "research models" of large LLMs). As AI teacher assistants are being developed for teachers and students, education researchers and their research assistants, if any, could also benefit from GenAI research assistants performing some of their routine tasks and combining some of the affordances mentioned above.

Table 13.2. The impacts of GenAI on science: A synthesis table

Pros	Cons	Uncertain
Access and process new types of data: unstructured data (text, images etc.)	Subject to confabulations	Creativity
Process large quantities of data	Often not explainable	
Connect distant areas of knowledge	Paper flooding	
Analyse extremely complex systems	Risk of lesser diversity	
Automatise certain research tasks, notably bureaucratic ones, allowing higher productivity of scientists, notably more focus on ideation tasks		
Can improve reproducibility		

As in other scientific fields, the modelling of complex phenomena such as learning or student pathways (e.g. Pardos and Borchers, 2026[16]) has already started, it is likely to continue to expand thanks to the new possibilities of AI to explore large and complex datasets. This may allow for new types of analysis. One can also imagine that GenAI tools will be able to extract hypotheses from the research literature, improve some of these hypotheses thanks to the ability to manage more data and possibly generate new ones.

The generation of synthetic privacy-preserving datasets that largely replicate original datasets is also a new possibility offered by GenAI. This would indeed allow to explore and analyse datasets that are not shared, linked (or created) for privacy reasons. Education research would then potentially have broader access to both administrative data and data collected by digital learning platforms. Synthetic datasets require careful evaluation to ensure they properly replicate the statistical properties of the original dataset and really preserve privacy. As part of education research relies on survey data, which are said to be increasingly difficult to collect, the possibility to augment those data with simulated data also opens new possibilities, even though it is still unclear how promising this avenue is. While the development and analysis of standardised tests now routinely includes the generation of simulated answers (that is, simulated students), whether this technique can be extended to other subjects is still debated. One promising avenue in scientific research that could also be used in education research are AI multi-agent models: they could help address questions that benefit from inter-disciplinary research or that would involve the feedback from simulated social agents, as can be the case in policy research and development.

Last but not least, education research may benefit from the advances in other fields that are made possible by new scientific GenAI tools. One could indeed imagine that, with the right incentives new research could explore the chemistry of the brain while learning or that neuroscience and cognitive sciences will provide new insights on the biological development of children that could inform education research and policies. There is clear convergence between machine learning as a discipline and psychology, notably developmental psychology, with bidirectional knowledge flows, where psychology informs machine learning about training methods and architectures and machine learning offers psychology a testbed for its hypotheses (see e.g. Goddu, 2024[71]).

While it is difficult to predict how GenAI will become integrated in the fabric of education research, not as a subject but as a tool, one clear lesson of its use in other fields so far is that successful GenAI-powered research requires the tight supervision of humans and usually corresponds to an augmentation rather than a replacement model.

References

Ahmed, N., M. Wahed and N. Thompson (2023), "The growing influence of industry in AI research", *Science*, Vol. 379/6635, pp. 884-886, https://doi.org/10.1126/science.ade2420.   [65]

Anderson, B., J. Shah and M. Kreminski (2024), "Homogenization Effects of Large Language Models on Human Creative Ideation", *Creativity and Cognition*, pp. 413-425, https://doi.org/10.1145/3635636.3656204.   [52]

Anderson, C. (2008), *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, https://www.wired.com/2008/06/pb-theory/.   [67]

Anthropic (2025), *How we built our multi-agent research system*, https://www.anthropic.com/engineering/multi-agent-research-system.   [40]

Baek, J. et al. (2025), *ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models.*   [46]

Baker R.S., X. Liu, M. Shah, M. Pankiewicz, Y.J. Kim, Y. Lee and C. Porter (2026), "Generative AI as a teaching assistant", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en.   [44]

Bergstrom, C. and J. Bak-Coleman (2025), *AI, peer review and the human activity of science*, https://www.nature.com/articles/d41586-025-01839-w.   [62]

Biever, C. (2025), *AI scientist 'team' joins the search for extraterrestrial life*, https://www.nature.com/articles/d41586-025-01364-w.   [36]

Binz, M. (2025), A foundation model to predict and capture human cognition, https://www.nature.com/articles/s41586-025-09215-4.   [25]

Bran, A. (2025), "Revealing chemical reasoning in LLMs through search on complex planning tasks", A*I for Accelerated Materials Design-ICLR.*   [19]

