Front. Digit. Educ., 2024, 1(3): 223–245
https://doi.org/10.1007/s44366-024-0028-5

REVIEW ARTICLE

A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education

Xin Zhang[a], Peng Zhang[a], Yuan Shen[b], Min Liu[c], Qiong Wang[a], Dragan Gašević[d], Yizhou Fan[a,d]

[a] Graduate School of Education, Peking University, Beijing 100871, China

[b] Research Center for Data Hub and Security, Zhejiang Lab, Hangzhou 311121, China

[c] Teachers College, Beijing Language and Culture University, Beijing 100083, China

[d] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

Abstract Generative artificial intelligence (GenAI), achieving human-like capabilities in interpreting, summarising, creating, and predicting language, has sparked significant interest, leading to extensive exploration and discussion in educational applications. However, the frontline practice of education stakeholders or the conceptual discussion of theorists alone is not sufficient to deeply understand and reshape the application of GenAI in education, and rigorous empirical research and data-based evidence are essential. In the past two years, a large number of empirical studies on GenAI in education have emerged, but there is still a lack of systematic reviews to summarise and analyse the current empirical studies in this field to evaluate existing progress and inform future research. Therefore, this work systematically reviews and analyses 48 recent empirical studies on GenAI in education, detailing their general characteristics and empirical findings regarding promises and concerns, while also outlining current needs and future directions. Our findings highlight GenAI's role as an assistant and facilitator in learning support, a subject expert and instructional designer in teaching support, and its contributions to diverse feedback methods and emerging assessment opportunities. The empirical studies also raise concerns such as the impact of GenAI imperfections on feedback quality, ethical dilemmas in complex task applications, and mismatches between artificial intelligence (AI)-enabled teaching and user competencies. Our review also summarises and elaborates on essential areas such as AI literacy and integration, the impact of GenAI on the efficiency of educational processes, collaborative dynamics between AI and teachers, the importance of addressing students' metacognition with GenAI, and the potential for transformative assessments. These insights provide valuable guidelines for future empirical research on GenAI in education.

Keywords generative artificial intelligence (GenAI), empirical research, systemic literature review, education

1 Introduction

In the past decades, artificial intelligence (AI) technologies have started to influence teaching and learning by enabling innovative practices such as personalised learning (Tapalova & Zhiyenbayeva, 2022), intelligent tutoring systems (Alam, 2023), and automated assessments (Ercikan & McCaffrey, 2022). Traditional AI, often referred to as Weak AI or Narrow AI, is designed to perform specific tasks and solve particular problems within a limited scope, such as voice assistants like Apple's Siri or Amazon's Alexa (Russell & Norvig, 2016). This type of AI operates within a defined scope and excels in areas like pattern recognition and decision-making, based on pre-programmed rules and data-driven insights (Goodfellow, 2016). In contrast, Strong AI can be divided into two categories (Russell & Norvig, 2016): Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI). AGI refers to a system with self-awareness and the ability to solve problems and plan for the future, while ASI represents a hypothetical form of intelligence that surpasses human abilities. While AGI remains a theoretical goal, in November 2022, the debut of ChatGPT brought Generative AI (GenAI) into the spotlight (OpenAI, 2024a), marking a significant advancement in AI capabilities. GenAI,

Received July 31, 2024; accepted September 15, 2024

Yizhou Fan (✉)
E-mail: fyz@pku.edu.cn

based on large language models (LLMs) and artificial neural networks (ANNs), processes vast datasets to achieve human-like capabilities in understanding, summarising, creating, and predicting language (Tedre et al., 2023). Unlike previous AI technologies, GenAI models such as ChatGPT (OpenAI, 2024b), Claude (Anthropic, 2024), and Llama (Meta AI, 2023) significantly enhance content generation through techniques like in-context learning and reinforcement learning from human feedback (Wu et al., 2023). These advancements extend far beyond mere response generation, thus broadening their applicability across various fields, such as artistic media (Epstein et al., 2023), journalism (Pavlik, 2023), medicine (Lang et al., 2024), and finance (Chen et al., 2023). The rapid adoption of ChatGPT underscores the widespread appeal and significant potential pivotal role in shaping future technological and academic landscapes (Baidoo-Anu & Owusu Ansah, 2023).

In the early days of GenAI, there were considerable concerns and excitement about its implications for human learning, teaching and education, yet empirical evidence regarding its value remained scarce. Since the launch of ChatGPT, several studies have emerged, providing valuable empirical insights into GenAI's impact. Despite the significance of these studies, there has been limited effort to systematically organise the existing empirical evidence. Some researchers (Baidoo-Anu & Owusu Ansah, 2023; Bannister et al., 2023; Bozkurt, 2023; Law, 2024; de la Torre & Baldeon-Calisto, 2024) have conducted systematic literature reviews or bibliometric analysis to enhance understanding of the application of GenAI and its potential benefits in education. Furthermore, it should be noted that most review articles in the literature discuss the promises and concerns of ChatGPT (Lo, 2023; Karthikeyan, 2023; Montenegro-Rueda et al., 2023; Pradana et al., 2023; Rahman & Watanobe, 2023; Vargas-Murillo et al., 2023). However, most previous reviews focused on specific areas such as medical education (Xu et al., 2024), healthcare education (Sallam, 2023), and programming education, with minimal emphasis on empirical studies. This was in part because empirical research was very scarce at the time when researchers conducted previous reviews (Law, 2024; Luo, 2024). However, empirical evidence is crucial in the field of GenAI in education because it provides robust data that can evaluate the effectiveness and quality of GenAI in various educational settings (Law, 2024). Systematising evidence in empirical research can help identify gaps in the literature and establish clear directions and priorities for future research (Michel-Villarreal et al., 2023).

According to Yan et al. (2024a), GenAI offers promises and poses challenges to human learning (see Figure 1). GenAI offers extensive promises across various educational tasks. It supports learning by generating contextually relevant responses and adjusting teaching strategies based on student reactions. This aligns with Vygotsky's sociocultural theory by actively guiding students within their Zone of Proximal Development (Vygotsky & Cole, 1978). For teaching support, GenAI can be used to create educational resources, such as image generation in STEAM classes (Lee et al., 2023) and personalised reading materials for middle school English learners (Xiao et al., 2023), reducing the preparation burden for educators and potentially enhancing the quality of instruction. It can also be used to provide personalised, timely feedback on different questions, allowing for human-like interaction (Lu et al., 2024). Furthermore, GenAI holds a promise to improve the assessment processes by implementing adaptive scoring systems that accurately reflect student performance and introducing innovative tools such as conversational assessment (Latif & Zhai, 2024; Nguyen Thanh et al., 2023) and automated essay grading (Latif & Zhai, 2024; Yancey et al., 2023). Despite these promises, the rise of GenAI in education also raises concerns such as ethical issues (Alier et al., 2024; Ferrara, 2024), data privacy protection (Gupta et al., 2023; Michel-Villarreal et al., 2023), and academic integrity (Sevnarayan & Potter, 2024; Yusuf et al., 2024). Beyond promises and concerns, Yan et al. (2024a) emphasise that to effectively integrate GenAI in education, three critical needs must be addressed: enhancing AI literacy among educational stakeholders, prioritising evidence-based decision-making, and ensuring methodological rigour in educational research on GenAI integration.

To date, there have been very few reviews of empirical research on the application of GenAI in education (AlBadarin et al., 2024; Lo et al., 2024). However, AlBadarin et al. (2024) conducted a systematic review covering only 14 empirical studies published between 2022 and April 10 2023, revealing the integration and impact of ChatGPT in educational settings. A significant limitation of the review is that there had not been a large amount of rigorous empirical research accumulated at that time, especially research using large models more advanced than ChatGPT 3.5 (i.e., ChatGPT 4.0 was released in March 2023). Recently, we witnessed the emergence of large multimodal foundation models (LMFMs), such as ChatGPT-4 Turbo by OpenAI or Gemini by Google, along with the rapid growth of the applications of their application in education. The potential implications of these LMFMs on education demands further exploration as they could offer deeper integrative capabilities (Küchemann et al., 2024). Although Lo et al. (2024) effectively synthesises empirical research on the influence of ChatGPT on student engagement using a

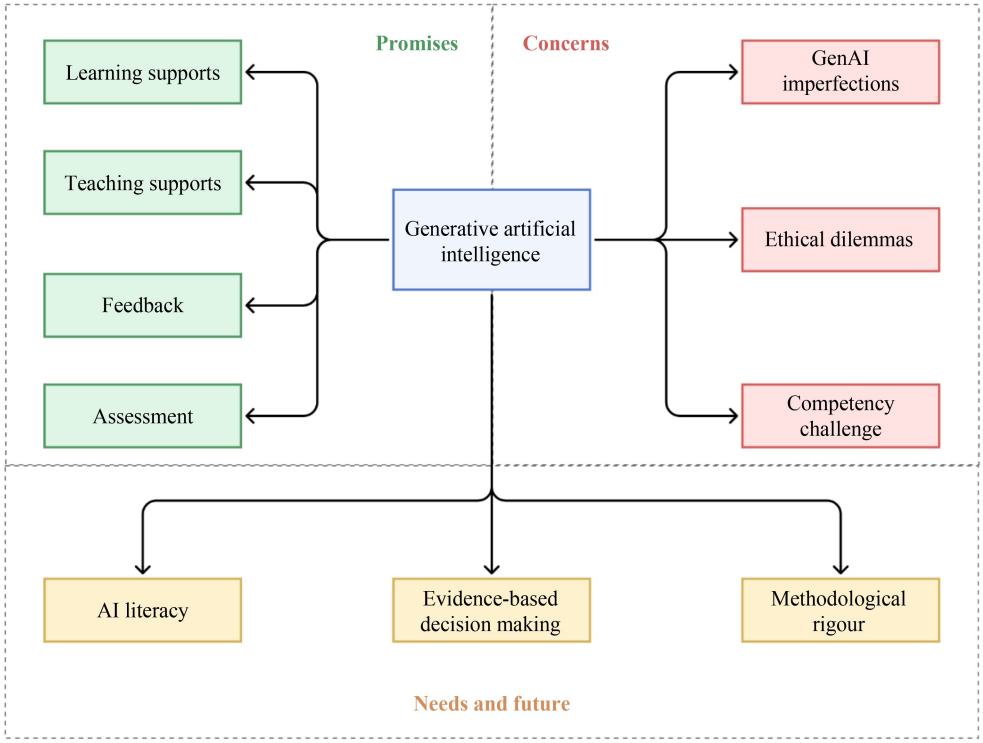

Figure 1 Yan et al. (2024a)'s framework of promises, concerns, and future directions for GenAI in education.

multidimensional framework, it lacks discussion on assessment, insights from longitudinal studies, and practical implementation guidelines.

Therefore, a systematic review of empirical studies in education application is urgently needed. This review clarifies how GenAI changes learning and teaching practices while critically addressing the ethical and practical challenges it brings, thereby contributing to establishing a future research agenda for GenAI's application in education. This review uniquely contributes to the field by offering an updated, comprehensive analysis of empirical studies that capture the advancements brought by these new LMFMs, providing insights into their broader implications for educational practices. Our work thus differentiates itself by both its timeliness in addressing the latest developments and its exclusive focus on empirical evidence, filling a critical gap in the existing literature. The following research questions (RQ1–RQ4) are posed to guide the review:

RQ1: What are the overarching characteristics and the landscape of empirical research on the application of GenAI in education?

