Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-025-09923-w

ORIGINAL RESEARCH

Unlocking the Future: A Comprehensive Review of ChatGPT in Education

Deoksoon Kim[1] · Chuqi Kiki Wang[1] · Katrina Borowiec[2] · Noa Rein[1] · Jun-Hyeop Alex Cho[1] · Jiayu Jocelyn Liu[1]

Received: 20 July 2025 / Revised: 13 October 2025 / Accepted: 28 October 2025

Keywords ChatGPT · Personalized learning · Equitable learning · Critical AI literacy · Learner motivation · Innovative pedagogy

1 Introduction

As technology continues to reshape our lives, the emergence of generative artificial intelligence (AI) is having an increasingly important impact on education. ChatGPT—a text-generative chatbot that can simulate conversations with human users—was launched by the AI research company OpenAI in 2022 (OpenAI, 2025), and it has gained considerable attention for its potential to transform all aspects of teaching and learning (Tate et al., 2023; Williams, 2023).

Throughout this review, we use the term generative AI rather than the broader term AI to emphasize our focus on large language models such as ChatGPT. While artificial intel-

Deoksoon Kim
deoksoon.kim@bc.edu

Chuqi Kiki Wang
wangevv@bc.edu

Katrina Borowiec
katrina.borowiec@bc.edu

Noa Rein
reinn@bc.edu

Jun-Hyeop Alex Cho
choabo@bc.edu

Jiayu Jocelyn Liu
liucbk@bc.edu

[1] Teaching, Curriculum, and Society Lynch School of Education & Human Development, Boston College, Campion 127, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

[2] Measurement, Evaluation, Statistics, and Assessment (MESA), Lynch School of Education & Human Development, Boston College, Campion 127, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

ligence encompasses a wide range of technologies (e.g., predictive analytics, adaptive learning systems, machine vision), our analysis is limited to generative AI tools that create new content, such as text, images, or code, in response to user prompts. This clarification prevents confusing generative AI with other AI applications in education and ensures that our findings are specific to ChatGPT. ChatGPT is a type of generative AI designed as a chatbot that can generate and understand text in a way that closely mirrors human language.

Many educators and others have serious concerns about ChatGPT: eroding academic integrity and ethics (Siegle, 2023), the spread of misinformation (Cooper, 2023), the adoption of misleading information (Kılınç, 2023), and the loss of human values and skills (Rahman & Watanobe, 2023). How far should educators allow generative AI to shape education? What are the implications for teachers and students?

On the other hand, many believe that ChatGPT's potential contributions to education are immense, with opportunities for creating personalized learning experiences (Chan & Hu, 2023), generating lesson ideas (Siegle, 2023), summarizing complex concepts (Tlili et al., 2023), facilitating student–teacher communication (Kılınç, 2023), and supporting language acquisition (Yan, 2023), among many other possibilities. The emergence of large language models (LLMs) might transform the education industry, helping learners focus on higher-order thinking skills and develop more advanced techniques (Bitzenbauer, 2023; Siegle, 2023; Zhu et al., 2023). Much like the introduction of calculators and the early days of the Google search engine, ChatGPT is seen as a double-edge sword—both a threat to human cognitive development and a tool for educational innovation.

Because advanced generative AI technology is so new, educators, researchers, and policymakers have not developed adequate policies and practices to address ChatGPT's use in education. To develop such policies, we need more knowledge about the potential uses, and misuses of the technology. Researchers have moved quickly, with many studies published about generative AI in education in the past two years. Many studies indicate that ChatGPT is shaping pedagogical practices quite broadly. For example, ChatGPT has been rapidly developing its capacity to understand and address educational queries and integrate with existing educational platforms, which enables educators to incorporate ChatGPT into their pedagogical practices (Kılınç, 2023; Tate et al., 2023). It is important to review what has currently been published on this topic, to help clarify researchers', educators', and policymakers' thinking and to provide directions for future work.

This paper provides a comprehensive review of the latest literature published in educational journals. By compiling and analyzing articles on ChatGPT in education we have been able to extract themes and areas of disagreement, with the goal of working toward a productive synthesis. By drawing on sociocultural theory, self-determination theory, and the concepts of accessibility, personalized/differentiated learning, and critical AI literacy, our research questions were designed to examine not only *what* practices and challenges emerged in the use of ChatGPT, but also *why* these patterns matter for teaching and learning.

