Taylor & Francis
Taylor & Francis Group

Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

The promise and challenges of generative AI in education

Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis & Bart Rienties

To cite this article: Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis & Bart Rienties (2025) The promise and challenges of generative AI in education, Behaviour & Information Technology, 44:11, 2518-2544, DOI: 10.1080/0144929X.2024.2394886

To link to this article: https://doi.org/10.1080/0144929X.2024.2394886

Published online: 02 Sep 2024.

BEHAVIOUR & INFORMATION TECHNOLOGY
2025, VOL. 44, NO. 11, 2518–2544
https://doi.org/10.1080/0144929X.2024.2394886

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

The promise and challenges of generative AI in education

Michail Giannakos[a,b], Roger Azevedo[c], Peter Brusilovsky[d], Mutlu Cukurova[e], Yannis Dimitriadis[f], Davinia Hernandez-Leo[g], Sanna Järvelä[h], Manolis Mavrikis[e] and Bart Rienties[i]

[a]Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway; [b]Department of Information Systems, University of Agder, Kristiansand, Norway; [c]School of Modeling Simulation and Training, University of Central Florida, Orlando, FL, USA; [d]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA; [e]UCL Knowledge Lab, Institute of Education, University College London, London, UK; [f]School of Telecommunications Engineering, Universidad de Valladolid, Valladolid, Spain; [g]Department of Information and Communication Technologies, Universitat Pompeu Fabra (UPF), Barcelona, Spain; [h]Learning and Educational Technology (LET) Research Lab, University of Oulu, Oulu, Finland; [i]Institute of Educational Technology, The Open University, Milton Keynes, UK

ABSTRACT
Generative artificial intelligence (GenAI) tools, such as large language models (LLMs), generate natural language and other types of content to perform a wide range of tasks. This represents a significant technological advancement that poses opportunities and challenges to educational research and practice. This commentary brings together contributions from nine experts working in the intersection of learning and technology and presents critical reflections on the opportunities, challenges, and implications related to GenAI technologies in the context of education. In the commentary, it is acknowledged that GenAI's capabilities can enhance some teaching and learning practices, such as learning design, regulation of learning, automated content, feedback, and assessment. Nevertheless, we also highlight its limitations, potential disruptions, ethical consequences, and potential misuses. The identified avenues for further research include the development of new insights into the roles human experts can play, strong and continuous evidence, human-centric design of technology, necessary policy, and support and competence mechanisms. Overall, we concur with the general skeptical optimism about the use of GenAI tools such as LLMs in education. Moreover, we highlight the danger of hastily adopting GenAI tools in education without deep consideration of the efficacy, ecosystem-level implications, ethics, and pedagogical soundness of such practices.

ARTICLE HISTORY
Received 30 March 2024
Accepted 9 August 2024

KEYWORDS
Generative AI in education; AI in education; large language models; commentary

Abstract

1. Introduction

In recent years, generative AI (GenAI) has emerged as one of the fastest technology take-ups in human history. Largely driven by large language models (LLMs), GenAI tools have become the talk of every school, every teacher, and every learning-related scientific venue, positioning GenAI in the epicentre of today's research, policy, and practice. As technology advances, educators are attempting to identify opportunities to help students learn in new ways, as well as determining the impact of GenAI tools on life, learning, and work. Today, there is a lack of consensus on whether and, if so, how GenAI should be used to support teaching and learning.

Problem

In most of the published works thus far (e.g. Bhandari, Liu, and Pardos 2023; Nguyen et al. 2023) the efficacy of GenAI tools was promising, but always accompanied by several limitations, shortcomings, and ethical implications (Hamilton, Wiliam, and Hattie 2023). Therefore, though there is no doubt that GenAI will have both positive and adverse impacts on education in the coming years, further work is needed to understand its opportunities and challenges, as well as the ways it will affect contemporary practices in terms of assessment, course creation, learning design, learning objectives, and so on. With this background, we seek to shed light on the two following research questions:

> RQ1) What are the opportunities, challenges, and implications related to GenAI technologies, in the context of learning technologies and education?

> RQ2) What are the most important research topics related to GenAI technologies, in the context of learning technologies and education?

To do this, we call on nine experts who work in the intersection of learning and technology (learning technology) research from eight leading learning technology units across five countries to share their views and

CONTACT Michail Giannakos michailg@ntnu.no

Approach

provide critical reflections on the opportunities, challenges, and implications related to GenAI technologies, in the context of learning technologies and education. GenAI refers to deep-learning models that can generate content beyond just textual data, including images, videos, and even music. In parts of this commentary, we focus on LLM technology and even on specific LLM tools; this is done due to their wide adoption by educators and the existence of early research on LLMs and specific tools. We employ purposeful use of terminology[1] while considering the overarching capabilities of GenAI and keeping the discussion inclusive of content beyond text. In essence, the main objective of this commentary is to draw on learning technology experts' experience and provide a summary and synthesis of their insights, discuss the major challenges and opportunities, and provide an agenda for future research in the area of GenAI.

The paper is organised as follows: In Section 2, we provide a short overview of recent developments in the intersection of artificial intelligence (AI) and human learning. In Section 3, we list the individual perspectives of learning technology experts with different focus areas, such as learning design (LD), self-regulated learning (SRL), feedback generation, necessary capabilities, and necessary skills, as well as domain areas such as math and computing education. In Section 4, based on these contributions, we provide a synthesis and a short concluding thought on the potential of GenAI to advance learning technology research and practice, as well as directions for future research.

2. AI and human learning

Since the debut of AI in education (AIED) more than three decades ago (McCalla 2023), various AI approaches have been considered to foster innovative teaching and learning practices, presenting opportunities that would have otherwise been impossible to materialise. AI equips systems with reasoning, allowing them to learn from experience, adjust to new inputs, and perform human-like tasks. Similar to other application areas of AI, during the first years the focus of AI was mainly on using labelled (supervised learning) and unlabelled (unsupervised learning) data to identify patterns and make predictions (Duan, Edwards, and Dwivedi 2019). Traditional AI algorithms, such as decision trees, random forests, support vector machines, and k-means clustering, provided useful but limited capabilities (Duan, Edwards, and Dwivedi 2019). Nowadays, in the AIED literature, one can find different techniques, such as natural language processing (NLP), neural networks (NNs), machine learning (ML), deep learning, and genetic algorithms (Ouyang and Jiao 2021). Techniques that power contemporary learning technologies with AI capabilities in different ways – such as intelligent tutors, adaptive learning analytics and interfaces, and automated content generation and feedback – and ultimately support teaching and learning in various educational arenas (e.g. Gobert 2023; Neumann et al. 2021).

In recent years, AI has significantly impacted the way humans learn and the way the respective institutions operate (Bond et al. 2024). In particular, GenAI tools can automatically generate outputs such as text and images and synthesise speech and audio, as well as create original video content and generate datasets, which requires large training datasets, NNs, and deep learning architectures (Nirala, Singh, and Purani 2022). Largely driven by LLMs that use deep NN models to effectively analyze complex linguistic structures, GenAI is currently at the epicentre of policy and research. In this context, a major milestone came in November 2022, when OpenAI introduced a chatbot called ChatGPT (generative pre-trained transformer). ChatGPT is a generative conversational AI interface that uses natural language to interact in a realistic way and even answers follow-up questions; it admits 'its mistakes, challenges incorrect premises, and rejects inappropriate requests' (OpenAI 2023). Although the goal of ChatGPT is to mimic human conversation and provide requested information, its capabilities extend to teaching and learning practices, such as solving exercises, creating essays, stories, poems, or acting like anything within its capability. Following ChatGPT's inception, various GenAI tools (particularly LLMs similar to ChatGPT) were either initiated or advanced and reached good efficacy (e.g. BERT and GitHub Copilot).

With the rise of GenAI applications such as ChatGPT and GitHub Copilot, the use of AI-enabled systems in teaching and learning has gained increased interest. Especially the use of ChatGPT is exhibiting peak interest in education, with almost every educational institute having developed its own policy concerning its use. Some countries have developed 'sandboxes' that allow public and private organisations to try out LLMs in a risk-free manner[2]; they have also developed their own versions of LLMs that better support some languages or address potential privacy and security concerns.[3] Moreover, UNESCO has published a report on GenAI in education (Miao and Holmes 2023), and the UK Department for Education (2023) and the Council of Europe[4] have outlined position statements indicating guidelines and the need for regulation. Companies such as MagicSchool and Eduaide have developed student and teacher assistants based on OpenAI's LLM technology, whereas others such as Turnitin have

Contributions

developed plagiarism detection tools for LLMs. In the past months, LLM studies have indicated promising areas of use, such as generating help messages and feedback (Nguyen et al. 2023), as well as questions in math (Bhandari, Liu, and Pardos 2023), including help-seeking and code improvement in programming education (Prather et al. 2023). Recently, textbooks have featured the use of GenAI (GitHub Copilot and ChatGPT) in teaching and learning programming (Porter and Zingaro 2024).

3. Perspectives from leading experts in learning technologies

In accordance with previous expert viewpoints on a diverse range of topics, such as the Metaverse (Dwivedi et al. 2022) and the future HCI grand challenges (Stephanidis et al. 2019), to mention a few, we examine the critical perspectives on the impact and challenges of GenAI on teaching and learning. Given the nature of this contribution (a commentary), we did not employ any strict protocol in synthesising the contributions. However, we adopt well-established processes developed for the Horizon reports[5] and Innovative Pedagogy.[6] In particular, to produce this contribution, a group of experts from eight institutions and five countries collaborated together from September 2023 to February 2024 via digital tools, meet-ups, and review processes. A long list of potential AI impacts on learning technology research and practice that has the potential to provoke major shifts in learning technology was discussed. Subsequently, individual authors and groups of authors worked on a range of contributions to share their views and provided critical reflections on these topics. These contributions were subsequently reviewed by other group members, then revised and further fine-tuned. Finally, the prospects and implications section was written by the first author (who went through all the contributions and wrote memos, from which they abstracted the high-level themes) and subsequently fine-tuned by various iteration cycles until no more comments or additions emerged from the contributing authors.

The full list of experts and their individual contributions are listed in Table 1. In particular, the contributors to this commentary include Michail Giannakos (NTNU, Norway), Roger Azevedo (UCF, USA), Peter Brusilovsky (Pitt, USA), Mutlu Cukurova (UCL, UK), Yannis Dimitriadis (UVA, Spain), Davinia Hernandez-Leo (UPF, Spain), Sanna Järvelä (U Oulu, Finland), Manolis Mavrikis (UCL, UK) and Bart Rienties (IET Open, UK). All authors are senior professors with more than 10 years of experience in the intersection of learning and technology, holding positions in prominent institutes and serving (or having served) as members of top-tiered relevant journals. The experts were selected based on their relevant expertise to account for important learning-related topics such as learning design, collaborative learning, metacognition, assessment, self-regulated learning; technology-related topics such as personalised learning, user\learner modelling, intelligent tutoring systems, recommender systems, and intelligent interfaces; as well as interdisciplinary topics such as AI literacy, ethics of AI in education, and hybrid intelligence and instruction. Although we did not intend, nor claim to have a complete coverage of topics, the lineup of experts accounts for a certain degree of plurality, representing several of the crucial domains in the intersection of learning and technology. The biographies of each contributor are included in the appendix of the commentary.

Table 1. Individual contributions.

Contribution Title	Author(s)
Contribution 1: LLMs: Upskilling, reskilling, or degrading human learning?	Michail Giannakos
Contribution 2: The metacognitive and SRL/SSRL issues with conversational GenAI in education	Roger Azevedo and Sanna Järvelä
Contribution 3: LLMs for computer science education: Early success and recognised challenges	Peter Brusilovsky
Contribution 4: Automatic content creation and learning design	Bart Rienties
Contribution 5: A human-centred perspective to GenAI and analytics layers in Learning Design	Davinia Hernández-Leo and Yannis Dimitriadis
Contribution 6: The use of LLMs in learning diagnosis and feedback content generation	Mutlu Cukurova
Contribution 7: Leveraging AIED foundations in the age of GenAI: The case of mathematics education	Manolis Mavrikis

3.1. Contribution 1: LLMs: upskilling, reskilling, or degrading human learning? by Michail Giannakos

3.1.1. Summary

LLMs is capable of generating human-like text based on context, prompts, and past conversations. Although such technology has been around for several years, it has only rapidly grown in the past months and become widely adopted. Despite the fact that several national and international institutes have already devised different policies for the use of LLM technology, its role in future learning technologies (and the education landscape) remains a topic of discussion and contention. Current discussion papers and early research papers acknowledge LLMs' capabilities to enhance teaching and learning, suggesting that it is likely to offer significant gains in education. To accelerate the current research debate and support the future research agenda, this contribution offers four provocations that depict certain future research challenges. First, the skills required in the world will likely look different. Second, LLM-like technology will impact current teaching and instruction practices. Third, contemporary and future learning technologies must embrace LLMs' capabilities if they wish to stay relevant. Fourth, as with any technological advancement, LLMs will be misused, and certain restrictions or legislations will be needed.

3.1.2. Introduction

LLM technologies, such as ChatGPT, have become the talk of every educational arena. It allows teachers and learners to 'ask anything,' and 'it may have a good answer' – in fact, in most cases, it does. LLMs have already managed to disrupt several teaching and learning practices, and a volume of early studies and media outlets have reported the advantages and best practices of LLMs in education (Baidoo-Anu and Owusu Ansah 2023; Kasneci et al. 2023).

Although it is not a surprise that something like this could have happened, the speed at which it has occurred has caught most of us off guard. Technology is, by definition, disruptive. It enables scientific knowledge to support the achievement of practical goals of human life, but, as a byproduct, it also reshapes activities and behaviours, and thus it must be regulated. When it comes to education, technology has offered several opportunities and disrupted our practice several times in the past. In the 1960s, electrical engineering students used a mechanical device called a slide rule to do their calculations. Slide rules allowed them to do multiplication, find squares and square roots, and conduct other calculations needed for their domain. By the 1970s, the introduction of electronic calculators made slide rules obsolete. This development resulted in reskilling (i.e. learning a new skill: how to use a calculator) and upskilling (i.e. expanding their existing skill set), and it even contributed to the degradation of certain electrical engineering skills (e.g. how to use a slide rule). The slide rule was heavily used for nearly 400 years and was the most commonly used calculation tool in science and engineering. More recent examples include the use of physical libraries to access research papers and other information, or the use of books for finding examples and solutions for certain exercises. This required going to different university buildings, lending different books and magazines, and making notes and copies. Today, most of these books and articles are readily available on the internet, and services such as StackOverflow allow students to find a greater range of information that has more plurality and is more up-to-date. All in all, these advancements have differentiated the development of students and scientists. The skills of finding and having a collection of relevant books and articles are no longer needed, and new skills such as managing large volumes of information and critical thinking have emerged.