RQ2: What promises does the application of GenAI in education hold based on empirical evidence?

RQ3: What concerns arise from the application of GenAI in education as evidenced by empirical research?

RQ4: What are the needs and future directions for GenAI in education identified in the empirical research?

2 Methodology

2.1 | Search Strategy

To conduct a comprehensive systematic review, this study adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). The literature search was carried out using six major digital libraries: (1) ACM Digital Library, (2) Education Research Complete, (3) ERIC, (4) IEEE Xplore, (5) Scopus, and (6) Web of Science. These databases were selected because of their extensive coverage of research related to technology and education. The search strategy involved querying these databases with a specific search string designed to capture a broad range of studies concerning GenAI. The search string used was (*ChatGPT OR LLM OR large language model* OR GenAI OR gai OR generative artificial intelligence OR generative AI*) AND (*quantitative OR qualitative*) aimed at identifying articles that contained these terms within their abstracts. This approach was chosen to ensure comprehensive retrieval of relevant quantitative or qualitative studies that discuss and investigate GenAI in education.

2.2 | Search Criteria

The initial search yielded 833 articles as of April 30, 2024. Given the likelihood of overlap across databases,

(1) duplicate entries were removed. Following the removal of duplicate entries, 445 articles remained. (2) Further scrutiny was necessary due to the multiple meanings associated with the acronym LLM, which also stands for language learning motivation, lipid lowering medication, leg lean mass, and Local Linear Mapping, among others. Articles unrelated to GenAI under these ambiguities were excluded. Meanwhile, (3) we imposed exclusion criteria on the resultant papers to remove short papers, and those that did not actually report on as empirical research. (4) Literature reviews and bibliometric studies, if retrieved, were used as background references, therefore not included in the review of empirical research. (5) In addition, articles that focused primarily on exploring viewpoints or perceptions regarding GenAI without providing empirical evidence were excluded. This criterion aimed to filter out qualitative studies by simply collecting GenAI perceptions among educators and students. Instead, the focus was on empirical studies that rigorously examined the actual application and effects of GenAI in educational settings. In June 2024, we used the snowballing method to identify additional relevant papers, which led to the inclusion of five more empirical studies. Ultimately, after applying all exclusion criteria, the search was narrowed down to 48 pertinent articles, which were deemed suitable for in-depth analysis in this systematic review. These articles were expected to provide valuable insights into the promises, concerns, needs and future of the application of GenAI in education. The review process is illustrated in Figure 2.

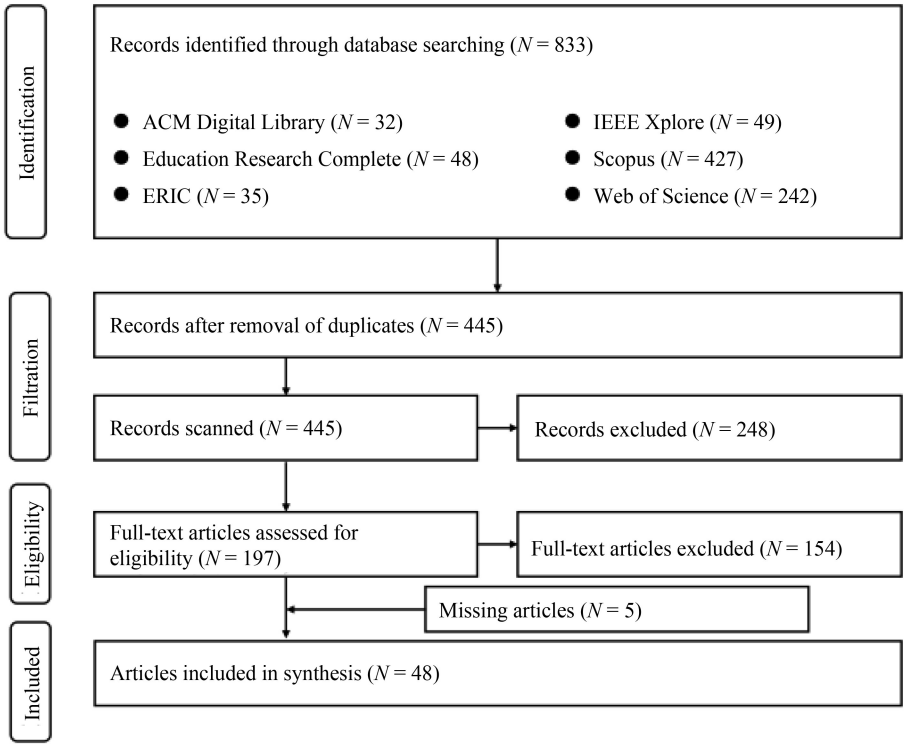

Figure 2 Systematic literature review methodology.

2.3 | Coding System

To explore our research question (RQ2–RQ4), this study mainly drew on the insights from Yan et al. (2024a) regarding the potential of GenAI in education and its associated ethical and practical challenges (see Figure 1). Specifically, to systematically explore the promises, concerns, and needs and future directions Yan et al. (2024a) mentioned in empirical studies included in the review, we proposed a coding framework and categorised these codes into specific coding categories as detailed in Table 1. This framework was carefully refined through multiple rounds of discussion among the researchers, ensuring it reflects a high level of agreement and consistency across all categories.

In order to address **RQ1**, six specific codes were used to categorise and analyse the general trends in the empirical educational research of GenAI. **Types of research subjects** categorised the subjects of the study into three groups: *GenAI*, *students*, and *teachers*. **Education level** explored research contexts across different educational stages, such as *K–12 education* and *higher education*. **Task** categorised educational scenarios, such as *writing*, *program*, and other specific tasks for which GenAI had been applied. Given the

Table 1 Codification by RQ1–RQ4

RQ	Coding	Categories
1	Types of research subjects	Student, teacher, GenAI
	Education level	K–12 education, higher education, vocational education
	Task	Writing, program, language learning, question answer, research, class, creative content, reading, teacher training, life
	GenAI type	ChatGPT, ChatGPT 3.5, ChatGPT 4.0, other else
	Data collection method	Questionnaire, interview/feedback, content generation, think-aloud, score rubric, performance records
	Research design	Descriptive analysis, inferential statistical analysis, experimental research, survey research, observational research, case studies, content analysis
2	Learning support	Interactive learning, cognitive facilitator, performance improvement, correct answer provider
	Teaching support	Curriculum development, content generation
	Feedback	Timeliness, informativeness, engagement, emotional interaction, high density, usefulness
	Assessment	Adaptive assessment, authentic assessment
3	GenAI imperfections	Hallucination, overgeneralised, randomness, data processing limitations, inability to generate original ideas, lack of emotional intelligence
	Ethical dilemmas	Beneficence, equality, over-reliance, transparency, academic integrity
	Competency challenge	Utilisation gap, skill atrophy, assessment validity
4	AI literacy	Training programs for educators, comprehensive AI literacy curriculum, ethical awareness training, AI literacy measurement tools
	Integration of GenAI	Integrating GenAI across scenarios, integrating GenAI with human experts/teachers, integrating GenAI with students, integrating GenAI with other technologies
	Evidence-based decision-making	Robust learning impacts evidence, metacognitive and cognitive skills engagement, multi-party collaborative efforts
	Methodological rigour	Evidence quality appraisal instrument, transparent prompting and generation disclosure, rigorous methodological standards

varying capabilities of different LLMs, **GenAI** distinguishes between unspecified versions of *ChatGPT*, the regular 3.5 version, the upgraded 4.0 version, and other GenAI tools such as Bing and Bard. **Data collection method** identified the methods used to collect data, such as *surveys* and *interviews*. Lastly, **research design** focused on the specific qualitative and quantitative methods used in the research. By coding the empirical articles on GenAI's application in education through RQ1, we aimed to explore the general characteristics of current research in this field.

In order to address **RQ2**, four codes were used as the promise of GenAI in **learning support** to characterise how GenAI contributes to learning. *Performance improvement* was defined by the ability of GenAI to assist students in completing and enhancing performance on their tasks. This involved generating content which students could integrate into their work, saving time and enhancing efficiency. *Cognitive facilitator* was characterised by GenAI acting as a capable peer or tutor. *Interactive learning* was defined by the ability of GenAI to engage students in text-based dialogues. This feature allowed students to ask further questions based on GenAI's responses, leading to a more engaging learning experience and a deeper understanding of the material. *Correct answer provider* was characterised by GenAI's high accuracy and completeness in answering questions across various domains. This ensured reliable support for students by providing them with correct and comprehensive answers. As for the promise of GenAI in **teaching support**, it encompassed *content generation* and *curriculum development*, each characterised by specific features. *Content generation* involved creating both textual and interactive elements, such as coding exercises and quizzes, and multimedia content, such as videos and simulations. Curriculum development meant GenAI acting as a subject matter expert and learning designer. It provided valuable insights to enhance educational materials, while also offering practical suggestions for course design. Feedback and assessment were also key areas where GenAI showed significant potential. **Feedback** involved GenAI-produced content characterised by *timeliness*, *informativeness*, *high density*, and *usefulness*, as well as its interaction with students, including *engagement* and *emotional interaction*. **Assessment** focuses on using GenAI as an evaluation tool in educational scenarios, shifting from traditional, often cumbersome methods to more *adaptive and authentic processes*. By examining these codes, we aimed to highlight the practical benefits and promises GenAI can bring to educational settings.

In order to address **RQ3**, three codes were used. Firstly, **GenAI imperfections** related to the inherent technical shortcomings of GenAI, such as *hallucination* and *randomness*, which could significantly affect the

accuracy and usability of the content generated by GenAI. Articles extracted also identified other technical limitations of GenAI, including *issues of overgeneralised*, *data processing limitations*, *inability to generate original content*, and *lack of emotional intelligence*. Secondly, the **ethical dilemmas** category highlighted the ethical issues arising from GenAI replacing tasks traditionally performed by humans. This was particularly concerning regarding *beneficence*, *fairness*, and *integrity*. This category indicated that concerns about the application of GenAI stemmed not only from its technical characteristics but also from the inconsistency between these characteristics and the ethical principles emphasised in educational settings. Thirdly, the **competency challenge** category referred to the difficulties and risks arising from the lack of ability among users, such as teachers and students, to use GenAI correctly and appropriately during its application. In addition to *utilisation gap* described above, the code of competency challenge also included issues such as *skill atrophy* due to over-reliance on GenAI and *assessment validity* crisis resulting from unclear delineation of roles in human–machine collaboration.