Sociocultural theory illuminated how ChatGPT supports learning through scaffolding, while self-determination theory explained patterns of learner autonomy and motivation. The principles of accessibility and personalized learning clarified findings on equity and inclusion, and critical AI literacy helped us interpret concerns about plagiarism, bias, and misinformation. Together, these frameworks ensured that our findings were theoretically grounded and offered deeper insights into the role of generative AI in education.

The following questions guided our review:

1. What are the central purposes of empirical research studies focused on ChatGPT in educational contexts?
2. What practices integrating ChatGPT in educational contexts have been identified through empirical research?
3. What challenges has empirical research identified when using ChatGPT in educational contexts?
4. What research methods have been used in empirical studies of ChatGPT in educational contexts?

2 Theoretical and Conceptual Background

2.1 Sociocultural Theory

One of Vygotsky's (1978) central ideas was the "zone of proximal development" meaning the knowledge and skills that students can learn given appropriate supports from cognitive tools and those with more knowledge. Often a person—a parent, teacher, or peer—provides this support, but technology could do this also. Sociocultural theory (Vygotsky, 1978) focuses on tools or technologies which mediate human thought and action, allowing us to accomplish tasks that we would otherwise be unable to complete. From this perspective, ChatGPT can mediate students' learning in educational contexts (Thorne, 2024) by providing students with technology-enhanced learning opportunities (Ormrod, 2011; Thorne, 2024).

2.2 Self-Determination Theory

Students' learning is also impacted by motivation. Self-determination theory (SDT) provides a framework for understanding human motivation (Deci & Ryan, 2012; Ryan & Deci, 2000). Individuals have higher motivation when three basic needs are met: (1) autonomy, (2) competence, and (3) relatedness. Autonomy refers to one's feeling that they are acting willfully, while competence refers to one's belief that they can accomplish their goals (Ryan & Deci, 2000). Relatedness refers to one's need to belong in a social community (Ryan & Deci, 2000).

One subtheory of SDT is cognitive evaluation theory (CET), which focuses on the interconnection of autonomy and competence (Deci & Ryan, 2012). Intrinsic motivation can be thwarted when students receive feedback that undermines their competence or when they receive rewards that seem outside their control. Intrinsic motivation can be enhanced when students receive positive feedback that supports their autonomous goals, making them feel more confident, and/or when they receive rewards that acknowledge their efforts. CET also indicates that some learning environments are autonomy-supportive, and others are autonomy-controlling. In autonomy-controlling environments, students have comparatively lower intrinsic motivation because they perceive less control over their goals.

2.3 Accessibility and Personalized/Differentiated Learning

Learners have unique needs that should be addressed in the classroom. The U.S. Department of Education (2010) underscores the importance of learning environments that are accessible to the needs of diverse learners, including multilingual learners, students with disabilities, and adult learners. Accessibility in education means that all students can utilize the educational resources in a particular learning environment (AEM Center, 2023).

The Universal Design for Learning (UDL) framework is grounded in three standards for designing accessible learning environments (Meyer et al., 2014). The first standard involves increasing student motivation by providing multiple opportunities for student engagement. Pertinent strategies include providing students with opportunities for self-reflection, fostering collaboration, and optimizing the amount of challenge and autonomy in learning. The second standard involves offering multiple opportunities for representation, meaning multiple instructional tools. Relevant strategies include providing digital media, culturally relevant textbooks, and data visualization. The final standard involves providing multiple opportunities for students to demonstrate their learning. Teaching strategies include incorporating performing or visual arts-based assessment and utilizing speech recognition software (Meyer et al., 2014).

In relation to UDL, individualized, differentiated, and personalized learning can increase accessibility for learners with diverse needs (U.S. Department of Education, 2010). These three approaches are incrementally more adaptive in terms of pacing, teaching method, and responsiveness to students' interests. While all three approaches allow modifications in pacing, only differentiated and personalized learning allow changes in methods across students and only personalized learning encourages teachers to tailor instruction to students' interests.