3.1.3. Provocations

Along the same lines, due to LLMs, skills reorganisation will permeate various spheres of education. For instance, the rise of LLMs has direct effects on assessment and examination, making certain types of assessment obsolete. In recent months, teachers have been discussing whether and how they can either use or restrict LLMs, whereas some universities (and countries) have already banned them over fears of student plagiarism. This is because some contemporary practices (e.g. assessment practices and assignments) fail to safeguard the principles of our academic integrity (e.g. students might pass assignments and courses without obtaining the necessary competence). Indeed, this is a great opportunity for the learning technology community to intensify its efforts and help society embrace LLM technologies. Instead of using them as 'systems that hinder or avoid human learning' (e.g. just copying a solution without understanding it), we should use them as 'systems that help humans learn important skills' (e.g. as a readily available personal tutor). To support this line of work, I offer a number of provocations (P) that depict certain research directions.

P1: The skills required in a world powered by LLM technologies will be different.

The disruption caused by technologies challenges established assumptions about the skills required in society and the way domains function. For instance, LLMs can be fine-tuned on a specific domain to assist learners. We have already seen examples in the software industry where developers are writing and testing their code alongside LLM technologies (Deng et al. 2023). Such a shift has the potential to increase efficiency, but it also requires different skills from professionals (e.g. the ability to formulate effective prompts and train an LLM using reliable data sources).

P2: Teaching and instruction will be impacted by LLMs and will require a transition to stay relevant.

Several teaching tasks can be performed by LLMs (Sabzalieva and Valentini 2023), and it is inevitable that, in the near future, more teaching tasks will be 'mastered' by LLMs. Today, we see LLM-like technologies that automatically generate math word problems

(Wang, Lan, and Baraniuk 2021), which comes with the challenge of understanding equations and putting them into the appropriate context. Teaching and instruction should embrace LLMs in human – machine hybrid instruction; this will allow teachers to leverage LLMs to deliver high-quality teaching tasks, which, in turn, will free time for them to nurture learners' critical thinking (or do other tasks that LLMs perform poorly).

P3: Contemporary learning technologies will be challenged by LLMs, and embracing LLM capabilities will be critical for learning technologies to stay relevant.

LLM technology can provide novel support to students. For instance, in the context of programming education, LLM-like technologies are efficient in producing content, solutions to assignments, and automated code explanations (Sarsa et al. 2022). At the same time, Prather et al. (2023) highlighted numerous challenges in utilising LLMs in computing education, ranging from reliable to responsible and ethical integration. Thus, future learning technologies are likely to provide LLM capabilities as a widget or different service; for instance, LLMs can act as round-the-clock support for students, playing different roles such as a personal tutor, a study buddy, or an assessor. It is important that this enhancement is implemented in ways that follow our values and reliably augment the learning experience; this is likely to both support learners but will also allow teachers to better allocate their teaching resources. As with any other AI technology, this will require proper integration with existing systems and processes, which comes with different challenges (e.g. technical issues, data privacy, and ethics).

P4: We need to be ready that LLMs will be misused in the field of education.

Learning interactions and routines will clearly change, perhaps even more so than they did with the rise of social media and search technology such as Google. These changes will result in certain disruptions, and it is up to us to clearly understand and align them with our values and goals as a society where education plays a very important and responsible role in the development of civilisation. Understanding how the new reality of education will look and what skills will be needed will allow us to develop proper policies, frameworks, and competence opportunities. Thus, as LLM technology is established and our understanding and expertise grow, there will be a gradual leaning toward strategic embeddedness, which, in turn, will contribute toward strategic, efficient, and sustainable adoption.

Based on these provocations, I suggest that 'we,' as relevant and responsible researchers, need to go to work. After all, this is what we have always wanted in the learning technology field. Learning technology has the potential to greatly advance human learning, offering a great opportunity and responsibility for research and for society at large.

3.2. Contribution 2: the metacognitive and SRL/SSRL issues with conversational GenAI in education by Roger Azevedo and Sanna Järvelä

The presence of AI is growing in all areas of life, and it plays an increasingly important role in both students' learning and their future work lives. We believe that learners themselves should be active participants when learning and working with AI, and, to do this, they must develop their SRL skills to monitor and control their own learning. SRL is an agentic process where learners strategically take control of their learning engagement and situations through active, dynamic, and temporally unfolding cycles of planning, performance, and reflection (Azevedo et al. 2022; Winne 2018). Socially shared regulation in learning (SSRL) extends individual SRL to group-level regulatory processes and refers to a group's deliberate, strategic, and transactive planning, as well as task enactment, reflection, and adaptation. It involves groups taking metacognitive control of the task together through negotiated, iterative fine-tuning of cognitive, behavioural, motivational, and emotional conditions as needed (Järvelä et al. 2021). We argue the potential of GenAI to empower learners' SRL and SSRL processes in new ways and support the development of novel theoretical and empirical grounds.

Although GenAI presents opportunities in conceptual, theoretical, methodological, analytical, and educational issues, our interdisciplinary research community must still face various challenges (Azevedo and Wiedbusch 2023; Järvelä, Nguyen, and Hadwin 2023). The problem is that, despite the well-documented benefits of SRL knowledge and skills (Winne and Azevedo 2022) and the numerous opportunities students have in education to acquire, learn, and practice them, SRL knowledge and skills remain underdeveloped (Kistner et al. 2010). We believe that SRL/SSRL theory-guided AI development and adaptive learning technology design can use learning process data and AI algorithms to empower the agency of learners and teachers in terms of agency (Taub and Azevedo 2023). Although GenAI tools do not have any conceptual knowledge or conscious understanding, it is critical to create theory-grounded interventions in this new technology space. Below, we share some examples of how theoretically based and empirically driven approaches to GenAI can be used to trigger, induce, support, and foster both SRL and SSRL.

GenAI has emerged as a powerful educational tool, significantly influencing SRL through various mechanisms. One notable example is personalised learning pathways, where AI algorithms analyze individual learner data to generate tailored educational content (Tankelevitch et al. 2023). By adapting to a student's pace, engagement, interests, and self-regulatory behaviours, GenAI can promote autonomy, allowing learners to take control of their learning, especially if they are capable of verbally expressing their self-regulatory needs using NLP while interacting with advanced learning technologies (e.g. game-based simulation). Furthermore, AI-driven feedback systems play a crucial role in measuring and enhancing self-regulation, especially when utilising and fusing multimodal trace data, which can provide more accurate student models of learners' SRL knowledge and skills and real-time assessments, thereby highlighting strengths and weaknesses and enabling learners to reflect on their performance and adjust their strategies accordingly.

Moreover, GenAI contributes to creating immersive and interactive learning environments. AI-powered virtual simulations can offer dynamic scenarios that challenge learners and promote problem-solving skills, learning, and reasoning while building their SRL knowledge and skills. These simulations engage students and provide a safe space for experimentation and learning from mistakes. For example, they are ideal for learners to experiment with their SRL knowledge and skills that are (1) prompted by an external agent (e.g. a conversational agent), (2) acquired but not fully automated and thus require more guided practice with scaffolding, and (3) fully automated but require consideration of how to apply the SRL knowledge and skills to other similar tasks, domains, and contexts (i.e. developing their metacognitive conditional knowledge), as well as skills that (4) rely on GenAI's meta-reasoning skills to propose different methods and approaches to learning, problem solving, and reasoning. In summary, GenAI has become a cornerstone in learning technology, providing multifaceted support for SRL through personalised content delivery, real-time feedback, immersive simulations, collaborative platforms, and the cultivation of cognitive, metacognitive, affective, and motivational skills.

As collaborative interactions are mediated by technology in increasingly enriched ways (e.g. enabling learners to utilise movement, gesture, and gaze to support learning when co-located within virtual learning environments), additional data are available from collaborative learning interactions. Recently, researchers have recognised the potential of gathering and analyzing multimodal multichannel trace data during collaborative interactions and computer-supported collaborative learning (CSCL) processes to investigate cognitive, affective, metacognitive, motivational, and social processes within and across individuals (e.g. Malmberg et al. 2022). These data include eye movements (e.g. attentional allocation to relevant contextual cues), learner system logs (e.g. sequence of learners' interactions with a game-based simulation), screen recordings (e.g. showing the dynamics between learners and a game-based simulation), video and audio (e.g. at various scales showing individual learner contributions and dynamics between collaborating learners), discourse (e.g. illustrating timing, sequence, and dynamics of self – and co-regulated learning processes), and physiological data (e.g. electrodermal activity and heart rate showing physiological reactivity to pertinent aspects of collaborative learning contexts). These data increase our understanding of the nature, dynamics, timing, triggers, and duration of 'unobservable' shared phenomena, such as the role of affect and emotions, social – emotional interactions, metacognitive level processes, and SSRL in collaborative learning. These data not only increase researchers' understanding of SSRL processes, but they can also be represented to augment and empower peers, learners, teachers, and AI agents to increase their awareness, monitoring, regulation, and reflection of SRL and SSRL.

For example, in the context of CSCL, multimodal analytics generated from real-time (or near real-time) multimodal data can be presented to peers as actionable data, based on which one can make decisions about one's self-regulatory behaviour (Azevedo and Wiedbusch 2023). Similarly, during CSCL with an immersive virtual learning environment, open learner models (OLMs) can be presented as part of the system's interface that shows learners the system's beliefs about their SSRL processes and offers them opportunities to negotiate its beliefs about their SSRL processes (Bull et al. 2022). In addition, during collaborative learning with an intelligent tutoring system, multimodal data can be used to control the behaviour of conversational agents and, with advances in NLP, allow learners to naturally ask them about their perceptions of one's metacognitive skills or efficacy in regulating other team members (Järvelä et al. 2024). Conversational agents can now become AI team members who can share the cognitive load, monitor emerging SRL and SSRL in the team, provide individualised and team scaffolding to ensure productive collaboration, and, depending on the CSCL or collaborative learning context, generate new problems, cases, and scenario, based on an amalgamation of each individual's SRL and the team's SSRL to accommodate challenges and accelerate development, learning, problem solving, or reasoning within an individual and across the group.

Recent progress in advancing SRL research with AI will contribute to theory-guided GenAI design. In particular, multimodal data can be used to understand core human learning mechanisms, improve human – machine collaboration, and contribute toward the development of effective hybrid intelligence systems (Akata et al. 2020) that augment rather than replace human intelligence – systems that leverage our strengths and compensate for our weaknesses. Considering the rate at which AI is evolving, the SRL field is an active participant due to its strong understanding of learners' agency, leveraging current theories and developing new concepts to bring AI to SRL/SSRL research. Currently, the field has much to learn in terms of understanding AI and the power of GenAI beyond just using it 'as a new technological tool' but one that offers new research and learning opportunities for researchers, learners, and teachers.

3.3. Contribution 3: LLMs for computer science education: early success and recognised challenges by Peter Brusilovsky

Computer science education (CSEd) could be considered as a special domain for the educational application of LLMs. Indeed, an important step in the current LLM revolution was Open AI Codex, the LLM that powered GitHub Copilot between 2021 and 2023. Open AI Codex originated from text-trained GPT-3 and was additionally trained on 159 gigabytes of Python code from 54 million GitHub repositories. With this training, GitHub Copilot demonstrated a remarkable ability to produce code to solve simple programming problems in response to a problem statement (Chen et al. 2021). Needless to say, this functionality was immediately noticed by the instructors of programming courses and researchers in CSEd (Finnie-Ansley et al. 2022). Whereas the former group was mostly concerned with the disruption that the increased use of GitHub Copilot by students introduced to the traditional learning process, the latter team considered it an exciting opportunity to improve teaching and learning in computer science courses and began exploring it. By the end of 2022, when the release of ChatGPT opened access to LLMs to a much larger community, CSEd researchers had already accumulated experience with the use of LLMs in education. ChatGPT, powered by GPT 3.5 model, retained its code-trained component and helped to engage a broader cohort of CSEd researchers, who leveraged the work of pioneers and produced a range of interesting new results. This early start and the code-enriched nature of modern LLMs made the CSEd domain unique among other domains exploring the use of LLMs in education and facilitated the collection of many interesting results. In this sense, a brief analysis of successes and challenges collected in this area could offer a 'glimpse into the future,' demonstrating opportunities and challenges that have not yet been uncovered or encountered by other domains.

Early research on the use of LLMs in CSEd focused on testing the 'declared' capability of LLMs, (i.e. the ability to generate a programme in response to problem specification in natural language). In contrast to LLM researchers who tend to demonstrate the power of their models (e.g. Codex or AlphaCode) by testing how well their models can solve competition-level programming problems (Chen et al. 2021; Li et al. 2022), CSEd researchers began their exploration by checking how well LLMs could solve programming problems that are typically used as assignments and tests in programming courses (Finnie-Ansley et al. 2022; Finnie-Ansley et al. 2023; Nguyen and Nadi 2022). The results demonstrated that Codex is not a perfect problem solver; however, it outperforms the majority of students in solving typical course problems, produces reasonably understandable code, and can generate more than one correct solution for the same problem. For practitioners, these results were important to recognise that LLMs are likely to affect the integrity of traditional educational process, where teaching and testing are focused on solving small programming problems. For CSEd researchers, this was a clear call for innovation, both in rethinking the approach to teaching programming and producing a new generation of learning tools to support students.

Assessing how LLMs can solve typical educational problems is a natural 'testing the waters' stage in exploring the educational use of LLMs in many domains, but, in itself, the ability to solve a problem for students has relatively low educational value. Although researchers argued that this ability is still valuable, as it could be used to generate model solutions or demonstrate different ways to solve the same problem (Becker et al. 2023), an ideal learning support tool should assist students rather than replace them in the problem-solving process. Considering traditional intelligent tutoring systems as a 'proper' example of problem-solving support, we might expect that LLMs could provide similar levels of support, for example, diagnose errors in the middle of problem solving, provide several types of hints (i.e. explain what is wrong with the current solution or how to fix the current problem), and suggest a path to solving the problem.

In the context of solving a programming problem, these kinds of assistance are most needed when students are at an 'impasse' during problem-solving when they get stuck and do not know what to do next. Most frequently, this happens with student programmes either produce compilation errors or work incorrectly (as determined by a set of tests). In this situation, students usually seek help from instructors, teaching assistants, or friends, which takes time and breaks the process. As several research teams have demonstrated just within one year (2023), LLMs can handle this 'impasse' surprisingly well by delivering several types of support. Hellas et al. (2023) specifically explored how well Codex and ChatGPT can answer real help requests from students and demonstrated that prompting LLMs with a combination of problem statement, current state of the code, and text of the request could produce useful answers in the majority of cases, identifying at least one issue in 70% of cases for Codex and 90% for ChatGPT. In a similar work, Kiesler, Lohr, and Keuning (2023) attempted to classify the types of help that ChatGPT can provide in response to a simple prompt ('What's wrong with my code?'), followed by the code of the student's submission. They reported a range of helpful feedback, such as stylistic suggestions, explanations of how to fix the error, an explanation of the error, and a code with its fix.