In order to address **RQ4**, four codes were used. **AI literacy** encompassed a basic understanding of how AI systems function but also an intimate awareness of their potential impact, ethical considerations, and limitations. The need to adopt a *comprehensive AI literacy curriculum* and *training programmes for educators* is essential to realise the promises of GenAI and mitigate its challenges among all educational stakeholders. In addition, *ethics awareness training* is critical to prepare people to live, learn, and work with GenAI in the 21st century, and this process requires adopting *AI literacy measurement tools*. The **integration of GenAI** code indicated that effectively integrating GenAI into educational practices was essential in the present stage of technological development, ensuring that GenAI serves as an intelligent aid rather than a single driver. For example, *integrating GenAI across various educational scenarios*, *integrating GenAI with human experts*, *teachers* and *students*, and *integrating AI with other emerging technologies*. **Evidence-based decision-making** required a collaborative effort among researchers, practitioners, and policymakers to generate robust evidence that could guide the effective and responsible use of AI in education. The *metacognitive and cognitive skills enhancement category* indicated, for example, while GenAI could significantly improve students' writing proficiency, there was a risk for their metacognitive skills such as students may overestimate their understanding due to the ease of cognitive information processing. Therefore, we need further robust learning impacts evidence to achieve a nuanced understanding of both its potential benefits and limitations. Besides, *multi-party collaborative efforts* including researchers, practitioners, and policymakers were necessary to guide the effective and responsible use of AI in education. Building on the discussions of evidence-based decision-making, it is paramount to emphasise the importance of **methodological rigour** in the application of GenAI technologies within educational research, including *utilising various evidence quality appraisal instruments*, *transparent prompting and generation disclosure*, and *establishing rigorous methodological standards*.

3 Results

The results of the literature review, as of April 30, 2024, indicate a robust and growing body of research focusing on the application of GenAI in education. Starting at the end of 2022, only Pinargote et al. (2022) raised an empirical study using GenAI. Following this, the year 2023 saw the publication of 24 articles, while the first four months of 2024 already contributed 23 articles to the field. This trend suggests a continuing and heightened focus on exploring the capabilities and implications of GenAI technologies in educational settings.

3.1 | RQ1: What Are the Overarching Characteristics and the Landscape of Empirical Research on the Application of GenAI in Education?

3.1.1 Diversity in Participants, Educational Contexts, and Research Foci

In our sample (see Table 2), the majority of studies focused primarily on students and GenAI as the research object, with only a few specifically examining teachers as central participants. Regarding studies taking GenAI as object, Riedel et al. (2023) evaluated ChatGPT's proficiency in addressing obstetrics and gynaecology-related questions and found it performed well on multiple choice questions and simple case study responses. In terms of educational levels, higher education was the most frequent (75%) context for empirical research on GenAI applications, followed by K–12 and vocational education. The focus on higher education could be expected due to researchers' affiliations and the advanced infrastructure and resources available in these settings.

The range of tasks examined within these educational settings was diverse. Writing-related tasks were the most prevalent, covering a range of activities from academic writing (Lu et al., 2024), to English as a

Table 2 Participant settings

Coding	Categories	*F*	*RF* (%)	Examples
Types of research subjects	Student	27	56	AlGhamdi (2024); Wachira et al. (2023); Wan et al. (2024)
	GenAI	19	40	Joshi et al. (2024); Maitland et al. (2024); Theelen et al. (2024)
	Teacher	2	4	Matthews & Volpe (2023); Yeh (2024)
Education level	Higher education	36	75	AlGhamdi (2024); Baba et al. (2024); Joshi et al. (2024); Nguyen et al. (2024)
	K–12 education	9	19	Alneyadi & Wardat (2024); Vázquez-Cano et al. (2023); Zhang & Huang (2024)
	Vocational education	3	6	Hu et al. (2024); Theelen et al. (2024); Yeh (2024)
Task	Writing	15	31	de Vicente-Yagüe-Jara et al. (2023); Liu et al. (2024); Niloy et al. (2024)
	Question answer	10	21	Hu et al. (2024); Keiper et al. (2023); Kieser et al. (2023); Maitland et al. (2024); Plevris et al. (2023)
	Research	6	13	Dengel et al. (2023); Essel et al. (2024)
	Program	5	10	Hellas et al. (2023); Phung et al. (2023); Shirafuji et al. (2023)
	Class	4	8	Alneyadi & Wardat (2024); Dasari et al. (2024); Villan & Santos (2023); Wu et al. (2024)
	Else (language learning, creative content, reading, teacher training)	8	16	Baba et al. (2024); Matthews & Volpe (2023); Pinargote et al. (2022); Putjorn & Putjorn (2023); Urban et al. (2024); Vázquez-Cano et al. (2023); Yeh (2024); Zhang & Huang (2024)
GenAI type	ChatGPT	22	46	Bernabei et al. (2023); Theelen et al. (2024); Villan & Santos (2023)
	ChatGPT 3.5	22	46	Joshi et al. (2024); Lu et al. (2024); Niloy et al. (2024)
	ChatGPT 4.0	6	13	Kieser et al. (2023); Maitland et al. (2024); Phung et al. (2023); Plevris et al. (2023)
	Else	10	21	Kieser et al. (2023); Wachira et al. (2023)

Notes. *F* = frequency, *RF* = frequency ratio.

foreign language (EFL) writing (Liu et al., 2024; Yan,2023), and to creative writing activities (Niloy et al., 2024). Following closely were studies focusing on question answering, which typically evaluated GenAI's capability to engage with content-specific questions pertinent to standardised testing or textbooks. These tasks particularly shone in education in the fields of medicine (Gilson et al., 2023; Maitland et al., 2024; Riedel et al., 2023) and the sciences, including physics (Kieser et al., 2023), chemistry (Watts et al., 2023), mathematics (Latif & Zhai, 2024), and other areas (Hu et al., 2024; Uddin et al., 2023). Research-related tasks also featured prominently with studies using GenAI for education-related qualitative analyses such as theme identification and coding (De Paoli, 2023; Morgan, 2023; Tai et al., 2024; Theelen et al., 2024). GenAI tools such as ChatGPT have been increasingly utilised in programming education for tasks such as interaction strategy (Yan et al., 2024b), computational thinking skills (Yilmaz & Yilmaz, 2023), and code restructuring (Shirafuji et al., 2023), as well as in scientific subjects such as mathematics (Dasari et al., 2024; Wu et al., 2024) and physics (Alneyadi & Wardat, 2024) to enhance student performance and engagement in class. Beyond these core areas, GenAI was also explored for its potential in language learning (Yeh, 2024; Zhang & Huang, 2024), creative content generation (Putjorn & Putjorn, 2023), reading comprehension (Vázquez-Cano et al., 2023), teacher training (Matthews & Volpe, 2023), and as a campus life helper (Baba et al., 2024; Pinargote et al., 2022). These areas were categorised under the *else* category, highlighting GenAI's versatile applications across a broad spectrum of educational applications.

In the empirical research landscape of GenAI applications in education, ChatGPT remained the predominant tool, with version ChatGPT 3.5 being the most extensively used up to the moment of our review. Despite the superior performance and multimodal data capabilities of ChatGPT 4.0, the cost of large numbers of requests to the GPT-4 API is non-negligible (Savelka et al., 2023). However, the free release of ChatGPT 4.0 on May 13, 2024, holds significant implications for future research. Meanwhile, it should be noted that when studies did not specify the version of ChatGPT used, we simply categorised them as ChatGPT. Additionally, the research community explored other GenAI platforms, such as Bing (Liu et al., 2024) and Google's BARD (Dengel et al., 2023; Plevris et al., 2023), which were categorised as else. These platforms were often included in comparative studies to evaluate their features and effectiveness in education.

3.1.2 Data Collection Methods

As shown in Table 3, the most common data collection method in our sample was using scoring rubrics to assess the quality of content generated by GenAI systems. The dimensions of these rubrics varied based on the education level and the specific tasks evaluated. For instance, Keiper et al. (2023) developed an analytical binary rubric with three dimensions,

including grammar structure, completeness, and topical accuracy to assess the quality of sports management educational material generated by ChatGPT. In STEM disciplines, such as chemistry, Latif & Zhai (2024) crafted scoring standards that encompassed five different response perspectives, tailored to evaluate the scientific accuracy and relevancy of answers provided by GenAI. Typically, these rubrics were applied by multiple experienced researchers throughout the development process to ensure consistency and accuracy in data collection.

Interviews or gathering participant feedback was another prevalent method of data collection. AlGhamdi (2024) conducted a study examining the impact of ChatGPT-generated feedback on the writing skills of first-year computer science students at a Saudi university, comparing their reflections on feedback from both teachers and ChatGPT through qualitative coding analysis. Similarly, Yan (2023) recruited eight participants via case-by-case observation and screening for three in-depth interview sessions to explore their perceptions of using ChatGPT for writing tasks.

Questionnaires were also widely used either to collect basic information or for studies exploring perceptions. Wu et al. (2024) used Motivation Scale, Engagement Scale, and Self-Efficacy Scale as part of the post-test in a high school maths experiment. Furthermore, direct qualitative analysis of GenAI-generated content was a frequent approach, alongside think-aloud data collection (Liu et al., 2024) during experimental tasks. Performance analysis was another significant method, examining the effectiveness of GenAI applications in educational settings. Dasari et al. (2024) explored how students relying solely on ChatGPT for learning mathematics performed in comparison to those receiving teacher guidance with or without ChatGPT support, highlighting the impact of GenAIon learning outcomes.

It is important to highlight that 22 articles utilised multiple data collection methods. For example, Matthews & Volpe (2023) employed a scoring rubric to evaluate scholars' perspectives on written outputs generated by ChatGPT 3.5 and humans, focusing on accuracy, levels of confidence in decision-making, and instances of mind changing in the context of initial teacher education. Subsequently, interviews were conducted to delve deeper into scholars' views (Matthews & Volpe, 2023). These diverse methods collectively enable a comprehensive evaluation of GenAI's capabilities and effects, providing valuable insights into its practical implications for educational processes and outcomes.

3.1.3 Research Design

As shown in Table 3, experimental research was the most common method which usually involved experimental and control groups. Typically, the control group followed traditional educational methods, while the experimental group incorporated GenAI. For example, in Zhang & Huang (2024), the students in the experimental group were provided access to chatbots (based on LLMs) as a learning support tool for vocabulary acquisition in their EFL studies, whereas the control group used alternative digital resources. In addition to experimental research, survey research through questionnaires or interviews was also utilised to collect both qualitative and quantitative data, allowing researchers to explore participants' perspectives on GenAI.

Table 3 Data collection and research design

Coding	Categories	*F*	*RF* (%)	Examples
Data collection method	Score rubric	24	50	Juanda & Afandi (2024); Keiper et al. (2023); Uddin et al. (2023)
	Interview/feedback	20	42	AlGhamdi (2024); Baba et al. (2024); Yan (2023)
	Questionnaire	11	23	Bernabei et al. (2023); Putjorn & Putjorn (2023); Tossell et al. (2024)
	Content generation	9	19	Dengel et al. (2023); Hellas et al. (2023); Shirafuji et al. (2023)
	Think-aloud	8	17	Liu et al. (2024); Nguyen et al. (2024)
	Performance records	5	10	Dasari et al. (2024); Mansour et al. (2024)
Research design	Descriptive analysis	20	42	Hellas et al. (2023); Kieser et al. (2023); Phung et al. (2023)
	Experimental research	14	29	Dasari et al. (2024); Niloy et al. (2024); Urban et al. (2024)
	Inferential statistical analysis	13	27	de Vicente-Yagüe-Jara et al. (2023); Juanda & Afandi (2024); Riedel et al. (2023)
	Survey research	12	25	Lu et al. (2024); Matthews & Volpe (2023); Wan et al. (2024)
	Case studies	10	21	Keiper et al. (2023); Plevris et al. (2023); Tai et al. (2024); Villan & Santos (2023)
	Content analysis	9	19	Dasari et al. (2024); Nguyen et al. (2024); Wan et al. (2024); Yan et al. (2024b);
	Observational research	6	13	Liu et al. (2024); Wachira et al. (2023); Yeh (2024)

Notes. *F* = frequency, *RF* = frequency ratio.