ChatGPT provides students access to complex topics, allowing students to extend their knowledge, and create personalized learning experiences based on students' circumstances (Aktay et al., 2023; Chan & Hu, 2023; Kılınç, 2023). Teachers can also use ChatGPT to create differentiated learning materials, assessments, and feedback in response to students' inquiries, language proficiency levels, and special needs (Barrett & Pack, 2023; Jauhiainen & Guerra, 2023; Nikolic et al., 2023).

2.4 Critical AI Literacy

There are growing ethical and practical concerns about ChatGPT. In terms of ethics, there are concerns that ChatGPT's responses will be biased based on dominant perspectives that neglect marginalized groups (Huallpa et al., 2023) and about academic integrity issues due to increased plagiarism (Chaudhry et al., 2023; Kakhki & Gendron, 2024). Additionally, there are practical concerns that ChatGPT may produce inaccurate and misleading information (Gregorcic & Pendrill, 2023).

These concerns are connected to students' critical thinking and especially to their critical AI literacy. Critical thinking is one of the "twenty-first century skills" identified by political and business leaders as essential in the workplace (National Research Council, 2012). Thomas and Lok (2015) developed an operational framework for critical thinking with three inter-related attributes: (1) disposition (e.g., attitudes, habits of mind, intellectual virtues), (2) knowledge (e.g., general information, experience, specific content), and (3) skills (e.g.,

evaluation, reasoning, and self-regulation). Mills et al. (2023) similarly identified three dimensions of a related concept, critical AI literacy: (1) developing domain-specific critical thinking skills concerning AI-generated information, (2) questioning whether AI is reliable and/or possibly harmful, and (3) promoting critical pedagogies that encourage students to question the relationship between power, inequality, and AI use.

2.5 Current Study

In alignment with our research questions, we use a sociocultural lens to examine how educators have utilized ChatGPT as a tool so far, and its benefits and threats. We also use self-determination theory and conceptual knowledge concerning accessibility, personalized/differentiated learning, and critical AI literacy to further frame our understanding of the benefits and threats of ChatGPT.

We view these theories and conceptual frameworks as complementary and interrelated lenses through which we can derive deeper insights from our findings. Sociocultural theory, for instance, emphasizes that students can learn best when equipped with appropriate human or technology-based support. Self-determination theory, accessibility, personalized/differentiated learning, and critical AI literacy are pertinent to supporting students' learning goals and thus are directly related to sociocultural theory.

3 Methodology

This review adopts a content analysis method (Schreier, 2012). Qualitative content analysis allows us to examine the data to address specific research questions. It involves three phases: (1) examining and identifying relevant studies for analysis, (2) developing a coding framework to analyze the literature, and (3) identifying common themes from the studies included in the review (Schreier, 2012).

3.1 Phase 1. Examining and Identifying Relevant Studies

Guided by the objectives to explore the integration, benefits, challenges, and research scope of the selected articles, we sought all relevant articles published between January 2023 and December 2024, and we focused on the integration of ChatGPT in K-12 and higher education. The review identified literature mainly from two databases: (1) the ERIC database as the main source and (2) Google Scholar as an additional resource. ERIC is an online library of educational research, sponsored by the Institute of Education Sciences (IES) of the U.S. Department of Education. Google Scholar provides a variety of educational research, including journal articles, conference papers, and books. Additionally, we conducted a purposeful search of seven leading journals in the field of technology and education: *Computers and Education, Journal of Educational Computing Research, The Internet and Higher Education, Journal of Computers in Education, Journal of Computing in Higher Education, International Journal of Educational Research,* and *Journal of Educational Research*. Using each journal's search engine, we specifically searched for the term "ChatGPT and education" to identify more articles focusing on ChatGPT and its integration with education. The combination of the two databases and purposeful journal searching ensured a

comprehensive collection of relevant articles. In total, 112 articles were identified in this initial search process.

Following the initial search, 60 articles were excluded, using the following exclusion criteria: (a) duplicated articles (n = 2), (b) articles that do not have full-text availability (n = 2), (c) non-empirical articles (n = 41), (d) articles that were not published in peer-reviewed journals (n = 0), and (e) articles that did not explicitly focus on ChatGPT and education (n = 15). As shown in Fig. 1, 52 empirical articles were included in the final data analysis.