Several research teams have explored LLMs' ability to provide specific kinds of help from this list and beyond, such as explaining compiler error messages that students frequently fail to understand (Leinonen et al. 2023; Santos, Prasad, and Becker 2023) or 'repair' bugs in the current student solution (Koutcheme et al. 2023; Phung et al. 2023). Researchers have also demonstrated that the performance of LLM is constantly increasing (Santos, Prasad, and Becker 2023) and that GPT4 could provide a better and more reliable explanation of compiler errors than Codex (Leinonen et al. 2023). The early results cited above were obtained using datasets of past student code submissions and the quality of LLM-generated help was assessed manually by the research teams, but Pankiewicz and Baker (2023) reported results of using ChatGPT to produce 'impasse' feedback for student code in a semester-long classroom study. They demonstrated that the majority of LLM hints were positively assessed by students and that the presence of hints considerably increased the likelihood of students completing assignments without human help. We hope that more classroom studies will follow next year, bringing more reliable data about the value of LLM in the learning process.

Finally, CSEd researchers have explored the opportunity to use the power of LLMs to automatically create learning content. Naturally, some of this work followed a broader stream of work on using LLMs to generate educational questions (Bulathwela, Muse, and Yilmaz 2023; Tran et al. 2023; Wang et al. 2022). More interesting, however, are the attempts to use the code-understanding power of LLM to generate more complex content, such as programming exercises and code explanations (Oli et al. 2023; Sarsa et al. 2022). In particular, following early promising results, several teams attempted to leverage LLMs' code explanation ability to produce working code examples, that is, examples of programming problem solutions augmented with code explanations (Hassany et al. 2023; Jury et al. 2024; MacNeil et al. 2023).

The current research on LLMs in CSEd has revealed an important issue related to LLM performance. A comparison with other domain results shows that this issue is domain independent, and it is useful to discuss this in the context of this paper. First, in all tasks performed by LLM in the CSEd context, their performance was not perfect. Even on the tasks for which LLMs are trained (i.e. code generation in response to a problem statement), they can fail to solve some problems and even solve some incorrectly. According to an early study by Nguyen and Nadi (2022), correctness rates for GitHub Copilot-generated problem solutions were between 27% and 57% across four languages. A study assessing Codex performance in generating explanations of error messages (Leinonen et al. 2023) demonstrated that LLMs could provide an explanation for 84% of the provided programmes and error messages, with a correctness rate of 57%. A study using LLMs for programming problem generation (Sarsa et al. 2022) reported that only 75% of generated problems were sensible. Among LLM answers to student help requests (Hellas et al. 2023), 48% reported issues that did not actually exist in the student's code. For tasks such as code explanation generation, success and correctness rates might be higher, and newer models tend to perform better than older models (i.e. Codex vs. GPT 3.5 vs. GPT 4) on most types of tasks (Savelka et al. 2023). Yet, LLMs are still not perfect, and this is an important limitation in an educational context. In the original context for which LLMs were designed (i.e. working with professional programmers, which GitHub Copilot was designed for) this is not a serious problem. Professionals can tolerate the lack of an answer and know how to assess its correctness if it is delivered. In contrast, students (who are domain novices) frequently cannot assess the correctness and quality of code, explanations, and other artifacts generated by LLMs.

This problem could be addressed in two complementary ways. First, as already argued by several authors (e.g. Becker et al. 2023; Finnie-Ansley et al. 2022), the traditional CSEd focus on problem solving should be complemented by increased attention to code comprehension and interpretation. It should better equip LLM-assisted students to assess the correctness and relevance of both code suggestions and explanations generated by LLMs. In the area of AIED, the importance of this 'answer interpretation' knowledge has been long advocated by Ohlsson and Mitrovic, and the ability to assess this knowledge has been implemented in several domains through episodic learner modelling (Mitrovic and Ohlsson 1999; Ohlsson 1992).

Second, to prevent potential harm, the developers of LLM-based learning tools should strive to ensure high quality of generated artifacts and avoid using tools whose quality cannot assured. The quality assurance approaches could depend on the nature of the generated artifacts. For example, the correctness of 'exemplary code solutions' suggested by Finnie-Ansley et al. (2022) could be assured by their performance on tests. The correctness of learning content generated by LLM (i.e. programming problems, explained examples) could be assured by engaging human – AI collaboration where generated content could be checked and improved by human authors (Hassany et al. 2023). Assuring the correctness of dynamic feedback generated by LLM in assisting the student in problem solving is the most challenging case, as direct engagement of 'humans in the loop' is not feasible here. An interesting idea is to use LLM itself in 'reverse' mode to validate dynamic error explanations, which was explored in the PyFiXV system by Phung et al. (2023).

3.4. Contribution 4: automatic content creation and learning design by Bart Rienties

With the recent advancements of GenAI, one obvious area in which to rapidly implement its affordances is automatic content creation and LD (Balaban, Rienties, and Winne 2023; Kasneci et al. 2023; Yan et al. 2023). Most learning courses consist of substantial amounts of written texts, digital artifacts, and different learning materials. Various recent estimates from big data studies on how educators design learning activities have suggested that typically between 40% and 90% of blended and online teaching and learning activities consist primarily of written artifacts (e.g. Albuquerque, Rienties, and Divjak 2024; Rizvi et al. 2022; Toetenel and Rienties 2016). For example, an analysis of 12,749 teaching and learning activities designed by 165 educators from 40 + institutions via an LD tool called Balanced Learning Design (Albuquerque, Rienties, and Divjak 2024) indicated that 55% of designed activities were primarily online content or assessment activities. In an analysis of 10 MOOCs by Rizvi et al. (2022), 52% of learning materials and activities were classified as written articles, 22% as videos, and 7% as (written) quizzes.

GenAI tools such as ChatGPT can read, (re-)design, and (re-)create these learning materials and activities (Kasneci et al. 2023; Yan et al. 2023). In the commercial sector and among some large-scale providers of online learning platforms, rapid progress is being made in using such automatic content creation approaches to design quick and personalised learning content. For example, the commercial company Stellar Labs (https://www.stellarlabs.io/) provides human resources companies with personalised training programmes within minutes based on automatic content creation using GenAI. Similarly, Open University UK, the largest university in Europe, is currently experimenting with GenAI to provide different versions of the same LD to groups with different learning needs, including accessibility and neurodiverse needs.

The affordances of automatic content creation and of offering different LDs based on the same content is an attractive proposition for creating quick and personalised learning activities for learners, as well as for tailoring learning experiences based on different learning needs, but there might also be some substantial challenges.

3.4.1. GenAI learning design is an art and a science

Designing high-quality learning activities that are pedagogically sound and suitable for a particular context and group of learners takes substantial time, skills, and effort, as well as technological, pedagogical, and disciplinary content expertise (Yeh, Chan, and Hsu 2021). It is often argued that LD is both a science and an art (Drugova et al. 2023; Misiejuk et al. 2023). GenAI could make, or even provide early drafts of, some of the design, co-creation, and collation of these activities perhaps faster and easier. Although there are currently several approaches to designing short tasks or assessment activities using GenAI (Kasneci et al. 2023), designing sequences of teaching and learning activities that are coherent, meaningful, and appropriate for a given context might still require human expertise.

3.4.2. GenAI and glocalization

Although GenAI is being adopted in different languages and approaches (Yang et al. 2023), it still needs to consider glocalization issues (i.e. how to make content locally/contextually relevant). For example, the role of expertise, quality assurance, the rapidly changing policy environment, and the focus on (automatic) content rather than (automatic) pedagogy might require GenAI approaches to find appropriate ways of supporting the learning and teaching of diverse learners and educators (Rizvi et al. 2022). For example, recently the UK Government (2023) has provided strict guidelines that 'Schools and colleges should not allow or cause intellectual property, including pupils' work, to be used to train GenAI models, without appropriate consent or exemption to copyright.' Other national or regional governments, as well as publishers, might have different takes on how educators can use GenAI in practice, as this is an evolving narrative.

3.4.3. GenAI and the reuse of commercially sensitive content

Although educational institutions might be attracted to the notions of free or easy-to-use templates for LD and automatic content creation, there are obvious risks in terms of sharing commercially sensitive content with GenAI. For example, uploading one's carefully designed e-course on ML for undergraduate computer scientists to a GenAI platform could allow GenAI programmes to provide variations of that course back to the end user. At the same time, GenAI programmes might reuse and re-create this course for other institutions and other commercial enterprises. Some educators might welcome this open educational resource philosophy, but, at present, educators are not in control of who, how, and when other users can use and reuse their carefully designed content. Potentially, this could infringe copyright law or government guidance, as indicated by the UK Government (2023) example.

3.4.4. GenAI and authentic learning

Perhaps most importantly, though current GenAI approaches are very useful and powerful for generating written artifacts based on the current body of knowledge, for decades educational researchers have argued and found that deep and authentic learning is more than just the assimilation of written texts and artifacts (Kirschner and van Merriënboer 2013; Nguyen, Rienties, and Richardson 2020; Winne 2017). In particular, a range of educational models have highlighted that working on authentic tasks with others is essential for establishing deep and complex learning opportunities for learners. The verdict is still out on whether GenAI can be used to design, implement, and critically evaluate automatic content for course LDs.

3.5. Contribution 5: a human-centred perspective to GenAI and analytics layers in learning design by Davinia Hernández-Leo and Yannis Dimitriadis

LD is a field that has attracted much attention in recent decades, especially in the learning technology context (Michos and Hernández-Leo 2020), as it provides methods and tools that support the creative process of multiple stakeholders in designing for learning. However, it is highly complex and time consuming to make pedagogically informed decisions regarding the design of tasks to be undertaken, the tools and resources to be used, and the social environment in which students may learn (Goodyear, Carvalho, and Yeoman 2021). Due to this complexity, LD methods and tools have not been widely adopted, despite the high relevance of the field (Dagnino et al. 2018). The question addressed in this contribution is whether there can possibly be a function – or multiple functions – for GenAI to reduce its complexity.

In the Learning Design Life Cycle (Asensio-Pérez et al. 2017), multiple stakeholders (mainly teachers, but also instructional designers, learners, and even learning scientists) collaborate toward a pedagogically sound LD, subject to the constraints of the educational context and the stakeholders' technological pedagogical content knowledge (TPACK). Stakeholders are involved in the creative inquiry process of co-designing for learning, generating design artifacts that evolve over time and feed the successive phases. Moreover, learning, design, and (teacher) community analytics have been shown to provide relevant data-based evidence that supports the LD process (Hernández-Leo et al. 2019).

Given the high relevance and intrinsic complexity of the LD Life Cycle, LLMs and the derived conversational agents have a high potential for enhancing each phase of this human-centred creative design process (Demetriadis and Dimitriadis 2023) – a process that is intrinsically centred on humans, as it is under the responsibility of the stakeholders, and the needs of the stakeholders must be considered. This potential is increasing, as the use of LLMs can be integrated with other types of GenAI and tools to enable conversational browsing, analysis, or illustration. An analysis of the LD process and its connection to learning analytics through the lens of the affordances of GenAI unveils new opportunities, around facilitating a deep dive into the educational context, promotes brainstorming, and aids in crystallizing the envisioned solution or enriching actionable analytics indicators to improve LD. We formulate these opportunities as speculative functions in which GenAI has the potential to support and mediate the data-driven support of LD phases (Figure 1, Table 2).

These speculative functions can be illustrated through the following generic case: A course needs to be redesigned, and the teaching team invites former students of the course to a co-design session with the aim of improving the course LD. They prompt GenAI,

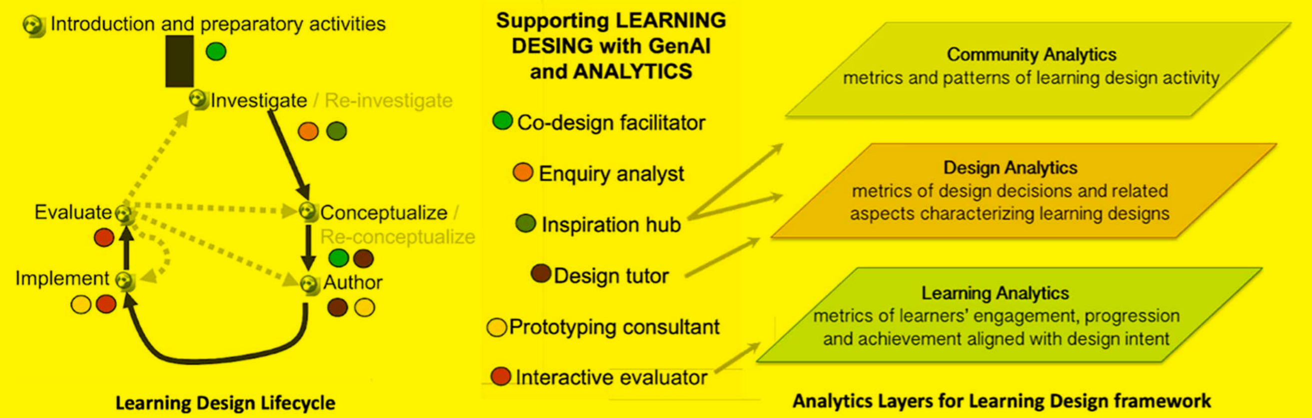

Figure 1. Speculative functions in which GenAI integrated with analytics layers (Hernández-Leo et al. 2019) may support the Learning Design Life Cycle (Asensio-Pérez et al. 2017).

acting as a **co-design facilitator**, to conduct a role-play discussion (Sharples 2023) with adaptive stimulus questions guiding problem identification (Hernández-Leo et al. 2017). The questions guide the stakeholders' exposition, integration, and summarisation of issues to help select the main redesign problem to tackle. In the process, while being assisted by GenAI functions, the stakeholders need to mindfully consider the core limitations of this technology (e.g. hallucinations and bias) by carefully ensuring that the produced outcomes actually reflect the essence of the discussion. Once the problem is identified, the team uses a GenAI **enquiry analyst** to elaborate on the problem and how it relates to the contextual facets of the educational situation. The GenAI may elaborate several options of, for example, why the problem may be due to specific characteristics of the students in this context and may also generate descriptions about specific needs. These options can be used as a starting point, when meaningful to the design team, the team can iterate those that are closer to their own assessment to further investigate and analyze their needs. Once the problem and needs are clear, the team may request ideas from the **inspiration hub** concerning how to tackle the problem. The inspiration hub offers summaries of results in the conversational browsing of relevant LDs shared in a community platform (Gutiérrez-Páez et al. 2023) or the generation of text by an LLM trained using those LDs, including the extraction of patterns in those designs (Ljubojevic and Laurillard 2011). The team can then critically select ideas and approaches for solving the design problem and use the **co-design facilitator** again to guide negotiations and a collaborative knowledge-building approach.

Table 2. Description of speculative functions in which GenAI and analytics layers may support LD.