Quantitative methods prominently featured descriptive and inferential statistical analyses. Descriptive statistics, including measures such as means, standard deviations, and frequency distributions, were commonly used to summarize key variables and provide an initial overview of the data. Inferential techniques, such as t-tests, ANOVA, and regression analysis, were employed to test hypotheses, identify relationships, and determine the statistical significance of the findings. These methods collectively provided a comprehensive understanding of the data, enabling researchers to draw meaningful conclusions about the application of GenAI in education. Qualitative methods including content analysis, case studies, and observational research were used less frequently than quantitative methods in research. It should be noted that 25 studies employed more than one method, with 10 studies utilising a mixed-methods approach, combining both qualitative and quantitative techniques. For example, the Song & Song (2023) study employed a mixed-methods approach, starting with a quantitative phase that utilised a pre-test and post-test design, using established scoring criteria to assess writing skills. This was followed by a qualitative phase where semi-structured interviews were conducted with select participants to delve into their writing motivation and experiences with GenAI-assisted learning. In summary, research on GenAI in education has employed a diverse array of research design, combining quantitative and qualitative analyses to evaluate its effectiveness.

3.2 | RQ2: What Promises Does the Application of GenAI in Education Hold Based on Empirical Evidence?

Among 48 empirical studies we reviewed, 46 studies demonstrated the promises of GenAI across various educational contexts and only two studies did not mention promises of GenAI in education.

3.2.1 GenAI as an Assistant and Facilitator in Learning Support

By serving as both an assistant in optimising student performance on learning tasks and a facilitator in cognitive growth, GenAI offered multifaceted support that addresses the diverse needs of students. Its applications in education fell into four key areas: performance improvement, cognitive facilitation, high-quality answer provision, and interactive learning.

In our review, we found that GenAI was reported to show potential in improving learning performance by helping students complete and optimise their tasks, especially in writing (AlGhamdi, 2024; Lu et al., 2024; Song & Song, 2023). As shown in Table 4, 40% of studies focused on this aspect. For example, Nguyen et al. (2024) conducted a study on doctoral students' academic writing activities supported by GenAI and found that students who frequently used GenAI scored higher on writing tasks compared to those who used it less frequently. The high-scoring students' workflow involved using GenAI to generate task content, reading and copying this content, pasting it into their documents, and then directly editing or adding their own input (Nguyen et al., 2024). While GenAI was generating the content, learners concurrently searched for information and revised their writing, which saved time and improved writing efficiency (Nguyen et al., 2024).

Furthermore, GenAI served as a cognitive facilitator (35%), offering insights and perspectives beyond students' current capabilities (Essel et al., 2024; Phung et al., 2023; Villan & Santos, 2023). This aligned with Vygotsky's Zone of Proximal Development theory (Vygotsky & Cole, 1978), where students were guided to higher levels of understanding. For example, Villan & Santos (2023) utilised ChatGPT as a co-mentor for science project learning, offering students interdisciplinary knowledge and supporting them in tackling more advanced concepts and complex topics. A study conducted by Essel et al. (2024) involving university students in Ghana found that those who used ChatGPT for in-class tasks demonstrated notable improvements in critical (effect size = 0.229), creative (effect size = 0.075), and reflective (effect size = 0.160) thinking skills compared to peers using traditional research tools, highlighting the GenAI's role in enhancing cognitive abilities.

Providing high-quality answers was another identified promise of GenAI, accounting for 25% of the studies. GenAI has been effective in various fields such as sports management issues (Keiper et al., 2023), medical testing (Maitland et al., 2024; Riedel et al., 2023), mathematical logic problems (Plevris et al., 2023), certified engineer exam questions (Hu et al., 2024), and programming problems (Hellas et al., 2023; Joshi et al., 2024). Empirical evidence indicated that GenAI was found to be promising in providing good-quality content (Hu et al., 2024; Keiper et al., 2023), which may support its effectiveness in learning contexts. For example, the study by Keiper et al. (2023) investigated the performance of GenAI in sports exams. The results showed that the model provided a completeness score of 4.75 (5.00) for answers.

GenAI also fostered interactive learning through text-based dialogues, allowing students to engage more deeply with the material. This was highlighted in 21% of the studies wereviewed. These interactions enabled students to ask follow-up questions and receive immediate feedback, enhancing their engagement and understanding. For example, studies

Table 4 Promises of GenAI in education

Coding	Categories	*F*	*RF* (%)	Examples
Learning supports	Performance improvement	19	40	Essel et al. (2024); Lu et al. (2024); Nguyen et al. (2024); Putjorn & Putjorn (2023); Song & Song (2023); Villan & Santos (2023)
	Cognitive facilitator	17	35	Essel et al. (2024); Liu et al. (2024); Phung et al. (2023); Villan & Santos (2023); Zhang & Huang (2024)
	Correct answer provider	12	25	Hellas et al. (2023); Keiper et al. (2023); Riedel et al. (2023); Theelen et al. (2024); Watts et al. (2023)
	Interactive learning	10	21	AlGhamdi (2024); Dasari et al. (2024); Song & Song (2023); Villan & Santos (2023); Wan et al. (2024)
Teaching supports	Content generation	7	15	Dasari et al. (2024); Essel et al. (2024); Phung et al. (2023); Shirafuji et al. (2023); Villan & Santos (2023); Yan (2023); Yeh (2024)
	Curriculum development	5	10	Alneyadi & Wardat (2024); De Paoli (2023); Hellas et al. (2023); Keiper et al. (2023); Villan & Santos (2023)
Feedback	Usefulness	10	21	Dasari et al. (2024); Juanda & Afandi (2024)
	Timeliness	4	8	Song & Song (2023); Putjorn & Putjorn (2023); Yan (2023); Yan et al. (2024b)
	Emotional interaction	4	8	AlBadarin et al. (2024); Essel et al. (2024); Hellas et al. (2023); Lu et al. (2024)
	Engagement	3	6	Baba et al. (2024); Song & Song (2023); Yan et al. (2024b)
	Else	3	6	Putjorn & Putjorn (2023); Riedel et al. (2023); Tossell et al. (2024)
Assessment	Adaptive assessment	2	4	Lu et al. (2024); Pinargote et al. (2022)
	Authentic assessment	2	4	Latif & Zhai (2024); Mansour et al. (2024)

Notes. *F* = frequency, *RF* = frequency ratio.

by Villan & dos Santos (2023) and Dasari et al. (2024) showed that interactive learning with feedback from GenAI led to better performance and deeper comprehension.

3.2.2 GenAI as a Subject Expert and Instructional Designer in Teaching Support

Empirical research on GenAI's promise in teaching support is relatively scarce compared to its role in learning support, according to Table 4. In existing studies, GenAI acted as a subject matter expert and instructional designer in content generation and curriculum development to support teachers in their teaching (Alneyadi & Wardat, 2024; Dasari et al., 2024; Essel et al., 2024; Hellas et al., 2023; Keiper et al., 2023; Phung et al., 2023; Villan & Santos, 2023).

GenAI contributes to the generation of content (15%) by creating both textual content and interactive elements that enrich the educational experience. For example, Phung et al. (2023) explored the application of GenAI in programming education, generating hints for students and correcting programming results. Beyond providing text-based feedback, GenAI can generate interactive activities, such as coding exercises and quizzes (Phung et al., 2023), which could engage students and enhance their understanding of programming concepts. Additionally, GenAI can create multimedia content, including videos and simulations, that can cater to various learning needs. For example, Yeh (2024) highlighted GenAI's effectiveness in assisting teachers in conducting inquiry-based learning in English. GenAI did not only provide textual explanations, but also produced personalised picture books and karaoke exercises (Yeh, 2024). These personalised interactive resources helped teachers create a dynamic student-centred learning experience that addressed the diverse needs of students. Moreover, while meeting students' personalised needs, GenAI-generated interactive content can also reduce teachers' workload. For example, Shirafuji et al. (2023) utilised GenAI to automatically generate refactored programs, thereby reducing the amount of time teachers spent guiding students through code refactoring. By automating the generation of diverse content types, GenAI can enable educators to expand their knowledge and save time, allowing them to focus on more creative and interactive aspects of teaching, thus enhancing the overall learning experience.

In addition to content generation, GenAI showed promise in the realm of curriculum development (10%) by acting as both a subject expert and a learning designer. According to research by Keiper et al. (2023), ChatGPT's responses to short-answer questions in sports management were found to be grammatically accurate, content-complete, and thematically aligned with the subject. This indicates that GenAI could serve as an effective subject matter expert, providing valuable insights that can enhance the quality of materials development. Moreover, Keiper et al. (2023) suggests that GenAI's capabilities extended beyond mere content provision to include practical

suggestions that aided in course design (Keiper et al., 2023). By generating well-structured content and offering pedagogical advice, GenAI can support educators in developing comprehensive and engaging curriculum. Villan & dos Santos (2023) further illustrated this point through their action research in elementary education, where the integration of ChatGPT increased student engagement and reduced teacher resistance to new technologies. This study emphasised the role of ChatGPT in project-based learning, showcasing its ability to design interactive and student-centred learning experiences. In general, these findings highlight the dual role of GenAI in curriculum development, enhancing both content quality and the pedagogical design of educational programmes.

3.2.3 Diverse Feedback Promises and Emerging Assessment Opportunities

GenAI may transform educational settings by enhancing feedback and assessment, differentiating it from traditional AI applications. Unlike traditional AI, which typically focuses on static data analysis and pattern recognition, GenAI actively engages with content generation, making it a dynamic participant in educational processes.

In feedback mechanisms, GenAI may go beyond simple error correction, highlighting the learning process itself. For example, according to Table 5, Putjorn & Putjorn (2023) demonstrated how GenAI could foster creativity among highschool students, and Song & Song (2023) reported improvements in engagement in EFL writing due to GenAI interventions. Furthermore, at the higher education level, Yan (2023) highlighted the positive reception of ChatGPT's in-depth feedback among undergraduates, enhancing their writing skills through detailed analyses and personalised insights, a nuanced approach rarely achieved by traditional AI tools.

Table 5 Challenges of GenAI in education

Coding	Categories	*F*	*RF* (%)	Examples
GenAI imperfections	Hallucination	26	54	De Paoli (2023); Shirafuji et al. (2023)
	Overgeneralised	10	21	AlGhamdi (2024); Bernabei et al. (2023)
	Randomness	6	13	Hu et al. (2024); Wan et al. (2024)
	Data processing limitations	3	6	Alneyadi & Wardat (2024); Hu et al. (2024)
	Inability to generate original ideas	1	2	Keiper et al. (2023)
	Lack of emotional intelligence	1	2	Lu et al. (2024)
Ethical dilemmas	Beneficence	6	11	Maitland et al. (2024); Pinargote et al. (2022)
	Equality	4	8	Baba et al. (2024); Hellas et al. (2023)
	Over-reliance	3	6	Dasari et al. (2024); Song & Song (2023)
	Transparency	3	6	Bernabei et al. (2023); Tossell et al. (2024)
	Academic integrity	3	6	Matthews & Volpe (2023); Yan (2023)
Competency challenges	Utilisation gap	4	8	Urban et al. (2024); Yan et al. (2024b)
	Skill atrophy	3	6	Niloy et al. (2024); Putjorn & Putjorn (2023)
	Assessment validity	1	2	Matthews & Volpe (2023)

Notes. *F* = frequency, *RF* = frequency ratio.