3.2 Phases 2 and 3: Developing a Coding Framework and Identifying Themes

We used qualitative content analysis (Schreier, 2012) to analyze the 52 articles and identify common themes. As shown in Table 4 (Appendix A), the following criteria were outlined for each article: (a) purpose statement; (b) participant sample size, participant demographics, research contexts, and curriculum; (c) methodology, research design, data collection, and data analysis; and (d) findings.

To ensure rigor in our coding process, we adopted an iterative and highly collaborative approach. After developing initial codes grounded in our research questions and theoretical frameworks, we divided the research team into two coding groups. Each group independently coded a subset of the articles, with one team focusing on studies of language and literacy (e.g., English Language Learners, feedback, writing development) and the other team focusing on STEM and broader educational contexts (e.g., personalized learning, accessibility, academic integrity). Guided by the theoretical frameworks and the content of the studies, the two teams identified categories such as *personalized learning experiences, accessible learning experiences, supporting English Language Learners, potential inaccuracies, overreliance on technology,* and *generative AI plagiarism.* The groups then came together for joint meetings to compare coding decisions, resolve discrepancies, and refine the coding framework until consensus was achieved. This structure enhanced inter-coder reliability by bringing diverse perspectives to the analysis and reduced the influence of individual bias. We also engaged in peer debriefing sessions and maintained an audit trail

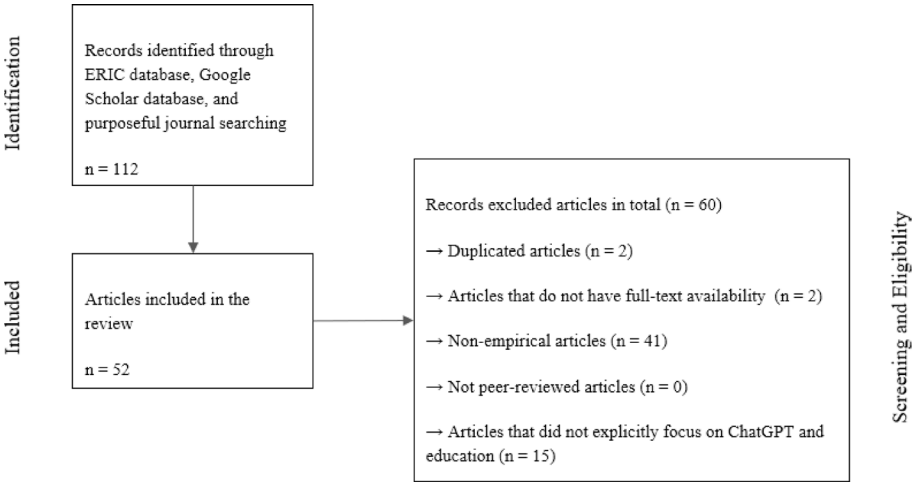

Fig. 1 Data Collection Process

of coding decisions, further strengthening trustworthiness and transparency. Through this rigorous two-team approach, our initial codes evolved into broader, integrative themes that allowed us to systematically interpret shared findings across the literature in a theoretically informed way.

4 Results

Findings include (1) the purposes of the studies, (2) the practices of ChatGPT, (3) the challenges of ChatGPT, and (4) the research methods used in empirical studies of ChatGPT in education.

4.1 The Purpose of the Studies

We addressed the first research question by examining the central purposes of these studies. There are four common research purposes: (a) effectiveness of using ChatGPT (n = 21); (b) perceptions of using ChatGPT (n = 18), (c) concerns of using ChatGPT (n = 7), and (d) other purposes (n = 6) (see Fig. 2).

4.1.1 Effectiveness of Using ChatGPT

Twenty-one studies focused on exploring ChatGPT's effectiveness in supporting students' learning, such as providing feedback and improving English writing and reading (C. Liu et al., 2024a; Çelik et al., 2024; Escalante et al., 2023; Evmenova et al., 2024; M. Liu et al., 2024b; Mizumoto & Eguchi, 2023; Punar & Yangin; 2024; Shabara et al., 2024; X. Wang et al., 2024b; Yang, 2024), building understanding and enhancing learning outcomes in STEM subjects (Cooper, 2023; Kortemeyer, 2023; Nikolic et al., 2023; Urrutia & Araya, 2024),