Co-design facilitator	Assist in the collection, integration, and summarisation of inputs from several stakeholders (i.e. their views on design problems in *preparatory activities*, and their views on design proposals during *conceptualization).*
Enquiry analyst	Help analyze enquiries for better clarity (e.g. understanding the context of the learning situation in the *investigation* stage).
Inspiration hub	Guide the exploration of design problems during the *investigate* stage (e.g. based on analysis of data extracted from an LD community platform [*community analytics*] and the LD patterns extracted in available previous and relevant designs [*design analytics*]).
Design tutor	Provide timely feedback on progress while *conceptualizing* and *authoring* a design based on the *design analytics* of the LD being created.
Prototyping consultant	Advise in the process of prototyping the LD beginning from its *authoring* to its *implementation*, including revising text, taking care of visual (e.g. facilitating inclusion) and technical coding, as well as formatting (e.g. for deployment in virtual learning environments).
Interactive evaluator	Offer analytics of students' data for interactive exploration during ongoing *implementation* and the *evaluation* of the implementation. This enhances the ease of interpretation and actionability of the learning analytics while teachers orchestrate the learning scenario (as a runtime 'orchestration partner') and later plan the scenario redesign for future cycle iterations (as a 'reflection assistant').

In the actual process of describing the approach and producing the detailed task descriptions and materials, as well as making decisions on space, tools, and social facets (Goodyear, Carvalho, and Yeoman 2021), the **design tutor** may generate immediate feedback about how to increase the potential pedagogical rigour and quality of the designs, detecting unclear tasks descriptions, unbalanced consideration of content, or learning methods (Albó et al. 2022). Again, the team – who is aware of the limitations of the GenAI behind the supporting tools – critically considers the feedback to improve the design when applicable and considers professional (human) support when needed. In the authoring process, a **prototyping consultant** would be able to help in finalising design creation, revising text, enriching with proposed images, and adapting to several formats considering universal design for learning principles. A GenAI prototyping consultant has also the potential to help in design coding (Ebert and Louridas 2023) in terms of the technical languages and formats of virtual learning environments for their interoperable deployment across platforms (Prieto et al. 2013). Finally, a GenAI conversational **interactive evaluator** could support teachers while implementing the learning scenario, prompting different analyses and human-readable explanations (Amarasinghe et al. 2023; Susnjak 2023) of student progress until a good understanding of their current knowledge is achieved (in alignment with their own criteria) to identify the correct interventions and feedback on the fly. The **interactive evaluator** is also useful once the implementation has finished to further understand the impact of the implementation and inform future redesigns of the same activity or the designs of forthcoming activities for the same cohort of students (Amarasinghe et al. 2022).

We are aware that the formulation of these speculative functions offers an optimistic perspective about the potential opportunities. Yet, as mentioned in the example, the users of the functions must be aware of the core limitations of the supporting technology (i.e. GenAI). Moreover, there exist multiple challenges that must be addressed with respect to the use of GenAI in LD (see also Contribution 4). However, this viewpoint is shaped by the rapidly changing nature of GenAI, marked by the continual introduction of new models, versions, and tools, as well as the integration between tools and the increasing facility of intervening in their training. The view is coupled with recent research into their capabilities and the knowledge about how advances in educational technology have been seeking to aid in LD. Yet, future research in the field must assess the possibilities and limits of the proposed functions. If these functions are to be offered to support actual practice, the LD tools should be transparent about their limitations and respect human centrality and agency in the design process (Hernández-Leo 2022). Human-centred AI (Shneiderman 2020), human – AI collaboration (Akata et al. 2020), and hybrid intelligence (Holstein, Aleven, and Rummel 2020) are probably the most relevant pillars to ensure that GenAI may serve as a productive and ethical companion augmenting, rather than replacing, the human intelligence of the stakeholders.

3.6. Contribution 6: the use of LLMs in learning diagnosis and feedback content generation by Mutlu Cukurova

This section explores the potential and limitations of LLMs in diagnosing students' learning challenges and appropriate feedback generation. LLMs have demonstrated promise in students' discourse analysis to be able to accurately detect student challenges and potential misconceptions, as well as use this information to generate appropriate feedback (Suraworachet, Seon, and Cukurova 2024). However, they also have some significant limitations making them unlikely to be the panacea for major challenges of AIED.

Recent advancements in LLMs have marked significant evolution in AI, demonstrating that these models can not only pass significant human professional exams but also outperform human counterparts in some instances. Notably, advanced LLMs have surpassed most law school graduates on the bar exam (Katz et al. 2023), successfully completed medical, law, and business exams (Achiam et al. 2023), and have even excelled in a challenging US medical licensing exam, albeit not yet at the level of human doctors (Brin et al. 2023). These are significant achievements for the state of the art in AI. However, for AIED research and practice, our goal is not necessarily to improve the state of the art in AI or to build optimal AIs to pass our existing frequently criticised exams, but to build AI systems that would support human learning. The first outcomes might be considered as initial steps toward the main goal. Nevertheless, however impressive these achievements of LLMs are, they do not mean much for AIED research unless the impact on the latter goal is evaluated and evidenced.

What is potentially more meaningful for AIED, is the use of LLMs as diagnostic tools to detect students' knowledge gaps and challenges, as well as generate relevant feedback to support their learning. Indeed, LLMs provide certain advantages on these fronts as well. For instance, in our recent work, we found that advanced LLMs (GPT4) can perform comparably well to traditional NLP approaches (e.g. support vector machines and random forest algorithms with feature engineering) in detecting and identifying student challenges in their discourse without the need for resource – and time-intensive feature engineering work and model training expertise (Suraworachet, Seon, and Cukurova 2024). We also tested LLM models to create relevant feedback for students in higher education social science contexts and found out that the models have some potential to generate relevant feedback that is likely to have a positive impact on students' learning (Leiker et al. 2023a). However, this is yet to be evidenced at a scale with longitudinal impact evaluations. Expanding feedback generation to further content generation for adult learning in learning management systems and to other modalities with multimodal foundation models can also generate learning materials, and this has the potential to be as effective as traditionally produced expert content (Leiker et al. 2023b). Despite these promising findings, unless significant resources, time, effort, and data are used to patch them, the lack of stable world models in transformer-based LLMs limits their ability to reason reliably and plan effectively, making their performance in learning research and practice inconsistent. Therefore, at these early stages of development, caution is required in real-world implementations at scale. It is our responsibility to validate any LLM-based educational interventions before their release in the real world. We should not be mesmerised and stop at evaluations of the human likeness of LLMs' outputs. We need to consider the advantages and disadvantages of the various LLMs for teaching and learning, particularly for their potential to provide meaningful feedback to students and teachers to close the feedback loop. We do not have to be particularly impressed with the evaluations of the models' performance on existing exams for humans. Rather, we need long-term evaluations of the impact of the use of LLMs in diagnosing students' learning challenges, as well as the impact of their feedback on students' learning outcomes and competencies. Supporting students in their learning involves more than just providing correct responses to questions for passing exams. Rather, it requires motivating learners to engage with the feedback in the first place, sustaining learners' engagement with the feedback, and ensuring support is aligned with the learners' affective and metacognitive states; it also requires an understanding of the context in which the learning takes place and the nurturing of key thinking skills among learners so that they do not only learn the answers to the questions posed but also progress toward learning how to learn. Although AI excelling in exams demonstrates computational prowess, this does not necessarily equate to a capacity to independently support broader educational objectives (Bulathwela et al. 2024). This observation does not diminish AI's potential role in these areas but, rather, emphasises that this role itself should be evaluated before we get too excited about the real-world impact on education. In addition, AI does not have to support all aspects of teaching on its own, and some of these goals require hybrid intelligence approaches, where these broader objectives are met by systems that synergistically combine the complementary strengths of humans and AI (Cukurova 2024).

On this front, there might actually be some disadvantages of foundational LLMs compared with more traditional rule-based and supervised ML models. For instance, in the latter modelling approaches, feature importance and keyword dictionaries could provide evidence of students' challenges and potential solution examples in the form of lists of keywords, word clouds, and so on, which could be clearly understood by learners and teachers. Not only could this potentially promote trustworthiness by helping learners and educators understand the rationales behind the models, but it could also provide an additional layer of visualising the student's ideas and help identify common challenges. For instance, this would be particularly useful for helping teachers decide whether to give whole-class feedback (if the same issues emerge for multiple students) rather than trying to provide individualised support to each group to address a commonly faced challenge in practice. Previous work has indicated that, when people are presented with content framed as coming from AI, they tend to judge it as less credible compared with the same content framed as products of educational psychology or neuroscience due to their mistrust of AI (Cukurova, Luckin, and Kent 2020). Similar results may be observed when AI-generated content or feedback is presented to students and teachers, and they might judge its quality lower if they know that it is AI-generated content.

Transformer-based LLMs can also provide evidence of their decision-making process regarding student learning gaps with a stochastic model for word prediction when prompted. Currently, these outputs on diagnosis and feedback suggestions are generated based on the prompt the users give it; the most likely next word is based on its training data and a random element, as well as controlled patch training with reinforcement learning with human feedback. For stochastic LLMs, unlike deterministic rule-based approaches or other modelling techniques, the same prompts posed might indeed lead to different outputs. This is why tools such as ChatGPT are sometimes referred to as 'stochastic parrots' (Bender et al. 2021) – stochastic in that they generate content based on probability analysis, and parrot because they do not necessarily have any understanding of the meaning of anything they generate. Given this 'black box' and probability-based nature of LLM decision-making processes, users would be less likely to legitimize or rationalise the model's outputs, potentially lowering the trustworthiness of the system, particularly when it makes a mistake (e.g. the LLM fails to perceive the students' challenges accurately or generates inappropriate feedback). Preliminary research on AIED has already indicated that teachers have unrealistic expectations of AI, specifically that its recommendations should always be accurate (Nazaretsky et al. 2021).

Surely, a rule-based and supervised ML approach requires higher development time – and domain-specific expert knowledge to engineer features for the models to achieve satisfactory models. In contrast, 'off-the-shelf' LLMs, which have been pre-trained, require only minimal effort to achieve similar/better results, and, thus, their development time is notably shorter. This also links to another key advantage of LLMs, which is their high accessibility. LLMs require no prior background in programming to create a model, as the prompt is written in human language, whereas the rule-based and supervised ML approaches require designers to construct a model using a programming language. In addition, one main drawback of the traditional approaches in general is in the domain specificity of the lexical corpus. In other words, traditional models tend to be tied to the terminologies presented in the contexts, resulting in low model generalizability. However, LLMs gain advantages over this issue through the utilisation of a 'large' corpus to pre-train the model, assuming that this will mitigate the model generalisation problem. Studies have shown that LLMs have high applicability to perform tasks in a wide range of domains. Hence, LLMs have a higher capability to be generalised into other contexts than the traditional rule-based and ML approaches.

Obviously, all modelling approaches, including LLMs, have certain advantages and disadvantages. The decision of which models should be used predominantly depends on the use case, goals, and expertise of the designers and users, as well as on multiple social and ethical considerations. For example, on the one hand, a novice teacher with no programming background who wants to set up a short-term analytics system to assist students during collaboration could consider constructing prompts for LLMs to perform the task. On the other hand, an expert in programming who wants to deploy long-term learning analytics to study students' struggling moments and provide feedback on their patterns might consider deploying the rule-based or ML approaches. It is also essential to highlight that this is not a mutually exclusive approach where users have to select one method, but, rather, they can experiment with different approaches and justify what is best for which task to further complement the model advantages in particular settings. However, consideration of LLMs as the panacea to all challenges of the field is oversimplistic, and overemphasis on one particular approach is likely to lead to stagnation in progress. It is important to remember that our goal is to create AI that supports human learners to make them more competent, not to construct the best AI students (i.e. AI capable of excelling in our exams). In our journey toward this goal, multiple AI techniques are beneficial to the toolkit of AIED researchers and practitioners. LLMs are a valuable asset and tend to outperform more traditional modelling techniques in learning diagnosis and feedback content generation, but are unlikely to be the only approach used in the future.

3.7. Contribution 7: leveraging AIED foundations in the age of GenAI: the case of mathematics education by Manolis Mavrikis

3.7.1. Introduction

This contribution emerged from an observation that the at-scale availability of GenAI surfaced discussions about applications of AI in Education that seem, at best, to reinvent the wheel and, at worst, to overlook or even undermine years of foundational research in the field of AIED. Hastily adopting GenAI technologies in educational settings raises concerns about the efficacy and pedagogical soundness of such approaches. The objective of this section is to motivate leveraging the rich foundations of AIED to inform and enhance the integration of GenAI. To provide a concrete context, we focus on the case of mathematics education in K-12. We discuss how lessons learned from AIED and hybrid approaches of 'traditional' AI and GenAI have the potential to not only address the limitations inherent in, for example, current LLMs but also enrich the pedagogical strategies employed to support mathematics learning.

3.7.2. Learning from AIED

A short section cannot do justice to the extensive history of research under the umbrella of AIED. Regardless, we refer the reader to reviews by McCalla (2023) and Mavrikis et al. (2021). In brief, and narrowing the lens to mathematics education, past AIED research has primarily focused on the design, development, and evaluation of systems designed to enable adaptive learning experiences for learners (Aleven et al. 2016; Koedinger et al. 1997) and support systems for teachers (Holstein, McLaren, and Aleven 2017; Mavrikis et al. 2019). Among the most frequently used approaches in the field, as highlighted by Aleven et al. (2023), we focus on three relevant to mathematics education: tutored problem solving, OLMs, and support for exploratory learning. We explore what these approaches offer and how GenAI can be integrated.

First, **tutored problem solving** refers to a variety of techniques that aim to guide the learner through a problem-solving process by providing real-time adaptive feedback and scaffolding. A prerequisite for achieving this is the monitoring of students' progress in relation to specific 'knowledge components' (Aleven and Koedinger 2013). This allows either selecting specific problems for 'deliberate practice' (Ericsson, Krampe, and Tesch-Römer 1993) or enables breaking down problems and providing feedback in specific steps during the problem-solving process (Heffernan, Koedinger, and Razzaq 2008). For the latter in particular, there is a need for the system to have an accurate representation of the solution space of the problem. In complex cases, this can be achieved by solving the problem through, for example, accessing a computer algebra system (e.g. Melis et al. 2001) or even through conventional problem solvers (c.f. Newell, Shaw, and Simon 1959). These various approaches to accurately representing the solution space can offer a potential solution to one of the key limitations of GenAI models: potentially generating incorrect solutions. More importantly, they offer a way for the system to provide precise and pedagogically sound feedback rather than a solution. An example where such a hybrid approach shows promise, is a recent pilot we are undertaking that involves the integration of Wolfram Alpha's computational engine with the natural language interaction provided by OpenAI's GPT-3 model within a single notebook. Rather than directly providing the steps to solve the problem, the notebooks are designed based on research in mathematics education about mathematical modelling (Blum and Leiß 2007), resulting in a prototype support system that guides students through a structured process of mathematical problem solving (Mavrikis et al. 2024).