Callaway, E. (2025), *AlphaFold is running out of data - so drug firms are building their own version*, https://www.nature.com/articles/d41586-025-00868-9.   [15]

Castelvecchi, D. (2024), R*esearchers built an 'AI Scientist' - what can it do?*, https://www.nature.com/articles/d41586-024-02842-3 (accessed on 19 December 2025).   [23]

Chenglei Si, D. (2024), *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers.*   [51]

Chiou, L. (2025), "At Secret Math Meeting, Researchers Struggle to Outsmart AI", *Scientific American*, https://www.scientificamerican.com/article/inside-the-secret-meeting-where-mathematicians-struggled-to-outsmart-ai/.   [42]

DeepMind (2024), *AI achieves silver-medal standard solving International Mathematical Olympiad problems*, https://deepmind.google/blog/ai-solves-imo-problems-at-silver-medal-level/.   [22]

Doshi, A. (2024), "Generative AI enhances individual creativity but reduces the collective diversity of novel content", *Science Advances*, Vol. 10, no. 28/eadn5290.   [57]

Duede, E., W. Dolan, A. Bauer, I. Foster and K. Lakhani (2024), *Oil & Water? Diffusion of AI Within and Across Scientific Fields.*   [2]

Evans, J., Q. Hao, F. Xu and Y. Li (2024), *Artificial Intelligence Expands Scientists' Impact but Contracts Science's Focus*, https://arxiv.org/pdf/2412.07727#page=1.70.   [3]

Gans, J. (2025), *Growth in AI Knowledge, National Bureau of Economic Research*, Cambridge, MA, https://doi.org/10.3386/w33907.   [53]

Gibney, E. (2025), *AI tools are spotting errors in research papers: inside a growing movement*, https://www.nature.com/articles/d41586-025-00648-5.   [13]

Goddu, M. (2024), *The Development of Human Causal Learning and Reasoning*, https://www.nature.com/articles/s44159-024-00300-5.   [71]

Gottweis, J. (2025), *Towards an AI co-scientist*, https://arxiv.org/pdf/2502.18864.   [37]

Gruda, D. (2024), *Three ways ChatGPT helps me in my academic writing*, https://www.nature.com/articles/d41586-024-01042-3 (accessed on 5 November 2025).   [9]

Hayes, T. (2025), "Simulating 500 million years of evolution with a language model", *Science*, Vol. 387/6736, pp. 850-858.   [20]

Heidt, A. (2025), *AI for research: the ultimate guide to choosing the right tool*, https://www.nature.com/articles/d41586-025-01069-0 (accessed on 5 November 2025).   [7]

Isungset, M.A. D. Conley, H.D. Zachrisson, E.Ystrom, A. Havdahl, P.R. Njølstad and T.H. Lyngstad (2022), "Social and genetic associations with educational performance in a Scandinavian welfare state", *Proceedings of the National Academy of Sciences*, Vol. 119/25, https://doi.org/10.1073/pnas.2201869119.   [17]

Jones, N. (2025), *OpenAI's 'deep research' tool: is it useful for scientists?*, https://www.nature.com/articles/d41586-025-00377-9 (accessed on 5 November 2025).   [33]

Kapania, S., W. Agnew, M. Eslami, H. Heidari, and S.E. Fox (2025), "Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants", *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1-17, https://doi.org/10.1145/3706598.3713220.   [26]

King, R. (2024), *From Adam to Genesis: Robot Scientists*, https://www.mondodigitale.org/en/news/adam-genesis-robot-scientists.   [47]

Kitano, H. (2016), "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery", *AI Magazine*, Vol. 37/1, pp. 39-49, https://doi.org/10.1609/aimag.v37i1.2642.   [58]

Kobak, D., R. González-Márquez, E-A. Horvát and J. Lause (2025), "Delving into LLM-assisted writing in biomedical publications through excess vocabulary", *Science Advances*, Vol. 11/27, https://www.science.org/doi/10.1126/sciadv.adt3813.   [12]

Krenn, M., Y. Drori and A. Adhikari (2025), "Digital Discovery of Interferometric Gravitational Wave Detectors", *Physical Review X*, Vol. 15, https://journals.aps.org/prx/pdf/10.1103/PhysRevX.15.021012.   [24]

Kudiabor, H. (2024), *How AI-powered science search engines can speed up your research*, https://www.nature.com/articles/d41586-024-02942-0 (accessed on 5 November 2025).   [29]