In terms of assessment, GenAI can depart from the rigid scoring systems typical of traditional AI, adopting a more adaptive and authentic evaluation process. While feedback may inform assessment, GenAI's role as an evaluation tool is distinct, focusing on adapting assessments to the varied abilities of students and providing more contextual, nuanced evaluations. This shift is evidenced by the work of Lu et al. (2024), which found that ChatGPT provided comprehensive feedback, adapting to the varied abilities of students. Moreover, studies by Pinargote et al. (2022), Latif & Zhai (2024), and Mansour et al. (2024) underlined GenAI's ability to deliver precise and constructive assessments. This marks a significant advance over previous AI technologies, which often lacked the flexibility to adapt to diverse educational contexts and the insight to generate nuanced feedback tailored to individual learners' needs. Unlike traditional AI, which typically followed rigid, pre-programmed rules, GenAI can dynamically generate content and evaluations that reflect a deeper understanding of the material and the student's performance, leading to more personalised and effective educational outcomes.

In terms of empirical research quantity, GenAI has a substantial amount of studies on feedback in education, with a broad range of perspectives,

demonstrating clear advantages over traditional AI. However, there is a notable lack of empirical research focusing on the assessment, highlighting significant potential for future studies. Addressing this gap through the design of assessment-related experiments could make GenAI a key tool for adaptive, context-aware educational environments and enhance our understanding and support of student learning.

3.3 | RQ3: What Concerns Arise from the Application of GenAI in Education as Evidenced by Empirical Research?

Although in the previous section, we reviewed the empirical studies that found some promise of GenAI, these empirical studies also revealed the related challenges and concerns. In our review, we observed that 37 studies reported concrete concerns associated with GenAI use in education, and only 11 studies did not report any challenges. Based on our review framework, we categorised these challenges into three main types: GenAI imperfections, ethical dilemmas, and competency challenges, as shown in Table 5.

3.3.1 GenAI Imperfections Undermine Feedback Quality

Even though GenAI brings promises in learning and teaching support and feedback or assessment, in our review, we found that 29 studies mentioned that GenAI may also provide incorrect, inconsistent, or irrelevant guidance or suggestions to teachers and students. Essentially, LLMs use a statistical model to estimate the probability of occurrence of each word in a given context, and then select the most likely next word based on the probabilities the model learned (OpenAI, 2024b). With the reliance on patterns of word association rather than true comprehension of the content, LLM responses may not fully grasp the finer nuances of user questions or provide targeted answers (Liu et al., 2023).

Hallucinations were found to be the biggest challenge due to mismatches in training data and the complexity of language generation tasks, leading to outputs misaligned with factual information or lacking sufficient accuracy (Ji et al., 2023). Studies discovered that GenAI's suggestions lack contextual accuracy and provide a high proportion of errors in specific tasks, potentially misleading students and fostering misconceptions (Alneyadi & Wardat, 2024; Essel et al., 2024; Maitland et al., 2024; Song & Song, 2023). For example, Liu et al. (2024) found that Bing generated unnatural images such as "fingers seem distorted," which disappointed students and affected their efficiency in completing learning tasks. The main reason for hallucinations in GenAI tools lies in their technical rationale. Joshi et al. (2024) found that the GenAI was unable to retain and integrate the previous context, resulting in inaccurate responses that did not align with the overall discussion.

Studies discovered that in the process of helping learning, GenAI often generated feedback that was too general, which lacked personalised responses tailored to individual student characteristics. This limitation can be understood through Vygotsky's concept of the Zone of Proximal Development, which emphasizes the importance of tailored guidance in helping students progress from their current level of understanding to higher levels of competence. Vygotsky argued that effective learning occurred when a more knowledgeable other, such as a teacher, offered scaffolding that was specifically adapted to the learner's developmental stage (Vygotsky & Cole, 1978). However, GenAI often lacked a nuanced interpretation of critical learning elements such as knowledge gaps or contextual relevance. As summarised by Bernabei et al. (2023), LLMs tended to produce content that was "shallow and not of high quality." Compared with GenAI, human experts can provide real-time examples, demonstrations, and interactive discussions to help students grasp mathematical ideas; for example, Dasari et al. (2024) found that the mathematics performance of students taught by human experts or through human–AI co-teaching was better compared to students who solely rely on ChatGPT. This implies that while ChatGPT can offer information and answers based on its data, it may not have the same level of expertise or knowledge of individual student needs as human experts.

Studies reported randomness as a specific deficiency because GenAI relies more on language patterns than logical reasoning, making its ability to follow instructions unstable and leading to unpredictable and inconsistent outputs. Since LLMs are not machines based on logic, randomness is a built-in feature of the LLMs. LLMs do not follow the clear logic-based operations that traditional computer programs do (Hu et al., 2024; Joshi et al., 2024; Wan et al., 2024). In the Hellas et al. (2023)'s study of an online introductory programming course, the researchers explicitly asked GPT 3.5 not to produce model solutions and corrected code. However, GPT 3.5 still provided feedback, including revised code or model solution for the learning exercise. Shirafuji et al. (2023) discovered that GPT 3.5 revised an already straightforward, readable, or concise program that did not require refactoring. Plevris et al. (2023) compared three LLMs' abilities in solving mathematics and logic problems and found that LLMs all provided conflicting answers when given the question more than once. This randomness was problematic, especially for questions to which students did not know the exact answer and had to rely on the response of the LLMs to get it.

Existing studies mentioned GenAI's limitations in processing data, such as length constraints and data interpretation. Alneyadi & Wardat (2024) described how ChatGPT cannot read or interpret graphic equations directly, so students had to rely on other resources or traditional teaching methods to understand Quantum Physics graphs effectively. In the Hu et al. (2024)'s study, the limitation of the input data length (limit of 2,048 characters in GPT 3.5) made it difficult to preserve contextual cues within the narrative that may figure into the LLMs response. Some studies found that lack of emotional intelligence and inability to generate original ideas were challenges for GenAI. Keiper et al. (2023) discovered that if a question asked for three elements associated with sports analytics, then ChatGPT would find the answer with only three elements and would miss other elements that could have been critical for a comprehensive answer. Lu et al. (2024) uncovered that ChatGPT's feedback displayed a higher amount of summary and general suggestions, while the teacher's feedback featured a greater amount of praise referred by the students in the evaluation of academic writing.

3.3.2 Ethical Dilemmas in GenAI-Assisted Ill-Structured Tasks

With the advent of deep learning, GenAI, such as ChatGPT, can create new text or visual outputs from large amounts of training data and undertake more ill-structured tasks that were not automated before (Janse van Rensburg, 2024; Urban et al., 2024). An ill-structured task often lacks clear initial states, goals, and intermediate steps (Reed, 2016). The adoption of GenAI in the completion of these tasks raises ethical issues, especially academic integrity, over-reliance, equality, and students' beneficence. In our review, 13 studies have discovered ethical dilemmas, mainly based on qualitative evidence.

Six studies have highlighted concerns regarding the beneficence of the use of GenAI in education. For example, in Pinargote et al. (2022), GenAI was implemented to generate summaries of class activities, general group behaviours with respect to collaboration metrics, the roles of each student during the activity, and the main insights and areas of improvement for each student. However, the utilisation of GenAI raised concerns among students about the fairness of representation. A total of 35% of the participants in this study noted that the information provided by GenAI was unfair and someone could feel uncomfortable being consistently confined to a specific role (Pinargote et al., 2022). In the study of Maitland et al. (2024), ChatGPT showed cognitive bias, such as anchoring bias and confirmation bias, on medical exams, since ChatGPT was tuned to rely too much on pre-existing beliefs or trends in training data. These biases were deemed to negatively affect the trustworthiness of learners in using GenAI as a tool for exam preparation. These biases could lead medical students to make incorrect judgements during their careers, potentially impacting patient care. In Yan (2023), the students expressed their disagreement about treating GenAI as a "shortcut" or "secret recipe" because they would eventually be hurt if GenAI was not controlled and limited.

Studies discovered the lack of transparency in the use of GenAI tools in education. The level of transparency affects educators, students, and stakeholders, as the inner workings of these tools are primarily accessible only to AI experts. As discovered by Bernabei et al. (2023), If ChatGPT were trained on incorrect or misleading data, it would produce inaccurate results without providing an explanation, making it impossible to trace the specific data sources used during its training or to generate its answers. Tossell et al. (2024) investigated the perceived trustworthiness of ChatGPT and found that some participants remained sceptical about ChatGPT's capabilities due to the lack of metadata and demanded a clearer explanation.

Equality is another significant ethical issue presented in our sample. Yan (2023) investigated students' perceptions about GenAI-assisted writing. They found that participants expressed their worries about the power of ChatGPT to generate a piece of writing "in the blink of an eye," which gave students "enormous advantages" to outperform their peers. Baba et al. (2024) found that an AI-powered educational platform inadvertently perpetuated societal biases present in its training data. This AI system's recommendation and penalisation features may perpetuate existing disparities in academic achievement among students from diverse populations, raising concerns about impartiality. Instead of providing equal learning opportunities to all students, the system's recommendation and penalisation features may perpetuate existing disparities in academic achievement among students from diverse populations. Studies also found equality issues between different cultures. For instance, Hellas et al. (2023) found that LLMs' responses to prompts in a non-English language were slightly worse than responses to English prompts when offering help to beginner programmers. Wan et al. (2024) also discovered cultural misunderstandings presented by ChatGPT during human–AI collaborative creative writing.

Concerns about students becoming overly dependent on GenAI in education have been documented in three studies. Empirical evidence indicates that the concern of over-reliance mainly originated from the students themselves. Students who

participated in the studies by Dasari et al. (2024), Liu et al. (2024), and Song & Song (2023), all voiced concerns that excessive reliance on ChatGPT could hinder intellectual development, demonstrating a decline in students' ability to analyse and synthesise information independently. Academic integrity is another issue that raises concerns among students and teachers. In Liu et al. (2024)'s study, students noted that GenAI's ability to elude detection and recognition was shockingly impactful, potentially undermining academic standards. Students also argued that the chances for the AI-generated document to be detected by existing plagiarism detection software were minimal (Liu et al., 2024).

3.3.3 Mismatches in GenAI-Enabled Teaching Evolution and User Competency

Studies have discovered mismatches between the teaching changes brought about by GenAI and user capabilities. These mismatches mainly occur as utilisation gaps, challenges to assessment validity, and skill atrophy in our review. Although only seven articles in our sample reported this type of challenge empirically, more evidence may likely appear with appropriate utilisation in the long run.

We identified a utilisation gap related to the use of GenAI when users lack the necessary skills or knowledge. Nguyen et al. (2024) found that low-performing students interacted with the GenAI-powered tool in a linear way—prompting, copying, and pasting—without the depth of usage observed in the high-performing learners. This suggests a lack of critical engagement with AI-generated content, possibly due to limited executive control or insufficient integration of the AI tool into their cognitive writing process. In another example, Liu et al. (2024) mentioned that students were sometimes not satisfied with AI image styles because the students often forgot to express their style preferences in their text-to-image prompts. Urban et al. (2024) found that ChatGPT boosted self-efficacy of the learners, but the benefits of higher self-efficacy for actual performance may have been limited. Higher self-efficacy was associated with higher bias indices (i.e., greater overestimation) of one's performance. In the study by Yan et al. (2024b), students described their difficulties in correcting miscommunications with GenAI. They pointed out that identifying GenAI's mistakes and seeking solutions together was challenging. These tasks require self-regulated learning skills to ensure successful learning outcomes, such as setting learning goals, learning new codes, adjusting strategies, and finding the cause of an "error" with GenAI.