Second, **OLMs** (Bull and Kay 2010) and similar approaches are designed to 'open' (usually in the form of visualisations) a system's representation of a student's learning state for a student or teacher to scrutinise (Bodily et al. 2018), and they have been studied extensively in AIED. Such approaches have the potential to address criticisms related to a lack of self-regulation when learning with AI (Molenaar et al. 2019). In the age of GenAI, employing OLMs in appropriate ways has the potential to provide learners and educators with a transparent view of a learner's progress (Conati, Porayska-Pomsta, and Mavrikis 2018). For instance, consider the scenario above where students engage in mathematical problem solving aided by a GenAI-driven tutoring system, but their interaction is closely mapped to a learner model. In this case, the transparency offered by the OLM has the potential to encourage them to take more control over their educational journey. At the same time, their teachers can benefit from access to OLM. By including a question-driven design dashboard (Pozdniakov et al. 2022), the OLM can be queried in natural language to ensure an intuitive and accessible user experience for teachers.

Lastly, an important subset of AIED research focuses on open-ended or **exploratory learning environments** and follows a constructivist approach infused with AI to engage with concepts in a self-directed manner; the focus is on problem solving within simulations and virtual labs (Smetana and Bell 2012), microworlds (Mavrikis et al. 2013), or other contexts (Hannafin 1995). In general, such environments prioritise inquiry-based learning, allowing students to formulate their own questions and hypotheses while benefiting from real-time feedback and potential collaborative experiences. AI approaches have been effectively used to provide support in such environments (Gutierrez-Santos, Mavrikis, and Magoulas 2012), and previous research has suggested that combining guided exploration with tutored problem-solving tasks supports students' understanding in both conceptual and procedural knowledge just as effectively as those who engage in tutored problem-solving alone (Mavrikis et al. 2022). Integrating GenAI with such environments offers a pathway to enhance their exploratory and inquiry-based nature. For instance, apart from facilitating natural language interaction, complex scenarios, examples, or questions and hypotheses can be dynamically generated. Combined with learner modelling and tutoring approaches as mentioned above, these scenarios can be tailored to the learners' interests, curiosity, and skills.

3.7.3. Conclusion

This section looked into how the rise in interest in applying AI in education can benefit from the rich foundations of AIED. We only looked into three approaches focusing on mathematics education due to space limitations, but the field of AIED has several other methodologies and frameworks that could also offer valuable contributions when integrated with GenAI technologies (du Boulay, Mitrovic, and Yacef 2023). The approaches discussed have broader applicability and could inform AI applications in various other educational contexts.

4. Prospects and implications of GenAI in learning technology research and practice

In the following subsections, we summarise the emerging themes.

4.1. Utilising GenAI to learning design

LD refers to the process of designing effective learning experiences (Mangaroska and Giannakos 2019), which often requires the use of technological innovations and consists of substantial amounts of written texts, digital artifacts, and different learning materials. LD defines the learning objectives and pedagogical approaches that educators can reflect upon to make decisions and improvements. LD has been described as the ‘application of methods, resources, and theoretical frameworks to achieve a particular pedagogical goal in a given context’ (Mor and Craft 2012, 88). Effective LD is a cornerstone for traditional, online, and blended learning settings. It has been recognised as a key factor in any learning activity’s success and a major driver of the learning experience (Nguyen, Rienties, and Whitelock 2022). In recent years, LD has gained momentum due to its critical role in creating online and blended courses, as well as in the corporate training space.

Although most contributions in this article touch upon several LD aspects, Contributions 4 and 5 focus on the importance of LD and potential opportunities and implications for teaching and learning. First, it is important to highlight that, with the introduction of GenAI capabilities, LD will neither lose its importance nor its central goals (designing the learning experience with the learner in mind). Defining the key learning outcomes, creating appropriate learning materials, and orchestrating the learning approach will continue to be at the epicentre of LD. At the same time, as we can see a more detail account Contributions 4 and 5, LD needs to embrace GenAI capabilities, and it is likely to automate some LD routines (see initiatives from the Open University UK and companies such as MagicSchool and Eduaide); moreover, additional roles and processes will be required (see Contribution 5). In particular, GenAI could create some of the learning activities, probably faster than traditional methods. Although some of these LDs will be of high quality, it is almost certain that human expertise will be required. Accordingly, the craft of LD is likely to evolve with some of the functions becoming obsolete (e.g. early drafts of learning materials), whereas others will become enhanced or new ones will emerge (e.g. glocalization of the LD, human – GenAI co-creation, and development of deep and complex learning opportunities).

The aforementioned opportunities indicate how GenAI can support LD. For example, with affordances such as automatic content creation, content co-creation, and a plurality of LD we can develop quick and personalised learning activities. At the same time, there are also substantial challenges associated with these affordances. These include concerns about data privacy and security when training GenAI, as well as about intellectual property rights, including publishers’ and pupils’ work. Moreover, we are aware that deep and authentic learning is more than just the assimilation of written texts and artifacts and that models or decisions working in one context might not necessarily work in a different one. Therefore, the importance (and even burden) of proper contextualisation and quality assurance will still lay on human experts, and they may become more important than ever in the future. Moreover, the lifecycle of LD will greatly benefit from data derived through design, community, and learning analytics in conjunction with the affordances of human – GenAI integration. However, the verdict is still out on whether GenAI could be used to design, implement, and critically evaluate automatic content for complete LDs. We anticipate that GenAI can be used in conjunction with human experts to augment rather than replace human intelligence, and, in this context, GenAI will serve as a productive and ethical companion to human expertise.

4.2. Regulation of learning and GenAI

SRL refers to one’s ability to understand and control their learning progress. With the rise of GenAI, it has become more vital than ever for learners to be active participants, and it is critical to understand how they can expand their expertise and agency alongside the use of AI. We argue that GenAI has the potential to empower learners’ SRL and SSRL processes in new ways. For instance, such enhancement can utilise AI algorithms to detect learners’ progress and adjust the learning material accordingly. This approach can become particularly powerful if one considers the amount of multimodal data and the adaptivity capabilities of advanced technologies (immersive, game-based). These technological and data capabilities not only improve researchers’ understanding of the SRL processes, but they can also be utilised to augment and empower peers, learners, teachers, and AI agents to increase their awareness, monitoring, regulation, and reflection of SRL. Although early results clearly indicate this promise (see Contribution 2), future work is needed focusing on GenAI’s capabilities toward the development of novel theoretical and empirical grounds.

There are a number of challenges associated with the research and practice of SRL and GenAI. Some of these are long-standing challenges, whose importance has significantly increased in recent years. For instance, accurate measurement of SRL has been a challenge for decades, and, despite GenAI’s capabilities to improve this, the stochastic nature of GenAI models has raised significant concerns about reliability (see Contribution 6). Another noteworthy challenge is associated with aligning the various analytics (especially when multimodal data come into play) for designing appropriate affordances (e.g. visualisations, adaptations, and dashboards) to support this new form of co-regulation (human – GenAI regulation). Another challenge raised by GenAI models has to do with the way we used to evaluate and validate effects on SRL skills. Taking into account one of the most prominent approaches in learning technology (i.e. design-based research), it becomes clear that the usual evaluation cycle (identifying evidence of their effectiveness [or lack thereof] and progressively revising the tool/practice) is strongly challenged due to the difficulties in identifying the shortcomings of the different GenAI models. Therefore, though GenAI presents opportunities in conceptual, theoretical, methodological, analytical, and educational issues, our interdisciplinary research community must still tackle various challenges (Azevedo and Wiedbusch 2023; Järvelä, Nguyen, and Hadwin 2023). Thus, future work should focus on the development of theory-grounded and empirically driven approaches that can overcome the aforementioned challenges and utilise GenAI to trigger, induce, support, and foster effective ways of empowering students and educators to regulate their learning and teaching.

4.3. Automated content

Educational institutions and professionals might be attracted to the notions of free or cheaper automatically created learning materials, especially due to the current developments of automated content generation and the fact that there are areas such as math and programming (Contributions 3 and 7) where GenAI is producing effective learning materials. Despite the potential of automated content (lower cost and increased content), there are also obvious risks associated with it, as elaborated in Contributions 4 and 5. In particular, GenAI-generated content is likely to encounter glocalization issues (i.e. lack of contextually relevant content) and difficulties in achieving authentic learning, and this limited reliability makes LLMs' performance in learning inconsistent. Moreover, there are challenges associated with potential infringement with copyright laws, university policies, or government guidance.

Given the potential for automated content and the different challenges associated with its use, an important question is whether (and, if so, how) GenAI's capabilities for automated content creation are going to be utilised to support teaching and learning. This question heavily relies on the role of the teacher in adopting such practices and how teachers can work together with those tools to further develop their learning materials and teaching practices. Another important question is whether such resources will be used as central components of teaching or simply to complement it. Hence, for the efficient utilisation of GenAI content, new roles, competencies, and processes are likely needed (see also Contribution 5). Similar to the other LD tasks, we are confident that GenAI can be used in conjunction with human experts and relevant data analytics to enhance human expertise insofar as appropriate policies, practices, and professional support are in place and available to educators and students.

4.4. New skills and competencies

The use of technology is, by definition, disruptive, enabling scientific knowledge to support the achievement of the practical goals of human life. When it comes to teaching and learning, during the last century, digital technologies have advanced human learning (e.g. Pressey 1926). In this time, necessary skills, competencies, and jobs, as well as the 'technologies' for teaching and learning, have been continuously changing. In the last 20 years, we can see how several developments have changed the way humans learn. For instance, the inception of advanced digital libraries and online learning has allowed, more or less, everyone access to information, and the inception of video providers such as YouTube has allowed everyone to access videos and lectures about almost everything. Even during the last decade, the advancements of open online courses and analytics have provided tremendous opportunities and changed the way teaching and learning occur in several spaces. Therefore, digital technologies have offered several opportunities, disrupted our practice, and forced us to update our skills and competencies (as both teachers and learners) several times in the past.

Today, with the inception of GenAI and powerful tools such as ChatGPT, we see a disruption of various teaching and learning practices (see Contribution 1). A volume of early studies and media outlets have reported the advantages and best practices of ChatGPT in education (Baidoo-Anu and Owusu Ansah 2023; Kasneci et al. 2023). Moreover, we see effective automated generation of math word problems (Wang, Lan, and Baraniuk 2021) and programming problems (see Contribution 3), as well as promising uses for course design (Contributions 4 and 5). The disruption caused by GenAI challenges established assumptions about the skills required and the way teaching and learning will function in the near future. It is indeed likely to see changes in how teachers design their courses (Contributions 4 and 5) and utilise automated learning materials (Contributions 3 and 7), as well as how students regulate their learning (Contribution 2) and are assessed (Contribution 6). Such shifts have the potential to increase efficiency, but this will require the development of different skills to ensure that GenAI capabilities will support human learning instead of simply gaming the learning process and contributing to the degradation of important human skills.

Therefore, focused work is needed to understand what the new reality of education will look like and what skills will be needed (to improve and not hinder human learning). This will allow us to develop proper policy, frameworks, and competence opportunities for both teachers and learners. Thus, as GenAI technologies such as LLMs are being established and our understanding and expertise grow, there will be a gradual leaning toward strategic embeddedness, which will eventually contribute toward strategic, efficient, and sustainable adoption in our society.

4.5. Feedback and assessment

GenAI has the capacity to provide personalised feedback to students based on the information provided by students or teachers. A recent literature review on LLMs in education has indicated that assessment and grading (e.g. both formative and summative and taking place in different forms, such as short answer grading, essay grading, subjective question grading, and student self-explanation) are promising areas for GenAI application (LLMs specifically; Yan et al. 2023). In particular, GenAI has demonstrated its ability to accurately detect student challenges and potential misconceptions and to use this information to generate appropriate feedback (Suraworachet, Seon, and Cukurova 2024). Recent research has evaluated LLMs' ability to create relevant feedback for students in different contexts and settings (e.g. Escalante, Pack, and Barrett 2023; Meyer et al. 2023), finding that LLMs have some potential to generate relevant feedback that is likely to have a positive impact on students' learning. In particular, LLMs exhibit impressive performance in typical programming tasks associated with introduction to procedural – and object-oriented programming (Finnie-Ansley et al. 2023). Such tasks are oftentimes used for exams and other assessments of CSEd students and graduates. Along the same lines, we see LLMs passing medical, law, and business exams (see Contribution 6). These early outcomes suggest that GenAI (and LLMs in particular) has a certain value in providing tailored feedback to students and that some rethinking of assessments is essential.

GenAI models can utilise students' input and provide tailored feedback or suggest materials that align with students' learning needs, making them a useful resource for teachers in helping provide personalised feedback for students in a much faster and sometimes more efficient manner. To be sure, the teachers should critically evaluate such feedback, but GenAI can still help them save time and allow them to focus on other important parts of instruction, such as engaging with authentic tasks that are essential for establishing deep and complex learning opportunities for learners. In addition, there is a need for further research into students' and teachers' trust in AI-generated content. Previous research has indicated that AI-framed content is considered less credible (Cukurova, Luckin, and Kent 2020) and that teachers have unrealistic expectations from AI-based educational technology (Nazaretsky et al. 2022). Therefore, we should identify ways to reinforce teacher – AI co-understanding and complementarity, which will help us tackle the increasingly complex demands of upcoming AI-rich settings. Although the potential of GenAI (and LLMs in particular) to support students with relevant feedback at scale is a genuine possibility, we are just scratching the surface of the value of these approaches in real-world educational scenarios.

4.6. Domain-specific use and knowledge transfer

At-scale use of AI to support different levels of education and domains is not a new topic. In today's discussion, we should ground our decisions on years of foundational research in the field of AIED (McCalla 2023). There are particular domains and contexts where AI has shown particular promise to improve learning outcomes, such as AI-assisted tutoring in algebra (see Pane et al. 2014). As elaborated in Contribution 7, domains such as mathematics education have achieved tremendous progress in the past, greatly advancing in a number of areas, such as tutoring, problem solving, OLMs, and open/exploratory learning environments. Therefore, if we want to seriously consider the use of GenAI in education, we should draw from previous AIED experience and examine how the insights gained from the different case studies can be utilised to inform GenAI use in other domains and contexts.

At the same time, we recognise the important recent advancements of GenAI (and LLMs in particular; e.g. Open AI Codex) and the potential impact in different domains. In particular, this commentary also elaborates on how LLMs can support the domain of CSEd, highlighting early success and identifying different challenges (Contribution 3). One important milestone for CSEd is when Open AI Codex was additionally trained on 159 gigabytes of Python code from 54 million GitHub repositories, which allowed GitHub Copilot to demonstrate a remarkable ability to solve programming problems (Chen et al. 2021). This advancement made the CSEd domain unique among other domains in exploring the use of LLMs (as elaborated in Contribution 7). For instance, the use of LLMs has several shortcomings in the educational context, but, in the training context (professional software developers), such problems are overcome by the experience of the professionals (who can assess the outputs of the models), and this lack of reliable outputs can create major problems on novices and students learning. Therefore, the domain should use LLMs cautiously and focus on students' code comprehension and interpretation (instead of merely solving a stated problem). This will equip students with the necessary competence to assess the correctness and relevance of both code suggestions and explanations generated by LLMs.