Kwon, D. (2025), *Is it OK for AI to write science papers? Nature survey shows researchers are split*, https://www.nature.com/articles/d41586-025-01463-8 (accessed on 5 November 2025).   [11]

Lee, H., A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks and N. Wilson (2025), "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers", *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1-22, https://doi.org/10.1145/3706598.3713778.   [56]

Lenharo, M. (2024), *ChatGPT turns two: how the AI chatbot has changed scientists' lives*, https://www.nature.com/articles/d41586-024-03940-y.   [8]

Liang, W., Y. Zhang, H., B. Cao, B. Wang, D. Ding, D., X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland and J. Zou (2024), *Can large language models provide useful feedback on research papers? A large-scale empirical analysis.*   [4]

Liu, Y. (2025), "Leveraging LLM respondents for item evaluation: A psychometric analysis", *British Journal of Educational Technology*, Vol. 56(3), pp. 1028-1052.   [27]

Li, W. (2023), *International conference on machine learning - Transactions on machine learning research.*   [60]

Ludwig, J. and S. Mullainathan (2023), *Machine Learning as a Tool for Hypothesis Generation*, https://doi.org/10.3386/w31017.   [35]

McElfresh, D. (2025), *The AI tool that can interpret any spreadsheet instantly*, https://www.nature.com/articles/d41586-024-03852-x.   [14]

Morris, T. T., S. von Hinke, L. Pike, N.R. Ingram, G. Davey Smith, M.R. Munafò and N.M. Davies (2022), "Implications of the genomic revolution for education research and policy", *British Educational Research Journal*, Vol. 50/3, pp. 923-943, https://doi.org/10.1002/berj.3784.   [18]

Munafò, M. (ed.) (2025), "Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database", *PLOS Biology*, Vol. 23/5, , https://doi.org/10.1371/journal.pbio.3003152.   [64]

Naddaf, M. (2025), "How are researchers using AI? Survey reveals pros and cons for science", *Nature*, https://doi.org/10.1038/d41586-025-00343-5.   [6]

Noy, S. (2023), "Experimental evidence on the productivity effects of generative artificial intelligence", *Science*, Vol. 381/6654, pp. 187-192.   [59]

OECD (2023), *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris, https://doi.org/10.1787/a8d820bd-en. [69]

OECD (2023), *Initial Policy Considerations for Generative Artificial Intelligence*, OECD Publishing, https://www.oecd.org/en/publications/initial-policy-considerations-for-generative-artificial-intelligence_fae2d1e6-en.html. [70]

O'Neil, C., T. Ghosal, R. Raileanu, M. Walmsley, T. Bui, K. Schawinski and I. Ciuca (2025), "Sparks of Science: Hypothesis Generation Using Structured Paper Data", https://arxiv.org/pdf/2504.12976. [34]

Pardos, Z. and C. Borchers (2026), "AI in institutional workflows: learning from higher education to unlock new affordances for education systems and institutions", in *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris, https://doi.org/10.1787/062a7394-en. [16]

Pearson, H. (2024), "Can AI review the scientific literature - and figure out what it all means?", *Nature*, Vol. 635/8038, pp. 276-278, https://doi.org/10.1038/d41586-024-03676-9. [32]

Penadés, J.R., J. Gottweis, L. He, J.B. Patkowski, A. Shurick, W-H. Weng,… and R.D. Tiago (2025), *AI Mirrors Experimental Science to Uncover A Novel Mechanism of Gene Transfer Crucial to Bacterial Evolution*, Cold Spring Harbor Laboratory, https://doi.org/10.1101/2025.02.19.639094. [43]

Persson, K. (2025), *The two-way relationship between AI and materials science*, https://www.nature.com/articles/d42473-025-00107-9 (accessed on 5 November 2025). [68]

Peters, U. and B. Chin-Yee (2025), "Generalization bias in large language model summarization of scientific research", *Royal Society Open Science*, Vol. 12/4, https://doi.org/10.1098/rsos.241776. [30]

Petrov, I., J. Dekoninck, L. Baltadzhiev, M. Drencheva, K. Minchev, M. Balunovic, N. Jovanovic and M. Vechev (2025), *Proof or bluff? Evaluating LLMs on 2025 USA Math Olympiad*, https://arxiv.org/pdf/2503.21934. [41]