Skill atrophy observed in several studies was viewed, especially in writing experiments. For example, Niloy et al. (2024) found the scores of the experimental group, who wrote with the assistance of ChatGPT, showed a significant decrease in the post-test compared to the pre-test. This suggests that using GenAI may have negatively impacted students' creative writing abilities. In the domain of university essays, Bernabei et al. (2023) investigated the students' trust in GenAI and found that most of the sample worried that ChatGPT posed a threat to creativity and originality, especially after experiencing it. Putjorn & Putjorn (2023) also discovered that students were concerned regarding GenAI's potential negative impacts of job displacement for creators.

GenAI also presents significant risks to the validity of educational assessment. Matthews & Volpe (2023) conducted a Turing imitation game to explore whether academics could reliably recognise the texts that AI had generated. They found that academics were only able to identify GenAI-generated texts a half of the time, which brings great challenges to the practice of assessment, implying the need for redesigning assessments that succeed in providing robust evidence of student achievement of learning outcomes.

3.4 | RQ4: What Are the Needs and Future Directions for GenAI in Education Identified in the Empirical Research?

Among the 48 studies included in the review, except for six studies that mainly focused on fine-tuning the LLMs and improving the performance of GenAI, other studies clearly identified future research needs and directions for applying GenAI in education. These needs and future directions mainly included AI literacy, integration of GenAI, evidence-based decision-making, and methodological rigour.

3.4.1 AI Literacy

AI literacy is consistently highlighted as a critical area for future development in the studies viewed. Twelve studies highlight the critical importance of training programs for educators, emphasising the need for acquiring fundamental AI skills, enhancing AI literacy, and cultivating a rational understanding of the limitations of GenAI to effectively integrate it into educational practices. For instance, according to Table 6, Yilmaz & Yilmaz (2023) specifically recommended that teachers acquire AI literacy skills to effectively teach students' prompt-writing techniques, while Tossell et al. (2024) argued that institutions must train educators to use AI both responsibly and creatively. Among the twelve studies, four note the need for educators to learn how to employ GenAI tools productively while avoiding potential pitfalls as seen in

Table 6 Needs and future directions of GenAI in education.

Coding	Categories	*F*	*RF* (%)	Examples
AI literacy	Training programs for educators	12	25	Lu et al. (2024); Yilmaz & Yilmaz (2023)
	Ethical awareness training	10	21	Bernabei et al. (2023); Matthews & Volpe (2023);
	Comprehensive AI literacy curricula	4	8	Juanda & Afandi (2024); Wachira et al. (2023)
	AI Literacy measurement tools	0	0	
Integration of GenAI	Integrating GenAI across scenarios	11	23	Villan & Santos (2023); Wan et al. (2024)
	Integrating GenAI with humans experts/teachers	8	17	AlGhamdi (2024); Kieser et al. (2023)
	Integrating GenAI with students	2	4	Yan et al. (2024b); Zhang & Huang (2024)
	Integrating GenAI with other technologies	2	4	Keiper et al. (2023); Yeh (2024)
Evidence-based decision making	Robust learning impacts evidence	12	25	Baba et al. (2024); Joshi et al. (2024)
	Metacognitive and cognitive skills engagement	10	21	Niloy et al. (2024); Putjorn & Putjorn (2023)
	Multi-party collaborative efforts	1	2	de Vicente-Yagüe-Jara et al. (2023)
Methodological rigour	Rigorous methodological standards	5	10	Tai et al. (2024); Theelen et al. (2024);
	Evidence quality appraisal instrument	2	4	Essel et al. (2024); Morgan (2023)
	Transparent prompting and generation disclosure	1	2	De Paoli (2023)

Notes. *F* = frequency, *RF* = frequency ratio.

Gilson et al. (2023), Hellas et al. (2023), Riedel et al. (2023), and Uddin et al. (2023).

A total of 19 studies underscore the significance of ethical awareness training, arguing that understanding the ethical implications and potential challenges associated with GenAI is essential for both educators and students. For example, Matthews & Volpe (2023) employed Turing's Imitation Game, revealing that even academics can struggle to distinguish between texts generated by ChatGPT and those written by humans. This finding highlights the need for institutions to prioritise the development of students' abilities to use GenAI technology both effectively and ethically. Similarly, Putjorn & Putjorn (2023) argued that exploring the ethical dimensions of AI usage and formulating responsible AI guidelines can strengthen AI-focused curricula, thus preparing students for an AI-driven future.

In our review, four studies emphasised the importance of developing comprehensive AI literacy curricula to enhance students' AI literacy. For example, Wachira et al. (2023) found that within just two months of ChatGPT's introduction, one-third of the students had already incorporated it into their daily routines. Wachira et al. (2023) argued that if the curriculum does not evolve to foster critical thinking skills, students may rely on AI models to circumvent academic challenges, potentially rendering themselves obsolete in the professional world. Similarly, Keiper et al. (2023) underscored the need for students to preserve human reasoning, the elegance of the human mind, and the core principles of knowledge generation, especially given AI's limitations due to data bias.

3.4.2 Integration of AI

The integration of GenAI with human educators, subject matter experts, and students remains a key focus in research, with 11 studies emphasising the need for future investigations to replicate findings across diverse educational contexts and explore collaborative synergies between AI and human participants. Juanda & Afandi (2024) and Essel et al. (2024) emphasised the importance of future research replicating their research findings across diverse educational settings, universities, countries, and student demographics. For example, Watts et al. (2023) noted the evolving capabilities of GenAI in image analysis and stressed the need to continuously assess chatbots' performance across different chemistry tasks. Urban et al. (2024) also advocated for future studies to include a wide-ranging participant pool encompassing various ages, backgrounds, education levels, disciplines, and cultural contexts. The replication is critical to validating consistency and exploring potential variations in GenAI utilization and effectiveness, thereby enhancing the generalizability of the presented findings.

Researchers generally posit that GenAI still necessitates integration with the expertise of human educators or subject matter specialists. In other words, human–GenAI collaboration. AlGhamdi (2024) highlighted ChatGPT's potential in augmenting traditional feedback mechanisms, particularly in classes of high student-to-teacher ratios. However, GenAI currently lacks the nuanced understanding and personalization offered by human instructors. In another study, Pinargote et al. (2022) demonstrated how GenAI can automate data narratives in Learning Analytics Dashboards for collaborative learning

scenarios, illustrating GenAI's potential in streamlining evaluation processes and emphasizing the importance of collaborative synergy between human and AI evaluators. Consequently, the study advocates for a synergistic approach wherein ChatGPT-generated feedback is utilized alongside human oversight (Dasari et al., 2024; de Vicente-Yagüe-Jara et al., 2023).

A total of two studies pointed out that future research avenues could explore the integration and collaboration between AI and students. Zhang & Huang (2024) explored the potential of GenAI to enhance cooperation and sharing among students and their ability to foster higher-order thinking skills, such as reasoning abilities. Yan et al. (2024b) highlighted the effective collaboration between students and GenAI that could enhance students' metacognition and self-regulated learning skills.

3.4.3 Evidence-Based Decision-Making

Research on the effects of artificial intelligence technology in education has yielded encouraging results, but stakeholders need to better gauge the effectiveness of the interventions and provide stronger evidence to support educational decision-making. For example, Liu et al. (2024) compared EFL student writers' digital multimodal composing and traditional writing processes. They found that students behaved differently in these two ways, and students created more transitional texts and examples and tended to use summarized search results from Bing Chat. However, as Liu et al. (2024) admitted, robust learning impact evidence was needed to probe students' cognitive process, such as their affective and cognitive accounts in getting involved in the GenAI-assisted writing. Joshi et al. (2024) and Wu et al. (2024) both planed to conduct controlled experiments inside and outside the classroom to further establish the validity of their findings.

Researchers are currently exploring how GenAI technologies support educational goals and foster the development of cognitive and metacognitive skills. According to Song & Song (2023), students with GenAI-assisted instruction demonstrated enhanced proficiency in various aspects of writing, including organisation, coherence, grammar, and vocabulary, but they concerned about contextual accuracy and over-reliance.

Only one study included emphasise multi-party collaborative efforts, appealing to researchers, practitioners, and policymakers collaboratively generate robust evidence that can guide the effective and responsible use of AI in education. de Vicente-Yagüe-Jara et al. (2023) stressed that changes in educational practice in the coming years will be determined by AI developments and would need to rely on research in full collaboration with teachers, educational leaders, and students to ensure that appropriate educational policies are put in place.

3.4.4 Methodological Rigour

Building on the discussions on evidence-based decision-making, it is essential to emphasise the importance of methodological rigour in the application of GenAI technologies within educational research. A total of 6 studies delve into the potential utilisation of LLMs as effective coding and inter-rater reliability research instruments due to their capacity to extract and analyse extensive datasets, ultimately converging toward solutions. This intrinsic attribute of LLMs aligns adeptly with the demands of qualitative research, particularly in discerning meaning and trends within non-ordinal data. For instance, Tai et al. (2024) envisioned that the primary application of the LLM would be to provide confirmatory data for a coding analysis, thereby facilitating more reliable and insightful findings. Most studies are optimistic about GenAI advancements. However, Keiper et al. (2023) also asserted the ongoing importance of human expert judgements in verifying responses provided by ChatGPT.

4 Discussion

4.1 | Factors Impacting the Effectiveness of GenAI in Education

Based on our review, we believe that the effectiveness of GenAI in education is influenced by several key factors. One critical element is the competency and AI readiness of students. As GenAI tools are designed to enhance the learning experience, their efficacy is largely dependent on the users' competency to effectively interact with these technologies (Lu et al., 2024; Nguyen et al., 2024). Students who possess a higher degree of digital or AI literacy and familiarity with GenAI-driven tools are more likely to utilise these technologies to their fullest potential (Lu et al., 2024). This readiness encompasses not only technical skills but also an understanding of how to critically engage with GenAI outputs (Yan et al., 2024b). Students who are adept at using GenAI tools can leverage them for improved learning outcomes (Yan et al., 2024b), while those lacking in this competency may struggle to derive the same benefits (Nguyen et al., 2024), thereby exacerbating existing educational inequalities. Addressing the challenge of varying AI competencies through targeted interventions can transform this challenge into a promise of more inclusive and equitable education (Yilmaz & Yilmaz, 2023).