Although there are opportunities and challenges across domains (e.g. reliability of the outcomes), it is important that domain experts (e.g. experts in CSEd, math education, and language learning) work to identify the proper practices for using GenAI in their respective domains. In this direction, we have already seen a working group in CSEd (Prather et al. 2023). Moreover, it is important to mention that GenAI does not have to support all domains and all aspects of teaching in the same way; some domains and teaching aspects will require hybrid intelligence approaches (e.g. support teacher – AI complementarity; Holstein, McLaren, and Aleven 2019), where these broader domain objectives are met by systems that synergistically combine the complementary strengths of human and machine intelligence.

4.7. Ethical dimensions

Although different universities, countries[7], and international organisations (e.g. ACM) have clear and detailed codes of ethics, it is not always clear what is considered the ethical (or unethical) use of GenAI in education. For instance, in the context of CSEd, Prather et al. (2023) surveyed a large number of instructors and identified major disagreement on what constitutes an unethical use of GenAI tools by students. At the same time, they identified several areas where the instructors agreed that the use of GenAI tools should be allowed or not. For instance, ethical use should be interpreted within the context of use, and generating an entire solution is considered unethical as long as the students lack an understanding of the provided solution. Moreover, they highlighted that it is not unethical to use GenAI tools to generate part of a solution, facilitate code debugging, and enhance the readability of their solution. In summary, the authors concurred that there are situations where GenAI tools can be used in an ethical way to help students (and teachers) save time and improve their solutions, without negatively affecting the learning outcomes; at the same time, the rise of GenAI has raised alarms about academic integrity issues, such as cheating, plagiarism, and falsification.

Another important ethical dimension of the use of GenAI in education has to do with data privacy and security, as student and teacher data are sensitive and there are certain rights and obligations on the ways these data can be used (e.g. universities need to have data processing agreements with the technology providers to clarify such matters). Therefore, the way student and teacher data are going to be used must follow certain security (e.g. on collection and storage) and transparency (e.g. consent from students, parents, or teachers) standards. Another important ethical challenge of the use of GenAI models is the lack of trustworthiness and reliability of their outcomes. GenAI models may indeed perform very well at certain tasks, but their output often contains errors, cites inappropriate or fabricated sources, and, in many cases, provides inaccurate, misleading, and even biased information. Therefore, caution is required concerning how GenAI tools are being used, and human involvement is still needed to verify the trustworthiness of the insights and personalised offers.

5. Concluding thoughts and the way forward

GenAI tools are some of the most transformative tools developed in recent years. The use of such tools in education is a promising area of both research and practice, offering many opportunities to revolutionise different aspects of teaching and learning. At the same time, as with all other revolutionary and transformative tools in the history of learning technology, GenAI presents significant challenges for educational institutions, educators, and individual learners. Therefore, the use of GenAI tools in education needs to, first and foremost, be put into practice in ways that follow our values and augment our teaching and learning capacities. However, with the current lack of evidence – and theory-based guidelines and regulations, there is a high likelihood of abuse and misuse of GenAI tools in education.

Therefore, to unleash GenAI's full potential both ethically and responsibly in support of human learning, it is crucial to approach its use with caution and critically evaluate both its strengths and limitations by considering practical, ethical, and policy/legal challenges. Currently, learning technology research is fully absorbed with understanding and empirically evaluating GenAI's capabilities to advance (and hinder) human learning. With several guidance papers and reports from prestigious international organisations and publishers becoming available (e.g. Dwivedi et al. 2023; Kasneci et al. 2023; Miao and Holmes 2023) and an increasing number of empirical results surfacing (e.g. see recent GAIED workshop in NeurIPS: https://

gaied.org/neurips2023/), the research community has started to portray GenAI's best practices for learning, as well as inappropriate practices. One direction that becomes clear is that the best results would emerge with appropriate augmentation and coordination of technological developments of tools such as LLMs with human intelligence. Another focal point of the ongoing discussions is that future work is needed to crystallize further and provide evidence and insights into GenAI's capabilities for human learning. In this direction, we provide five thematic areas where future research is needed to enrich our understanding of the use of GenAI in education; shedding light on these thematic areas will help us establish widely accepted and inclusive practices and accelerate its proper use.

- **The role of human experts:** Regardless of whether humans are involved directly (teachers) or indirectly (designers of an online course) in the learning process, it is important to remember that GenAI cannot fully replace human expertise. As GenAI still lacks nuance, context, and common sense, by keeping human experts in the loop, we can mitigate potential risks and ensure that LDs and instructions are properly contextualised and meaningful. Further work should identify optimal ways for GenAI tools to be used in conjunction with human experts (e.g. teachers or course designers, who modify and approve learning materials generated by GenAI) in ways that a certain level of ownership, agency, and control is maintained by the expert.
- **Strong and continuous evidence:** Although there is evidence that GenAI tools can be used in ways that increase productivity and support learning (e.g. Prather et al. 2023), this is yet to be evidenced at scale with longitudinal impact evaluations and across different contexts and content domains. Therefore, long-term evaluations of the impact of using GenAI on students' learning must be conducted to clarify whether, how, and under what circumstances students' learning outcomes and competencies are impacted over time. Future work should utilise established methods, including learner data and analytics, to develop theory-grounded and empirically driven knowledge and practice; this will help us overcome challenges and identify effective ways of leveraging GenAI to enhance teaching and learning.
- **Design of technology:** GenAI uses algorithms to create new content and make predictions and interfaces to deliver tailored feedback and recommendations to learners. Those interfaces and algorithms should take into consideration and actively reduce potential lack of transparency, accountability, privacy, and fairness, as well as bias (e.g. algorithmic bias and provided information bias). Instead of adopting any technocentric approach, the design of these technological components should always be human-centred and strive for regularly updated, accurate, and open-source models, valuing transparency of data use and allowing further modification/extension. Moreover, researchers and practitioners should always consider ways in which different variants of AI (including non-generative ones) can synergistically combine their strengths to reinforce efficient, ethical, and sustainable use of technology.
- **Policy frame:** There is an acute need for the learning technology community to contribute toward the enactment of new guidelines, regulations, and laws to govern GenAI tools in education. The global nature of GenAI requires international coordination and cooperation to ensure our values are properly guarded and can responsibly maximise the benefits of GenAI in education. In particular, future work should explore how the use of GenAI tools might impact justice, equity, diversity, and inclusion in education, as well as the potential impact on vulnerable, marginalised, and underserved groups.
- **Support and competence development:** As research on GenAI in education is progressing, it becomes clear that there is a set of competencies that are needed to enable individuals such as teachers, students, and parents to critically evaluate GenAI technologies and use them to directly learn/teach – or to use them in one of the many other ways humans learn indirectly (e.g. collaborate, socialise, and work). The development of training and resources for teachers and learners on how to use GenAI can help them interpret its results, as well as fact-check and corroborate the information provided. Therefore, there is a need for additional research investigating what new competencies will be necessary in a future where GenAI has transformed the way that we teach and learn.

We concur with the general skeptical optimism about the use of GenAI tools such as ChatGPT in education. In this commentary, we have highlighted the danger of hastily adopting GenAI tools in education without deep consideration of the efficacy and pedagogical soundness of such practices. We provided seven contributions focusing on certain areas and highlighted seven central learning technology topics that are likely to play a pivotal role in the use of GenAI in education. We emphasise the need for further work in understanding both the opportunities and risks of GenAI to support human learning by providing a research agenda. Such work will not only shed light on effective ways of employing GenAI tools in education but also identify technological and pedagogical frames appropriating GenAI's use in education (or mitigating potential misuse).

Before closing this commentary, it is important to reflect on the contribution and limitations of its content. The manuscript shares the views of nine experts and provides critical reflections on the opportunities, challenges, and implications related to GenAI technologies in the context of learning technologies and education. Based on experts' insights, the manuscript also provides an agenda for future research in the area of GenAI in education. Although the provided perspectives can inform readers about the recent developments and crucial topics of GenAI in education, it is also important to highlight that the commentary presents the authors' viewpoint. It indeed serves as an intellectual exercise to contemplate the potential opportunities and challenges of GenAI in education, and although efforts were made to ensure a certain degree of reliability[8], they may still be influenced by the individual contributors' dispositions and biases. Moreover, it is important to consider that GenAI continues to improve in tools' reasoning and other capabilities, which is likely to affect the educational uses of these tools and can address some of the limitations and critiques highlighted in this manuscript. While today's research, policy, and practice discussions are dominated by LLM technologies, and most of the early works on GenAI in education are based on LLM tools, we currently see a growth of multimodal GenAI technologies (e.g. Gemini) that is likely to bring in new opportunities as well as challenges for teaching and learning.

Notes

1. We use GenAI when referring to statements that are relevant to the general notion of Generative AI tools, while we will be using the term LLM when the statement refers to the subset of GenAI that focuses on producing text, finally we will be using specific tool names (e.g., ChatGPT, Gemini, BERT and GitHub Copilot) when the statement refers to this specific tool.
2. https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/
3. For Norway, see https://sikt.no/tjenester/sikt-ki-chat
4. https://rm.coe.int/regulating-artificial-intelligence-in-education-26th-session-council-o/1680ac9b7c
5. https://library.educause.edu/resources/2021/2/horizon-reports
6. https://www.open.ac.uk/blogs/innovating/
7. EU guidelines on ethics in AI: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf
8. the degree to which members of a designated community concur on interpretations (Krippendorff 2018).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Sanna Järvelä http:orcid.org/0000-0001-6223-3668

References

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, et al. 2023. "GPT-4 Technical Report." https://arxiv.org/abs/2303.08774.

Akata, Z., D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, et al. 2020. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence." *Computer* 53 (8): 18–28. https://doi.org/10.1109/MC.2020.2996587.

Albó, L., J. Barria-Pineda, P. Brusilovsky, and D. Hernández-Leo. 2022. "Knowledge-based Design Analytics for Authoring Courses with Smart Learning Content." *International Journal of Artificial Intelligence in Education* 32 (1): 4–27. https://doi.org/10.1007/s40593-021-00253-3.

Albuquerque, J., B. Rienties, and B. Divjak. 2024. *Unpicking the DNA of Learning Design Decisions.*

Aleven, V., and K. R. Koedinger. 2013. "Knowledge Component Approaches to Learner Modeling." In *Design Recommendations for Adaptive Intelligent Tutoring Systems*, Vol. I, edited by R. Sottilare, A. Graesser, X. Hu, and H. Holden, 165–182. Orlando, FL: US Army Research Laboratory.

Aleven, V., M. Mavrikis, B. M. McLaren, H. A. Nguyen, J. K. Olsen, and N. Rummel. 2023. "Six Instructional Approaches Supported in AIED Systems." In *Handbook of Artificial Intelligence in Education*, edited by B. du Boulay, T. Mitrovic, and K. Yacef, 184–228. Cheltenham: Edward Elgar Publishing. Accessed February 13, 2024. https://doi.org/10.4337/9781800375413.00019.

Aleven, V., E. A. McLaughlin, R. A. Glenn, and K. R. Koedinger. 2016. "Instruction Based on Adaptive Learning Technologies." In *Handbook of Research on Learning and Instruction*, edited by R. E. Mayer and P. Alexander, 522–560. Routledge.

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 *European Conference on Technology Enhanced Learning*, edited by O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, and T. Papathoma, 32–43. Cham: Springer Nature Switzerland.

Amarasinghe, I., K. Michos, F. Crespi, and D. Hernández-Leo. 2022. "Learning Analytics Support to Teachers' Design and Orchestrating Tasks." *Journal of Computer Assisted Learning*, https://doi.org/10.1111/jcal.12711.

Asensio-Pérez, J. I., Y. Dimitriadis, F. Pozzi, D. Hernández-Leo, L. P. Prieto, D. Persico, and S. L. Villagrá-Sobrino.

2017. "Towards Teaching as Design: Exploring the Interplay Between Full-Lifecycle Learning Design Tooling and Teacher Professional Development." *Computers & Education* 114:92–116. https://doi.org/10.1016/j.compedu.2017.06.011

Azevedo, R., F. Bouchet, 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* 13:813632. https://doi.org/10.3389/fpsyg.2022.813632.

Azevedo, R., and M. Wiedbusch. 2023. "Theories of Metacognition and Pedagogy Applied in AIED Systems." In *Handbook of Artificial Intelligence in Education*, edited by B. du Boulay, T. Mitrovic, and K. Yacef, 141–173. Cheltenham: Edward Elgar Publishing.

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." SSRN 4337484.

Balaban, I., B. Rienties, and P. H. Winne. 2023. "Information Communication Technology (ICT) and Education." *Applied Sciences* 13 (22): 12318. https://www.mdpi.com/2076-3417/13/22/12318.

Becker, B. A., P. Denny, J. Finnie-Ansley, A. Luxton-Reilly, J. Prather, and E. A. Santos. 2023. "Programming is Hard-Or at Least it Used to Be: Educational Opportunities and Challenges of AI Code Generation." In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 500–506.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bhandari, S., Y. Liu, and Z. A. Pardos. 2023. "Evaluating ChatGPT-generated Textbook Questions using IRT." *Workshop on Generative AI for Education (GAIED) at NeurIPS'23.* https://gaied.org/neurips2023/files/44/44_paper.pdf.

Blum, W., and D. Leiß. 2007. "How do Students' and Teachers Deal with Modelling Problems?" In *Mathematical Modelling: Education, Engineering and Economics*, edited by C. Haines, et al., 222–231. Chichester: Horwood.

Bodily, R., J. Kay, V. Aleven, I. Jivet, D. Davis, F. Xhakaj, and K. Verbert. 2018. "Open Learner Models and Learning Analytics Dashboards: A Systematic Review." In *LAK '18: Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 41–50. New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3170358.3170409.

Bond, M., H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, and G. Siemens. 2024. "A Meta Systematic Review of Artificial Intelligence in Higher Education: A Call for Increased Ethics, Collaboration, and Rigour." *International Journal of Educational Technology in Higher Education* 21 (1): 4. https://doi.org/10.1186/s41239-023-00436-z.

Brin, D., V. Sorin, A. Vaid, A. Soroush, B. S. Glicksberg, A. W. Charney, G. Nadkarni, and E. Klang. 2023. "Comparing ChatGPT and GPT-4 Performance in USMLE Soft Skill Assessments." *Scientific Reports* 13:16492. https://doi.org/10.1038/s41598-023-43436-9.