Prillaman, M. (2024), *Is ChatGPT making scientists hyper-productive? The highs and lows of using AI*, https://www.nature.com/articles/d41586-024-00592-w. [10]

Ren, Z., Z. Zhang, Y. Tian and J. Li (2023), *CRESt – Copilot for Real-world Experimental Scientist*, American Chemical Society (ACS), https://doi.org/10.26434/chemrxiv-2023-tnz1x-v4. [48]

Sakana.ai (2025), *The AI Scientist Generates its First Peer-Reviewed Scientific Publication*, https://sakana.ai/ai-scientist-first-publication/ (accessed on 5 November 2025). [63]

Saplakoglu, Y. (2024), "How AI Revolutionized Protein Science, but Didn't End It", *Quanta Magazine,* https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/ (accessed on 5 November 2025). [21]

Schmidgall, S., Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, M. Moor, Z. Liu and E. Barsoum (2025), *Agent Laboratory: Using LLM Agents as Research Assistants.* [55]

Shojaee, P., I. Mirzadeh, K. Alizadeh, M. Farajtabar, M. Horton and S. Bengio (2025), *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.* [50]

Si, C. (2024), *Large language models help humans verify truthfulness - except when they are convincingly wrong.* [54]

Simon, H. (2001), "Creativity in the Arts and the Sciences", *The Kenyon Review*, Vol. 23(2), pp. 203-220, http://www.jstor.org/stable/4338222. [49]

Skarlinski, M., S. Cox, J.M. Laurent, J.D. Braza, M. Hinks, M.J. Hammerling, M. Ponnapati, S.G. Rodriques and A.D. White (2024), "Language agents achieve superhuman synthesis of scientific knowledge", https://doi.org/10.48550/arXiv.2409.13740. [31]

So, R. (2025), *AI-based Scientific Research Assistants*, https://project-rachel.4open.science/Rachel.So.AI-based.Scientific.Research.Assistants.pdf. [45]

STM (2023), *Generative AI In Scholarly Communications - Ethical and Practical Guidelines for the Use of Generative AI in the Publication Process*, https://stm-assoc.org/document/stm-generative-ai-paper-2023/. [61]

Stocker, M., L. Snyder, M. Anfuso, O. Ludwig, F. Thießen, K.E Farfar, M. Haris, A. Oelen and M.Y. Jaradeh (2025), "Rethinking the production and publication of machine-readable expressions of research findings", *Scientific Data*, Vol. 12/1, https://doi.org/10.1038/s41597-025-04905-0. [66]

Swanson, K., W. Wu, N.L. Bulaong, J.E. Pak and J. Zou (2025), "The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies", *Nature*, Vol. 646/8085, pp. 716-723, https://doi.org/10.1038/s41586-025-09442-9. [39]

Vincent-Lancrin, S. and G. Jacotin (eds.) (2023), *Measuring Innovation in Education 2023: Tools and Methods for Data-Driven Action and Improvement*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/a7167546-en.   [1]

Wang, Q., D. Downey, H. Ji and T. Hope (2024), *SciMon: Scientific Inspiration Machines Optimized for Novelty*, https://aclanthology.org/2024.acl-long.18.pdf.   [38]

Wiley (2025), *ExplanAItions: An AI study by Wiley,* https://www.wiley.com/content/dam/wiley-com/en/pdfs/about/wiley-explanaItions-ai-study-february-2025vers3.pdf.   [5]

You, X. (2024), *AI analysed 1,500 policies to cut emissions. These ones worked*, https://www.nature.com/articles/d41586-024-02717-7 (accessed on 5 November 2025).   [28]

OECD Digital Education Outlook 2026

Exploring Effective Uses of Generative AI in Education

The *OECD Digital Education Outlook 2026* explores emerging research on the use of generative AI in education and presents innovative tools and applications that show promise. The report examines the use of generative AI in different teaching and learning scenarios: by students to learn subject knowledge, by students and teachers together as part of instruction, and by teachers alone to support their work in the classroom. The Outlook also showcases how generative AI can enhance the efficiency of education institutions and systems, such as research, analysis of learning pathways and support for study advisors. While highlighting the opportunities and benefits of generative AI, the report also notes key challenges and concerns. It provides a state-of-the-art overview of the impact and possibilities of generative AI in education and is an essential resource for education policymakers, institution leaders, researchers, as well as other stakeholders.