Another significant factor impacting the application of GenAI in education is the nature of the tasks to which GenAI is applied. GenAI excels in tasks that require constructive and expansive thinking, such as writing (Liu et al., 2024; Lu et al., 2024; Niloy et al., 2024; Yan, 2023) and content creation (Phung et al., 2023; Shirafuji et al., 2023; Yeh, 2024). However, its efficacy diminishes when applied to tasks that demand logical reasoning and critical analysis (Hellas et al., 2023; Plevris et al., 2023). For instance, while ChatGPT may generate contextually relevant essays or stories, it may not perform as effectively in solving complex mathematical problems (Plevris et al., 2023). This delineation between generative and logical tasks presents both challenges and promises. The challenge lies inappropriately aligning the capabilities of GenAI with the specific requirements of educational tasks. Educators must be discerning in their application of AI tools, ensuring they are used where they add the most value (Lu et al., 2024). In contrast, this task- specific effectiveness of GenAI can be harnessed as a promise to innovate educational practices. By leveraging GenAI for tasks that it handles well, educators can free up time to focus on areas where human insight and judgement are indispensable (Dasari et al., 2024; Pinargote et al., 2022). For example, GenAI can be used to generate first drafts of student essays, allowing educators to concentrate on providing detailed feedback and fostering critical thinking skills. However, across various activities, the quality of GenAI-generated content is predominantly influenced by the prompts. Moreover, writing is inherently a complex process, requiring not just relevant responses but also logical reasoning, critical thinking, and values-based decision-making, all of which are vital for fostering thoughtful and creative expression. Furthermore, understanding the limitations of GenAI in logical tasks encourages the development of hybrid approaches that combine AI-driven creativity with human analytical skills (AlGhamdi, 2024; Dasari et al., 2024). This symbiotic relationship may lead to more comprehensive educational experience, where the strengths of both AI and human intelligence are utilised to their fullest promise.

4.2 | Collaboration between AI and Teachers in Education

Following the last point discussed above, incorporating GenAI, such as ChatGPT, into education, in contrast to traditional teacher-centred teaching methods, offers both unique advantages and certain risks. A key difference is that while GenAI delivers information and support, it lacks the nuanced understanding and emotional engagement that human teachers provide (Dasari et al., 2024; Phung et al., 2023). This difference raises concerns that reliance on GenAI for creative tasks could dampen students' creativity due to GenAI's predictable responses, which may not encourage original thinking or problem-solving skills (Niloy et al., 2024). However, others argue that GenAI could actually foster creativity by introducing students to diverse perspectives and novel information, thus promoting innovative thinking (Essel et al., 2024; Putjorn & Putjorn, 2023; Song & Song, 2023). The collaborative benefits of integrating GenAI with human teachers are significant, combining GenAI's data processing capabilities with teachers' ability to provide context and adapt lessons to enhance learning outcomes (Dasari et al., 2024) and alleviate routine teaching burdens (Villan & dos Santos, 2023). Nonetheless, this integration must be carefully managed to prevent over-reliance on AI, which may undermine students' critical thinking and interpersonal skills. A strategic approach is essential to balance the use of GenAI in educational settings effectively.

4.3 | Important but Neglected Topic: Students' Metacognition When Learning with GenAI

As shown in our review, agents in education based on GenAI exhibit human-like cognitive abilities, including learning support, task planning, situational assessment, progress monitoring, and collaborative efforts between agents (Yan et al., 2024a). However, GenAI can also be a double-edged sword when off-loading cognitive and metacognitive load on students because this off-loading may reduce the difficulty of the task while also inhibiting students' active thinking and reflection. This concern is particularly significant in educational environments that prioritise outcomes over processes, where students maybe more prone to taking shortcuts with the help of GenAI. The above-mentioned over-reliance issue is particularly noteworthy in terms of metacognition and self-regulated learning, which has rarely been addressed in current empirical studies. For example, several studies were viewed reported that GenAI performed well in improving students' task performance; however, there is little discussion on whether such "performance improvement" can be the result of what we may call "AI-empowered learning skills" which optimize tasks performance through the use of AI capabilities at the expense of developing genuine human skills. Only one study discussed that, even though ChatGPT reduced students' perceived difficulty, they also experienced difficulty calibrating their self-assessment, which is a crucial part in both metacognition and self-regulated learning (Urban et al., 2024). Therefore, researchers should more focus on metacognitive aspects of human–AI interaction, and educational stakeholders should be more aware of

students' over-reliance on AI and potentially metacognitive laziness issue. As defined by Fan et al., (2024), metacognitive laziness is "students' dependence on AI assistance, offloading metacognitive load, and less effectively associating responsible metacognitive processes" during learning. It is critical that future educational strategies balance AI's capabilities with the need to cultivate independent metacognitive practices, ensuring that AI supplements rather than replaces critical human-mediated learning interactions. This approach can help mitigate the risk of metacognitive laziness and support sustainable, genuine skill development in students.

4.4 | Creative Usage of GenAI in Teaching and Learning

In our analysis, we discovered that despite various technical limitations, such as hallucination and randomness in GenAI, these factors do not impede the application of teachers' pedagogical knowledge. In fact, leveraging the shortcomings of GenAI for teaching purposes is a key area for teachers to explore as an innovative teaching strategy. For instance, hallucinations created by GenAI bring opportunities in specific situations. Wan et al. (2024) found that in the context of creative writing, the seemingly irrelevant information or erroneous outputs generated by ChatGPT, such as repetitions and loops, provided unexpected inspiration to students who wrote stories. Furthermore, when GenAI was used for qualitative coding, De Paoli (2023) found that the model hallucinated by assigning a code to them, but the researchers treated this hallucination as a chance to foster discussion rather than provide solutions. In addition to cleverly using hallucinations, the opacity and over-generalisation of GenAI feedback can also be used to encourage students to engage inactive exploration. For example, previous research found that eventhough ChatGPT was found to lack in-depth explanations for complex topics occasionally, this feature prompted students to seek additional support from textbooks, academic articles, and online resources to supplement their learning (Alneyadi & Wardat, 2024).

The above practices indicated that rather than applying GenAI to every aspect of teaching, researchers currently recommend teachers to use their creativity to employ GenAI in specific scenarios where the GenAI capabilities are most suitable. For example, teachers were encouraged to leverage ChatGPT's capabilities to create prompts for open-ended questions that help to elicit students' thinking. Teachers were also advised to provide students with quality examples and exemplars of responses and learning tasks and to generate glossaries of terms and definitions relevant to the syllabus or unit of study (Baidoo-Anu & Owusu Ansah, 2023; Herft, 2023). Additionally, ChatGPT can be used to visualise the creative content in students' minds. Teachers can also encourage students to identify errors in GenAI outputs and compare their own work with GenAI-generated content (Pavlik & Pavlik, 2024). These approaches can help students develop a deeper understanding of the subject matter and foster a critical perspective on the use of GenAI.

4.5 | Potentials in Transforming Assessments

The integration of GenAI in educational assessments is transforming traditional methods into more adaptive and authentic processes (Swiecki et al., 2022). GenAI enables the development of personalised assessments through autonomous agents such as AgentGPT (Schacht et al., 2024). AgentGPT exhibits advanced cognitive abilities and operates independently, exemplified by agents organising complex events autonomously in dynamic environments. Furthermore, GenAI shows potential in improving the realism in simulations for more meaningful assessments, integrating multi-modal models to closely mimic real-world tasks, thus improving training and evaluation in fields such as medicine (Song et al., 2022). However, despite these advancements, empirical research primarily focuses on GenAI for immediate feedback rather than comprehensive outcome-based assessments. There remains a significant gap in validating GenAI's effectiveness in measuring and enhancing learning outcomes, marking a crucial direction for future research.

4.6 | Future Works, Limitations, and Conclusions

This study systematically reviews 48 empirical studies that explore the applications of GenAI in education, summarising their general characteristics and empirical findings in terms of promises and concerns, while also highlighting current needs and future directions. The promises are divided into four key areas: learning support, teaching support, feedback, and assessment, showing GenAI's vast potential for future applications in education. However, the literature on teaching support and assessment is notably sparse compared to that on learning support and feedback, indicating an urgent need for further empirical research to address these gaps. Beyond promises, there are significant concerns, including compromised feedback reliability due to GenAI imperfections, ethical dilemmas in difficult-to-codify tasks, and mismatches between the evolution of AI-enabled teaching and user competencies. These issues serve as critical cautions for incorporating GenAI into educational settings. Finally, there

is a pronounced demand and potential for future exploration in areas such as AI literacy, AI integration, evidence-based decision-making, and methodological rigour, which are essential for advancing GenAI research and applications in education.

Several limitations of this review must be acknowledged. Firstly, due to the rapid development in the GenAI sector, this review may not encompass all related studies or technological advancements, potentially overlooking some new insights. Secondly, the brevity of this paper restricts a thorough comparative analysis between traditional AI approaches and GenAI might leave the reader with an incomplete understanding of their respective nuances. Lastly, the review lacks a comprehensive theoretical overview and a detailed discussion of contributions from the studies examined, which could hinder a deeper understanding of the theoretical and practical advancements. Future research or reviews should aim to address these gaps, enhancing both the breadth and depth of the literature in this swiftly evolving domain. Ultimately, it is hoped that this review will serve as a reference for the design of future empirical research in GenAI, promoting further exploration in critical areas such as AI literacy, AI integration, evidence-based decision-making, and methodological rigour, all essential for advancing GenAI research and applications in education.

Acknowledgement This study was funded by the National Social Science Fund of China (Grant No. VGA230012) and National Natural Science Foundation of China (Grant No. 62407001).

Conflict of Interest Qiong Wang is a member of the Editorial Board of *Frontiers of Digital Education*, who was excluded from the peer-review process and all editorial decisions related to the acceptance and publication of this article. Peer-review was handled independently by the other editors to minimise bias.

Data Availability Statements The authors confirm that all data generated or analysed during this study are included in this published article.

References

Alam, A. (2023). Harnessing the power of AI to create intelligent tutoring systems for enhanced classroom experience and improved learning outcomes. In: Rajakumar, G., Du, K. L., Rocha, Á., eds. *Intelligent Communication Technologies and Virtual Mobile Networks*. Singapore: Springer, vol. 171, 571–591.

AlBadarin, Y., Tukiainen, M., Saqr, M., & Pope, N. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, *3*, 60.

AlGhamdi, R. (2024). Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study. *Education and Information Technologies*, 1–26.

Alier, M., García-Peñalvo, F., & Camba, J. D. (2024). Generative artificial intelligence in education: From deceptive to disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(5), 5.

Alneyadi, S., & Wardat, Y. (2024). Integrating ChatGPT in grade 12 quantum theory education: An exploratory study at Emirate School (UAE). *International Journal of Information and Education Technology*, *14*(3), 38.

Anthropic. (2024, May 30). Retrieved from Anthropic website.

Baba, K., Faddouli, E., & Cheimanoff, N. (2024). Mobile-optimized AI-driven personalized learning: A case study at Mohammed VI Polytechnic University. *International Journal of Interactive Mobile Technologies*, *18*(4), 81–96.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62.

Bannister, P., Urbieta, A. S., & Peñalver, E. A. (2023). A systematic review of generative AI and (English medium instruction) higher education. *Aula Abierta*, *52*(4), 401–409.

Bernabei, M., Colabianchi, S., Falegnami, A., & Costantino, F. (2023). Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence*, *5*, 100172.

Bozkurt, A. (2023). Unleashing the potential of generative AI, conversational agents and chatbots in educational praxis: A systematic review and bibliometric analysis of GenAI in education. *Open Praxis*, *15*(4), 261–270.

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: The growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*, *21*(8), 471–496.