Bulathwela, S., H. Muse, and E. Yilmaz. 2023. "Scalable Educational Question Generation with pre-Trained Language Models." In *Proceedings of 24th International Conference on Artificial Intelligence in Education, AIED 2023, Part 1*, edited by N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, 327–339. Tokyo: Springer.

Bulathwela, S., M. Pérez-Ortiz, C. Holloway, M. Cukurova, and J. Shawe-Taylor. 2024. "Artificial Intelligence Alone Will not Democratise Education: On Educational Inequality, Techno-Solutionism and Inclusive Tools." *Sustainability* 16 (2): 781. https://doi.org/10.3390/su16020781

Bull, J., K. Bartimote, K. Kitto, B. Kummerfeld, and P. Reimann. 2022. "Enhancing Learning by Open Learner Models (OLM) Driven Data Design." *Computers and Education: Artificial Intelligence* 3:1000069.

Bull, S., and J. Kay. 2010. "Open Learner Models." In *Advances in Intelligent Tutoring Systems. Studies in Computational Intelligence (vol. 308)*, edited by R. Nkambou, J. Bourdeau, and R. Mizoguchi, 302–322. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14363-2_15.

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. Pinto, J. Kaplan, H. Edwards, et al. 2021. Evaluating Large Language Models Trained on Code, Report No. arXiv:cs.LG/2107.03374, arXiv.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M.. 2018. AI in Education Needs Interpretable Machine Learning: Lessons from Open Learner Modelling. In *Workshop on Human Interpretability in Machine Learning*. arXiv preprint arXiv:1807.00154.

Cukurova, M. 2024. The Interplay of Learning, Analytics, and Artificial Intelligence in Education Learning Analytics & Knowledge Conference, Kyoto. March, 18–22.

Cukurova, M., R. Luckin, and C. Kent. 2020. "Impact of an Artificial Intelligence Research Frame on the Perceived Credibility of Educational Research Evidence." *International Journal of Artificial Intelligence in Education* 30 (2): 205–235. https://doi.org/10.1007/s40593-019-00188-w

Dagnino, F. M., Y. A. Dimitriadis, F. Pozzi, J. I. Asensio-Pérez, and B. Rubia-Avi. 2018. "Exploring Teachers' Needs and the Existing Barriers to the Adoption of Learning Design Methods and Tools: A Literature Survey." *British Journal of Educational Technology* 49 (6): 998–1013. https://doi.org/10.1111/bjet.12695

Demetriadis, S., and Y. Dimitriadis. 2023. "Conversational Agents and Language Models That Learn from Human Dialogues to Support Design Thinking." In *International Conference on Intelligent Tutoring Systems*, edited by C. Frasson and P. Mylonas, 691–700. Cham: Springer Nature Switzerland.

Deng, G., Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass. 2023. PentestGPT: An LLM-empowered automatic penetration testing tool. arXiv preprint arXiv:2308.06782.

Department for Education. 2023. "Generative Artificial Intelligence (AI) in Education." https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education.

Drugova, E., I. Zhuravleva, U. Zakharova, and A. Latipov. 2023. “Learning Analytics Driven Improvements in Learning Design in Higher Education: A Systematic Literature Review.” *Journal of Computer Assisted Learning* 40:510–524. https://doi.org/10.1111/jcal.12894.

Duan, Y., J. S. Edwards, and Y. K. Dwivedi. 2019. “Artificial Intelligence for Decision Making in the era of Big Data–Evolution, Challenges and Research Agenda.” *International Journal of Information Management* 48:63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

du Boulay, B., A. Mitrovic, and C. Yacef. 2023. *Handbook of Artificial Intelligence in Education*. Cheltenham: Edward Elgar Publishing. https://doi.org/10.4337/9781800375413.00038.

Dwivedi, Y. K., L. Hughes, A. M. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M. M. Al-Debei, D. Dennehy, et al. 2022. “Metaverse Beyond the Hype: Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy.” *International Journal of Information Management* 66:102542. https://doi.org/10.1016/j.ijinfomgt.2022.102542

Dwivedi, Y. K., N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, et al. 2023. “So What if ChatGPT Wrote it?” Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy.” *International Journal of Information Management* 71:102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Ebert, C., and P. Louridas. 2023. “Generative AI for Software Practitioners.” *IEEE Software* 40 (4): 30–38. https://doi.org/10.1109/MS.2023.3265877

Ericsson, K. A., R. T. Krampe, and C. Tesch-Römer. 1993. “The Role of Deliberate Practice in the Acquisition of Expert Performance.” *Psychological Review* 100 (3): 363–406. https://doi.org/10.1037//0033–295X.100.3.363.

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* 20 (1): 57. https://doi.org/10.1186/s41239-023-00425-2

Finnie-Ansley, J., P. Denny, B. A. Becker, A. Luxton-Reilly, and J. Prather. 2022. “The Robots are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming.” In *Proceedings of Australasian Computing Education Conference*, edited by J. Sheard and P. Denny, 10–19. New York, NY: ACM.

Finnie-Ansley, J., P. Denny, A. Luxton-Reilly, E. A. Santos, J. Prather, and B. A. Becker. 2023. “My AI Wants to Know if This Will Be on the Exam: Testing Openai’s Codex on CS2 Programming Exercises.” In *Proceedings of the 25th Australasian Computing Education Conference,* 97–104.

Gobert, J. D. 2023. “Inq-ITS: Creating Rigorous Assessment and Real-Time Support of Science Learning.” *XRDS: Crossroads, The ACM Magazine for Students* 29 (3): 36–40. https://doi.org/10.1145/3589647

Goodyear, P., L. Carvalho, and P. Yeoman. 2021. “Activity-Centred Analysis and Design (ACAD): Core Purposes, Distinctive Qualities and Current Developments.” *Education Tech Research Dev* 69 (2): 445–464. https://doi.org/10.1007/s11423-020-09926-7.

Gutiérrez-Páez, N. F., P. Santos, D. Hernández-Leo, and M. Carrió. 2023. “A Study of Motivations, Behavior, and Contributions Quality in Online Communities of Teachers: A Data Analytics Approach.” *Computers & Education* 201:104829. https://doi.org/10.1016/j.compedu.2023.104829

Gutierrez-Santos, S., M. Mavrikis, and G. Magoulas. 2012. “A Separation of Concerns for Engineering Intelligent Support for Exploratory Learning Environments.” *Journal of Research and Practice in IT* 44 (3): 347–360. https://search.informit.org/doi/10.3316ielapa.224958418015124.

Hamilton, A., D. Wiliam, and J. Hattie. 2023. “The Future of AI in Education: 13 Things we can do to Minimize the Damage.” https://doi.org/10.35542/osf.io/372vr.

Hannafin, M. J. 1995. “Open-ended Learning Environments: Foundations, Assumptions, and Implications for Automated Design.” In *Perspectives on Automating Instructional Design*, edited by R. Tennyson, 101–129. Berlin: Springer-Verlag.

Hassany, M., P. Brusilovsky, J. Ke, K. Akhuseyinoglu, and A. B. Lekshmi Narayanan. 2023. Authoring Worked Examples for Java Programming with Human-AI Collaboration, Report No. arXiv:2312.02105, arXiv.

Heffernan, N. T., K. R. Koedinger, and L. Razzaq. 2008. “Expanding the Model-Tracing Architecture: A 3rd Generation Intelligent Tutor for Algebra Symbolization.” *International Journal of Artificial Intelligence in Education* 18 (2): 153–117.

Hellas, A., J. Leinonen, S. Sarsa, C. Koutcheme, L. Kujanpää, and J. Sorva. 2023. “Exploring the Responses of Large Language Models to Beginner Programmers’ Help Requests.” In *Proceedings of Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, 93–105. ACM.

Hernández-Leo, D. 2022. “Directions for the Responsible Design and Use of AI by Children and Their Communities: Examples in the Field of Education.” In *Artificial Intelligence and the Rights of the Child: Towards an Integrated Agenda for Research and Policy. EUR 31048 EN*, edited by V. Charisi, S. Chaudron, R. Di Gioia, R. Vuorikari, M. Escobar Planas, J. I. Sanchez Martin, and E. Gomez Gutierrez, 73–74. Sevilla, European Union: Publications Office of the European Union, JRC127564. https://repositori.upf.edu/handle/10230/58975.

Hernández-Leo, D., S. Agostinho, M. Beardsley, S. Bennett, and L. Lockyer. 2017. “Helping Teachers to Think About Their Design Problem: A Pilot Study to Stimulate Design Thinking.” In *EDULEARN17 Proceedings 9th International Conference on Education and New Learning Technologies,* edited by L. Gómez Chova, A. López Martínez, and I. Candel Torres, 5681–5690. Barcelona, Valencia: IATED. July 3–5

Hernández-Leo, D., R. Martinez-Maldonado, A. Pardo, J. A. Muñoz-Cristóbal, and M. J. Rodríguez-Triana. 2019. “Analytics for Learning Design: A Layered Framework and Tools.” *British Journal of Educational Technology* 50 (1): 139–152. https://doi.org/10.1111/bjet.12645

Holstein, K., V. Aleven, and N. Rummel. 2020. “A Conceptual Framework for Human—AI Hybrid Adaptivity in Education.” In *Artificial Intelligence in Education*, edited by I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Vol. 12163, 240–254. Ifrane: Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7_20.

Holstein, K., B. McLaren, and V. Aleven. 2017. "Intelligent Tutors as Teachers' Aides: Exploring Teacher Needs for Real-Time Analytics in Blended Classrooms." In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK 2017)*, 257–266. ACM.

Holstein, K., B. M. McLaren, and V. Aleven. 2019. "Co-designing a Real-time Classroom Orchestration Tool to Support Teacher-AI Complementarity." Grantee Submission.

Järvelä, S., J. Malmberg, M. Sobocinski, and P. Kirschner. 2021. "Metacognition in Collaborative Learning." In *International Handbook of Computer Supported Collaborative Learning*, edited by U. Cress, A. Wise, C. Rosé, and J. Oshima, 281–294. Cham: Springer. https://doi.org/10.1007/978-3-030-65291-3_15.

Järvelä, S., A. Nguyen, and A. F. Hadwin. 2023. "Human and AI Collaboration for Socially Shared Regulation in Learning." *British Journal of Educational Technology* 54 (5): 1057–1076. https://doi.org/10.1111/bjet.13325.

Järvelä, S., A. Nguyen, J. Lämsä, M. Sobocinski, J. Edwards, B. Dang, and R. Whitehead. 2024. "Learning Regulation together with AI Agent – promoting Adaptive Collaborative Learners." *Poster presented at the LAK 2024 Conference, Kyoto*.

Jury, B., A. Lorusso, J. Leinonen, P. Denny, and A. Luxton-Reilly. 2024. "Evaluating LLM-Generated Worked Examples in an Introductory Programming Course." In *Proceedings of the 26th Australasian Computing Education Conference*, 77–86. Association for Computing Machinery.

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103:102274. https://doi.org/10.1016/j.lindif.2023.102274

Katz, D. M., M. J. Bommarito, S. Gao, and P. Arredondo. 2023. "GPT-4 Passes the Bar Exam (March 15, 2023)." SSRN: http://doi.org/10.2139/ssrn.4389233.

Kiesler, N., D. Lohr, and H. Keuning. 2023. "Exploring the Potential of Large Language Models to Generate Formative Programming Feedback." In *IEEE Frontiers in Education Conference (FIE)*, 1–5. IEEE.

Kirschner, P. A., and J. J. G. van Merriënboer. 2013. "Do Learners Really Know Best? Urban Legends in Education." *Educational Psychologist* 48 (3): 169–183. https://doi.org/10.1080/00461520.2013.804395

Kistner, S., K. Rakoczy, B. Otto, C. Dignath-van Ewijk, G. Büttner, and E. Klieme. 2010. "Promotion of Self-Regulated Learning in Classrooms: Investigating Frequency, Quality, and Consequences for Student Performance." *Metacognition and Learning* 5 (2): 157–171. https://doi.org/10.1007/s11409-010-9055-3.

Koedinger, K. R., J. R. Anderson, W. H. Hadley, and M. A. Mark. 1997. "Intelligent Tutoring Goes to School in the big City." *International Journal of Artificial Intelligence in Education* 8:30–43.

Koutcheme, C., S. Sarsa, J. Leinonen, A. Hellas, and P. Denny. 2023. "Automated Program Repair Using Generative Models for Code Infilling." In *Proceedings of 24th International Conference on Artificial Intelligence in Education, AIED 2023, Part 1, Tokyo, Japan*, edited by N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, 798–803. Cham: Springer.

Krippendorff, K. 2018. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

Leiker, D., S. Finnigan, A. R. Gyllen, and M. Cukurova. 2023b. "Prototyping the use of Large Language Models (LLMs) for Adult Learning Content Creation at Scale." https://arxiv.org/abs/2306.01815.

Leiker, D., A. R. Gyllen, I. Eldesouky, and M. Cukurova. 2023a. "Generative AI for Learning: Investigating the Potential of Learning Videos with Synthetic Virtual Instructors." In *International Conference on Artificial Intelligence in Education*, edited by N. Wang, G. Rebolledo-Mendez, and V. Dimitrova. Cham: Springer Nature Switzerland. https://arxiv.org/abs/2304.03784.

Leinonen, J., A. Hellas, S. Sarsa, B. Reeves, P. Denny, J. Prather, and B. A. Becker. 2023. "Using Large Language Models to Enhance Programming Error Messages." In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 563–569. ACM.

Li, Y., D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, et al. 2022. "Competition-level Code Generation with AlphaCode." *Science* 378 (6624): 1092–1097. https://doi.org/10.1126/science.abq1158

Ljubojevic, D., and D. Laurillard. 2011. "Pedagogical Pattern Collector Software Tool." In *Arts and Science of Learning Design 2011 (ASLD 2011) Workshop*. London.

MacNeil, S., A. Tran, A. Hellas, J. Kim, S. Sarsa, P. Denny, S. Bernstein, and J. Leinonen. 2023. "Experiences from Using Code Explanations Generated by Large Language Models in a web Software Development e-Book." In *Proceedings of Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 931–937. ACM.

Malmberg, J., M. Saqr, H. Järvenoja, and S. Järvelä. 2022. "How the Monitoring Events of Individual Students are Associated with Phases of Regulation – A Network Analysis Approach." *Journal of Learning Analytics* 9 (1): 77–92. https://doi.org/10.18608/jla.2022.7429.

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* 12 (4): 516–534. https://doi.org/10.1109/TLT.2018.2868673

Mavrikis, M., M. Cukurova, D. Di Mitri, J. Schneider, and H. Drachsler. 2021. "A Short History, Emerging Challenges and co-Operation Structures for Artificial Intelligence in Education." *Bildung und Erziehung* 74 (3): 249–263. https://doi.org/10.13109/buer.2021.74.3.249.