Dasari, D., Hendriyanto, A., Sahara, S., Suryadi, D., Muhaimin, L. H., Chao, T., & Fitriana, L. (2024). ChatGPT in didactical tetrahedron, does it make an exception. A case study in mathematics teaching and learning. *Frontiers in Education*, *8*, 1295413.

de la Torre, A., & Baldeon-Calisto, M. (2024). Generative artificial intelligence in Latin American higher education: A systematic literature review. In: *Proceedings of 2024 12th International Symposium on Digital Forensics and Security*. 1–7.

de Vicente-Yagüe-Jara, M. I., López-Martínez, O., Navarro-Navarro, V., & Cuéllar-Santiago, F. (2023). Writing, creativity, and artificial intelligence: ChatGPT in the university context. *Comunicar: Media Education Research Journal*, *31*, 45–54.

De Paoli, S. (2023). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019.

Dengel, A., Gehrlein, R., Fernes, D., Görlich, S., Maurer, J., Pham, H. H., Großmann, G., & Eisermann, N. D. G. (2023).

Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education. *Informatics*, *10*(4), 78.

Epstein, Z., Hertzmann, A., Herman, L.M., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., Farid, H., Leach, N., Pentland, A., & Russakovsky, O. (2023). Art and the science of generative AI. *Science*, *380*, 1110–1111.

Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, *59*, 272–287.

Essel, H. B., Vlachopoulos, D., Essuman, A. B., Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence*, *6*, 100198.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., Gasevic, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*. (in press).

Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, *7*, 549–569.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*, e45312.

Goodfellow, I. (2016). *Deep learning*. Cambridge: The MIT Press.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245.

Hellas, A., Leinonen, J., Sarsa, S., Koutcheme, C., Kujanpää, L., & Sorva, J. (2023). Exploring the responses of large language models to beginner programmers' help requests. In: *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*. Chicago, 93–105.

Herft, A. (2023). *A teacher's prompt guide to ChatGPT aligned with 'what works best.'* CESE NSW What Works Best in Practice.

Hu, Y., Goktas, Y., Yellamati, D. D., & De Tassigny, C. (2024). The use and misuse of pre-trained generative large language models in reliability engineering. In: *Proceedings of 2024 Annual Reliability and Maintainability Symposium*. 1–7.

Janse van Rensburg, J. (2024). Artificial human thinking: ChatGPT's capacity to be a model for critical thinking when prompted with problem-based writing activities. *Discover Education*, *3*, 42.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*, 1–38.

Joshi, I., Budhiraja, R., Dev, H., Kadia, J., Ataullah, M. O., Mitra, S., Akolekar, H. D., Kumar, D. (2024). ChatGPT in the classroom: An analysis of its strengths and weaknesses for solving undergraduate computer science questions. In: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V*. 625–631.

Juanda, J., & Afandi, I. (2024). Assessing text comprehension proficiency: Indonesian higher education students vs ChatGPT. *XLinguae*, *17*, 49–68.

Karthikeyan, C. (2023). Literature review on pros and cons of ChatGPT implications in education. *International Journal of Science and Research*, *12*, 283–291.

Keiper, M. C., Fried, G., Lupinek, J., & Nordstrom, H. (2023). Artificial intelligence in sport management education: Playing the AI game with ChatGPT. *Journal of Hospitality, Leisure, Sport & Tourism Education, 33*, 100456.

Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, *19*, 020150.

Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., Kasneci, E., Kasneci, G., Kuhr, T., Kutyniok, G., Malone, S., Sailer, M., Schmidt, A., Stadler, M., Weller, J., & Kuhn, J. (2024). Are large multimodal foundation models all we need? On opportunities and challenges of these models in education. *EdArXiv*.

Lang, O., Yaya-Stupp, D., Traynis, I., Cole-Lewis, H., Bennett, C. R., Lyles, C. R., Lau, C., Irani, M., Semturs, C., Webster, D. R., Corrado, G., Hassidim, A., Matias, Y., Liu, Y., Hammel, N., & Babenko, B. (2024). Using generative AI to investigate medical imagery models and datasets. *EBioMedicine*, *102*, 105075.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 100174.

Lee, U., Han, A., Lee, J., Lee, E., Kim, J., Kim, H., & Lim, C. (2023). Prompt aloud!: Incorporating image-generative AI into STEAM class with learning analytics using prompt data. *Education and Information Technologies*, *29*, 9575–9605.

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, 104977.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017.

Lo, C. K. (2023). What is the impact of ChatGPT on education. A rapid review of the literature. *Education Sciences*, *13*(4), 410.

Lo, C. K., Hew, K. F., & Jong, M. S. Y. (2024). The influence of ChatGPT on student engagement: A systematic review and

future research agenda. *Computers & Education*, *219*, 105100.

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing. *Assessment & Evaluation in Higher Education*, *49*(5), 616–633.

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, *49*(5), 651–664.

Maitland, A., Fowkes, R., & Maitland, S. (2024). Can ChatGPT pass the MRCP (UK) written examinations. Analysis of performance and errors using a clinical decision-reasoning framework. *BMJ Open*, *14*, e080558.

Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? *arXiv Preprint*, arXiv:2403.06149.

Matthews, J. A., & Volpe, C. R. (2023). Academics' perceptions of ChatGPT-generated written outputs: A practical application of Turing's imitation game. *Australasian Journal of Educational Technology*, *39*, 82–100.

Meta AI. (2024, May 30). *Meta AI*. Retrieved from Meta Llama.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*, 856.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, *8*(5), 336–341.

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, *12*, 153.

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, *22*, 16094069231211248.

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human–AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864.

Nguyen Thanh, B., Vo, D. T. H., Nguyen Nhat, M., Pham, T. T. T., Thai Trung, H., & Ha Xuan, S. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59–81.

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, *40*(2), 919–930.

OpenAI. (2024a, May 30). How ChatGPT and our language models are developed. Retrieved from OpenAI Help Center.

OpenAI. (2024b, May 30). OpenAI. Retrieved from OpenAI website.

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*, 84–93.

Pavlik, J. V., & Pavlik, O. M. (2024). Art education and generative AI: An exploratory study in constructivist learning and visualization automation for the classroom. *Creative Education*, *15*, 601–616.

Phung, T., Pădurean, V. A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., & Soares, G. (2023). Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. In: *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, 41–42.

Pinargote, A., Calderón, E., Cevallos, K., Carrillo, G., Chiluiza, K., & Echeverria, V. (2022). Automating data narratives in learning analytics dashboards using GenAI. In: *Proceedings of 2024 Joint of International Conference on Learning Analytics and Knowledge Workshops*, Aachen: CEUR-WS, 150–161.

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI, 4*, 949–969.

Pradana, M., Elisa, H. P., & Syarifuddin, S. (2023). Discussing ChatGPT in education: A literature review and bibliometric analysis. *Cogent Education*, *10*, 2243134.

Putjorn, T., & Putjorn, P. (2023). Augmented imagination: Exploring generative AI from the perspectives of young learners. In: *Proceedings of 2023 15th International Conference on Information Technology and Electrical Engineering*. Chiang Mai: IEEE, 353–358.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*, 5783.

Reed, S. K. (2016). The structure of ill-structured (and well-structured) problems revisited. *Educational Psychology Review*, *28*, 691–716.

Riedel, M., Kaefinger, K., Stuehrenberg, A., Ritter, V., Amann, N., Graf, A., Recker, F., Klein, E., Kiechle, M., Riedel, F., & Meyer, B. (2023). ChatGPT's performance in German OB/GYN exams—paving the way for AI-enhanced medical education and clinical practice. *Frontiers in Medicine*, *10*, 1296615.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. 3rd ed. Pearson.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*, 887.

Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv Preprint*, arXiv:2306.13906.

Schacht, S., Kamath Barkur, S., & Lanquillon, C. (2024). Generative agents to support students learning progress. In: *Proceedings of the 5th International Conference Business Meets Technology*.

Sevnarayan, K., & Potter, M. A. (2024). Generative artificial intelligence in distance education: Transformations, challenges, and impact on academic integrity and student

voice. *Journal of Applied Learning and Teaching*, *7*(1), 1–11.

Shirafuji, A., Oda, Y., Suzuki, J., Morishita, M., & Watanobe, Y. (2023). Refactoring programs using large language models with few-shot examples. *arXiv Preprint*, arXiv:2311.11690.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843.

Song, W., Hou, X., Li, S., Chen, C., Gao, D., Sun, Y., Hou, J., & Hao, A. (2022). An intelligent virtual standard patient for medical students training based on oral knowledge graph. *IEEE Transactions on Multimedia*, *25*, 6132–6145.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 16094069241231168.

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIED for personalised learning pathways. *Electronic Journal of e-Learning*, *20*, 639–653.

Tedre, M., Kahila, J., & Vartiainen, H. (2023). Exploration on how co-designing with AI facilitates critical evaluation of ethics of AI in craft education. In: *Proceedings of Society for Information Technology & Teacher Education International Conference*. 2289–2296.

Theelen, H., Vreuls, J., & Rutten, J. (2024). Doing research with help from ChatGPT: Promising examples for coding and inter-rater reliability. *International Journal of Technology in Education*, *7*(1), 1–18.

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & de Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies*, *17*, 1069–1081.

Uddin, S. J., Albert, A., Ovid, A., & Alsharef, A. (2023). Leveraging ChatGPT to aid construction hazard recognition and support safety education and training. *Sustainability*, *15*, 7121.

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, *215*, 105031.

Vargas-Murillo, A. R., de la Asuncion, I. N. M., & de Jesús Guevara-Soto, F. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Educational Research, 22*, 122–135.

Vázquez-Cano, E., Ramirez-Hurtado, J. M., Saez-Lopez, J. M., & Lopez-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity*, *49*, 101380.

Villan, F., & dos Santos, R. P. (2023). ChatGPT as co-advisor in scientific initiation: Action research with project-based learning in elementary education. *arXiv Preprint*, arXiv:2311.14701.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Cambridge: Harvard University Press.

Wachira, K., Wachira, L. N., Mwangi, E., Absaloms, H. O., & Jeon, G. (2023). Tertiary education integrity in a ChatGPT conscious world: Preliminary Kenyan Observations. In: *Proceedings of 2023 IEEE AFRICON, Nairobi*. IEEE, 1–6.

Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., & Lu, Z. (2024). "It felt like having a second mind": Investigating human–AI co-creativity in prewriting with large language models. *Proceedings of the ACM on Human–Computer Interaction*, *8*, 1–26.

Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J. M. G. (2023). Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn assignments. *Journal of Chemical Education*, *100*, 3806–3817.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*, 1122–1136.

Wu, T. T., Lee, H. Y., Li, P. H., Huang, C. N., Huang, Y. M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research*, *61*, 3–31.

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 610–625.

Xu, X., Chen, Y., & Miao, J. (2024). Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: A systematic scoping review. *Journal of Educational Evaluation for Health Professions*, *21*, 6.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*, 13943–13967.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024a). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behavior*. (in press).

Yan, W., Nakajima, T., & Sawada, R. (2024b). Benefits and challenges of collaboration between students and conversational generative artificial intelligence in programming learning: An empirical case study. *Education Sciences*, *14*, 433.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 576–584.

Yeh, H. C. (2024). The synergy of generative AI and inquiry-based learning: Transforming the landscape of English teaching and learning. *Interactive Learning Environments*. (in press).

Yilmaz, R., & Yilmaz, F. G. K. (2023). The effect of generative

artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*, *4*, 100147.

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, *21*, 21.

Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, *10*(3), e25370.