Mavrikis, M., E. Geraniou, S. Gutierrez Santos, and A. Poulovassilis. 2019. "Intelligent Analysis and Data Visualisation for Teacher Assistance Tools: The Case of Exploratory Learning." *British Journal of Educational Technology* 50 (6): 2920–2942. https://doi.org/10.1111/bjet.12876.

Mavrikis, M., S. Gutierrez-Santos, E. Geraniou, and R. Noss. 2013. "Design Requirements, Student Perception Indicators and Validation Metrics for Intelligent Exploratory Learning Environments." *Personal and Ubiquitous Computing* 17 (8): 1605–1620. https://doi.org/10.1007/s00779-012-0524-3

Mavrikis, M., N. Rummel, M. Wiedmann, K. Loibl, and W. Holmes. 2022. "Combining Exploratory Learning with Structured Practice Educational Technologies to Foster Both Conceptual and Procedural Fractions Knowledge."

Education Tech Research Dev 70 (3): 691–712. https://doi.org/10.1007/s11423-022-10104-0.

Mavrikis, M., T. Thamsir, E. Geraniou, and C. Crisan. 2024. “Learning and Teaching Mathematical Modelling with AI and Computational Tools: Insights from a Pilot with Wolfram notebooks.”

McCalla, G. 2023. “The History of Artificial Intelligence in Education – the First Quarter Century.” In *Handbook of Artificial Intelligence in Education*, edited by B. du Boulay, T. Mitrovic, and K. Yacef. Cheltenham: Edward Elgar Publishing. Accessed February 13, 2024. https://doi.org/10.4337/9781800375413.00010.

Melis, E., E. Andres, G. Goguadse, P. Libbrecht, M. Pollet, and C. Ullrich. 2001. “Activemath: System Description.” In *Artificial Intelligence in Education*, edited by J. D. Moore, C. Redfield, and W. L. Johnson, 580–582. Amsterdam: IOS Press.

Meyer, J., T. Jansen, R. Schiller, L. Liebenow, M. Steinbach, A. Horbach, and J. Fleckenstein. 2023. “Using LLMs to Bring Evidence-Based Feedback Into the Classroom: AI-Generated Feedback Increases Secondary Students’ Text Revision, Motivation, and Positive Emotions.” *Computers and Education: Artificial Intelligence* 6:100199.

Miao, F., and W. Holmes. 2023. “Guidance for Generative AI in Education and Research.” https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

Michos, K., and D. Hernández-Leo. 2020. “CIDA: A Collective Inquiry Framework to Study and Support Teachers as Designers in Technological Environments.” *Computers & Education* 143:103679. https://doi.org/10.1016/j.compedu.2019.103679

Misiejuk, K., I. J. Ness, R. Gray, and B. Wasson. 2023. “Changes in Online Course Designs: Before, During, and After the Pandemic.” *Frontiers in Education* 7. https://doi.org/10.3389/feduc.2022.996006.

Mitrovic, A., and S. Ohlsson. 1999. “Evaluation of a Constraint-Based Tutor for a Database Language.” *International Journal of Artificial Intelligence in Education* 10 (304): 238–256.

Molenaar, I., A. Horvers, S. H. E. Dijkstra, and R. S. Baker. 2019. “Designing Dashboards to Support Learners’ Self-Regulated Learning.” http://hdl.handle.net/2066/201823.

Mor, Y., and B. Craft. 2012. “Learning Design: Reflections Upon the Current Landscape.” *Research in Learning Technology* 20:85–94.

Nazaretsky, T., M. Ariely, M. Cukurova, and G. Alexandron. 2022. “Teachers’ Trust in AI-Powered Educational Technology and a Professional Development Program to Improve it.” *British Journal of Educational Technology* 53 (4): 914–931. https://doi.org/10.1111/bjet.13232

Nazaretsky, T., M. Cukurova, M. Ariely, and G. Alexandron. 2021. “Confirmation Bias and Trust: Human Factors That Influence Teachers’ Attitudes Towards AI-Based Educational Technology.” In *CEUR Workshop Proceedings*, edited by E. Yacobson, T. Nazaretsky, A. M. Toda, A. I. Cristea, and G. Alexandron. Aachen: RWTH Aachen University.

Neumann, A. T., T. Arndt, L. Köbis, R. Meissner, A. Martin, P. de Lange, N. Pengel, R. Klamma, and H. W. Wollersheim. 2021. “Chatbots as a Tool to Scale Mentoring Processes: Individually Supporting Self-Study in Higher Education.” *Frontiers in Artificial Intelligence* 4:668220. https://doi.org/10.3389/frai.2021.668220

Newell, A., J. C. Shaw, and H. A. Simon. 1959. *Report on a General Problem-Solving Program*. Santa Monica, CA: RAND Corporation.

Nguyen, N., and S. Nadi. 2022. “An Empirical Evaluation of GitHub Copilot’s Code Suggestions.” In *Proceedings of Proceedings of the 19th International Conference on Mining Software Repositories (MSR’22)*, 1–5. ACM.

Nguyen, Q., B. Rienties, and J. T. E. Richardson. 2020. “Learning Analytics to Uncover Inequality in Behavioural Engagement and Academic Attainment in a Distance Learning Setting.” *Assessment and Evaluation in Higher Education* 45 (4): 594–606. https://doi.org/10.1080/02602938.2019.1679088.

Nguyen, Q., B. Rienties, and D. Whitelock. 2022. “Informing Learning Design in Online Education Using Learning Analytics of Student Engagement.” In *Open World Learning: Research, Innovation and the Challenges of High-Quality Education*, edited by B. Rienties, R. Hampel, E. Scanlon, and D. Whitelock, 189–207. Oxfordshire: Routledge.

Nguyen, H. A., H. Stec, X. Hou, S. Di, and B. M. McLaren. 2023. “Evaluating ChatGPT’s Decimal Skills and Feedback Generation in a Digital Learning Game.” In *European Conference on Technology Enhanced Learning*, edited by O. Viberg, I. Jivet, and P. Muñoz-Merino, 278–293. Cham: Springer Nature Switzerland.

Nirala, K. K., N. K. Singh, and V. S. Purani. 2022. “A Survey on Providing Customer and Public Administration Based Services Using AI: Chatbot.” *Multimedia Tools and Applications* 81 (16): 22215–22246. https://doi.org/10.1007/s11042-021-11458-y

Ohlsson, S. 1992. “Constraint-based Student Modeling.” *Journal of Artificial Intelligence in Education* 3 (4): 429–447.

Oli, P., R. Banjade, J. Chapagain, and V. Rus. 2023. “The Behavior of Large Language Models When Prompted to Generate Code Explanations.” In *Proceedings of Proceedings of the Workshop on Generative AI for Education (GAIED) at the Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, December 2023*.

OpenAI. 2023. “Introducing ChatGPT.” https://openai.com/blog/chatgpt/.

Ouyang, F., and P. Jiao. 2021. “Artificial Intelligence in Education: The Three Paradigms.” *Computers and Education: Artificial Intelligence* 2:100020. https://doi.org/10.1016/j.caeai.2021.100020

Pane, J. F., B. A. Griffin, D. F. McCaffrey, and R. Karam. 2014. “Effectiveness of Cognitive Tutor Algebra I at Scale.” *Educational Evaluation and Policy Analysis* 36 (2): 127–144. https://doi.org/10.3102/0162373713507480

Pankiewicz, M., and R. Baker. 2023. “Large Language Models (GPT) for Automating Feedback on Programming Assignments.” In *Proceedings of 31st International Conference on Computers in Education (ICCE’2023), Matsue, Shimane, Japan, December 4–8, 2023*, edited by J. L. Shih.

Phung, T., J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. 2023. “Generating High-Precision Feedback for Programming Syntax Errors Using Large

Language Models." In *Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023)*, 482–489. Worcester: International Educational Data Mining Society.

Porter, L., and D. Zingaro. 2024. *Learn AI-Assisted Python Programming: With Github Copilot and ChatGPT*. New York City: Simon and Schuster.

Pozdniakov, S., R. Martinez-Maldonado, Y.-S. Tsai, M. Cukurova, T. Bartindale, P. Chen, H. Marshall, D. Richardson, and D. Gašević. 2022. "The Question-driven Dashboard: How can we Design Analytics Interfaces Aligned to Teachers' Inquiry?" https://doi.org/10.1145/3506860.3506885.

Prather, J., P. Denny, J. Leinonen, B. A. Becker, I. Albluwi, M. Craig, H. Keuning, et al. 2023. "The Robots are Here: Navigating the Generative AI Revolution in Computing Education." In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*, edited by M. Monga, M. J. Laakso, S. Simon, and J. Sheard, 108–159. New York City, NY: ACM.

Pressey, S. L. 1926. "A Simple Device for Teaching, Testing, and Research in Learning." *School and Society* 23:373–376.

Prieto, L. P., J. I. Asensio-Pérez, J. A. Munoz-Cristobal, Y. A. Dimitriadis, I. M. Jorrin-Abellan, and E. Gomez-Sanchez. 2013. "Enabling Teachers to Deploy CSCL Designs Across Distributed Learning Environments." *IEEE Transactions on Learning Technologies* 6 (4): 324–336. https://doi.org/10.1109/TLT.2013.22

Rizvi, S., B. Rienties, J. Rogaten, and R. Kizilcec. 2022. "Beyond one-Size-Fits-all in MOOCs: Variation in Learning Design and Persistence of Learners in Different Cultural and Socioeconomic Contexts." *Computers in Human Behavior* 126:106973. https://doi.org/10.1016/j.chb.2021.106973.

Sabzalieva, E., and A. Valentini. 2023. "ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide." https://unesdoc.unesco.org/ark:/48223/pf0000385146.

Santos, E. A., P. Prasad, and B. A. Becker. 2023. "Always Provide Context: The Effects of Code Context on Programming Error Message Enhancement." In *Proceedings of the ACM Conference on Global Computing Education V. 1*, edited by V. Choppella, 147–153. New York City, NY: ACM.

Sarsa, S., P. Denny, A. Hellas, and J. Leinonen. 2022. "Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models." In *Proceedings of the 2022 ACM Conference on International Computing Education Research*, edited by J. Vahrenhold, 27–43. New York City, NY: ACM.

Savelka, J., A. Agarwal, M. An, C. Bogart, and M. Sakr. 2023. "Thrilled by Your Progress! Large Language Models (GPT-4) no Longer Struggle to Pass Assessments in Higher Education Programming Courses." In *Proceedings of the 2023 ACM Conference on International Computing Education Research V. 1*, edited by V. Choppella, 78–92. New York City, NY: Association for Computing Machinery.

Sharples, M. 2023. "Towards Social Generative AI for Education: Theory, Practices and Ethics." *Learning: Research and Practice* 9 (2): 159–167. https://doi.org/10.1080/23735082.2023.2261131.

Shneiderman, B. 2020. "Human-centered Artificial Intelligence: Three Fresh Ideas." *AIS Transactions on Human-Computer Interaction* 12 (3): 109–124. https://doi.org/10.17705/1thci.00131.

Smetana, L., and R. Bell. 2012. "Computer Simulations to Support Science Instruction and Learning: A Critical Review of the Literature." *International Journal of Science Education* 34 (9): 1337–1370. https://doi.org/10.1080/09500693.2011.605182.

Stephanidis, C., G. Salvendy, M. Antona, J. Y. Chen, J. Dong, V. G. Duffy, X. Fang, et al. 2019. "Seven HCI Grand Challenges." *International Journal of Human–Computer Interaction* 35 (14): 1229–1269. https://doi.org/10.1080/10447318.2019.1619259

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." In *The 14th Learning Analytics and Knowledge Conference (LAK '24), Kyoto, Japan. ACM, New York, NY, USA*. https://doi.org/10.1145/3636555.

Susnjak, T. 2023. "Beyond Predictive Learning Analytics Modelling and Onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT." *International Journal of Artificial Intelligence in Education* 34:1–31.

Tankelevitch, L., V. Kewenig, A. Simkute, A. Scott, A. Sarkar, A. Sellen, and S. Rintel. 2023. "The Metacognitive Demands and Opportunities of Generative AI." arXiv preprint arXiv:2312.10893.

Taub, M., and R. Azevedo. 2023. "Teachers as Self-Regulated Learners: The Role of Multimodal Data Analytics for Instructional Decision-Making." *New Directions in Teaching and Learning* 2023 (174): 25–32. https://doi.org/10.1002/tl.20545.

Toetenel, L., and B. Rienties. 2016. "Analysing 157 Learning Designs Using Learning Analytic Approaches as a Means to Evaluate the Impact of Pedagogical Decision-Making." *British Journal of Educational Technology* 47 (5): 981–992. https://doi.org/10.1111/bjet.12423.

Tran, A., K. Angelikas, E. Rama, C. Okechukwu, D. H. Smith IV, and S. MacNeil. 2023. "Generating Multiple Choice Questions for Computing Courses Using Large Language Models." In *Proceedings of Frontiers in Education Conference (FIE 2023)*, 1–8, College Station, TX, October 18–21, 2023, Session PP24.

UK Government. 2023. "The Bletchley Declaration by Countries Attending the AI Safety Summit." https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Wang, Z., A. Lan, and R. Baraniuk. 2021. "Math Word Problem Generation with Mathematical Consistency and Problem Context Constraints." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5986–5999.

Wang, Z., J. Valdez, D. B. Mallick, and R. Baraniuk. 2022. "Towards Human-Like Educational Question Generation with Large Language Models." In *Proceedings of 23rd International Conference on Artificial Intelligence in Education, Part 1*, edited by M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, 153–166. Durham: Springer.

Winne, P. H. 2017. "Leveraging big Data to Help Each Learner Upgrade Learning and Accelerate Learning Science." *Teachers College Record* 119 (13): 1–24. https://doi.org/10.1177/016146811711901312

Winne, P. H. 2018. "Cognition and Metacognition Within Self-Regulated Learning." In *Handbook of Self-Regulation of Learning and Performance*, edited by D. H. Schunk and J. A. Greene, 36–48. London: Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315697048-3.

Winne, P. H., and R. Azevedo. 2022. "Metacognition." In *Cambridge Handbook of the Learning Sciences*, 3rd ed., edited by K. Sawyer, 93–113. Cambridge, MA: Cambridge University Press.

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* 00:1–23. https://doi.org/10.1111/bjet.13370.

Yang, H., Y. He, X. Bu, H. Xu, and W. Guo. 2023. "Automatic Essay Evaluation Technologies in Chinese Writing; A Systematic Literature Review." *Applied Sciences* 13 (19): 10737. https://www.mdpi.com/2076-3417/13/19/10737.

Yeh, Y.-F., K. K. H. Chan, and Y.-S. Hsu. 2021. "Toward a Framework That Connects Individual TPACK and Collective TPACK: A Systematic Review of TPACK Studies Investigating Teacher Collaborative Discourse in the Learning by Design Process." *Computers & Education* 171:104238. https://doi.org/10.1016/j.compedu.2021.104238.