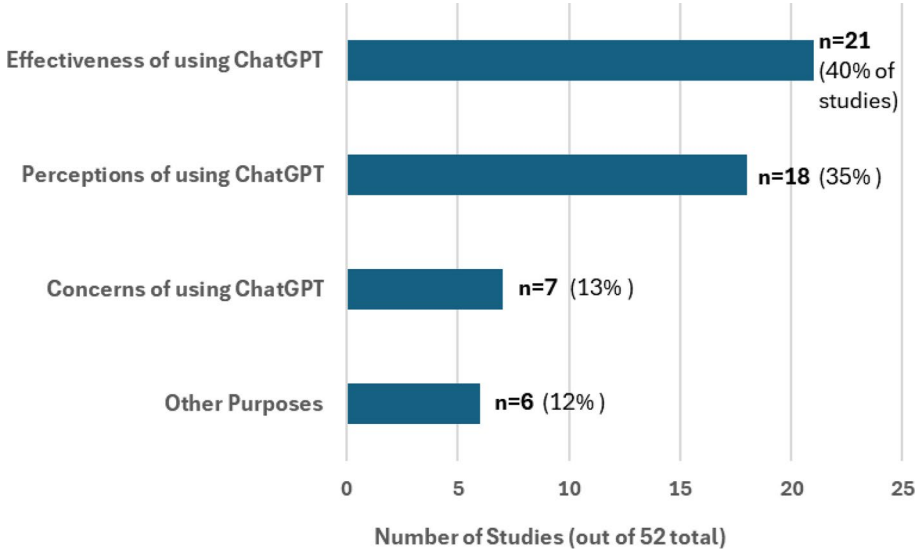

Fig. 2 Purpose of studies (N = 52)

and supporting self-regulated learning (Lee et al., 2024) and special education (Evmenova et al., 2024; Rakap, 2023). Nine studies investigated ChatGPT's performances in assisting teaching practice, such as supporting instructional and lesson design (Kılınç, 2023; Lee & Zhai, 2024) and grading and evaluating assessments (Dahlkemper et al., 2023; Perkins et al., 2024). For example, Escalante et al. (2023) investigated how AI-generated feedback supports students' linguistic progress in writing compared with teachers' feedback.

4.1.2 Perceptions of Using ChatGPT

There are 18 studies that explored students' and teachers' general perceptions and attitudes about integrating ChatGPT into education, specifically focusing on shifts before and after integrating ChatGPT and on identifying any acceptance or resistance. Subtopics include revealing students' general views on ChatGPT as a learning tool (Aktay et al., 2023; Bitzenbauer, 2023; Chan & Hu, 2023; Huallpa, et al., 2023; Jacob et al., 2024; Jauhiainen & Guerra, 2023; Mahapatra, 2024; Ngo, 2023; Romero Rodríguez et al., 2023; Tseng & Lin, 2024; Yan, 2023), teachers' perceptions about ChatGPT as a teaching tool (Barrett & Pack, 2023; Guo & Wang, 2024; Jeon & Lee, 2023; Kakhki et al., 2024; Mena Octavio et al., 2024; Mohamed, 2024), and public views of ChatGPT from social media posts (Mogavi et al., 2024). For instance, Bitzenbauer's (2023) research focused on high school students' initial thoughts on using ChatGPT for physics learning, and how students' attitudes shifted after the AI intervention.

4.1.3 Concerns of Using ChatGPT

Seven studies focused on concerns associated with ChatGPT including AI plagiarism and ChatGPT's exam performance (Alexander et al., 2023; Chaudhry et al., 2023; Cong-Lem et al., 2024; De Winter, 2024; Sullivan et al., 2023; Tlili et al., 2023; Yeadon et al., 2023). For example, Cong-Lem et al. (2024) investigated how EFL teachers perceive AI plagiarism in students' essays and possible strategies to effectively address this concern.

4.1.4 Other Purpose

Six studies evaluated ChatGPT's general implications and future integration in education using a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis or other methods (Gregorcic & Pendrill, 2023; Karakose et al., 2023; Rahman & Watanobe, 2023; Taktak et al., 2024; Tülübaş et al., 2023; Zhu et al., 2023). For instance, Taktak et al. (2024) explored the strengths and weaknesses of integrating ChatGPT in K-12 education using the SWOT analysis framework.

4.2 Enhancing Educational Practice Through ChatGPT

Four main ideas emerged across the literature concerning how ChatGPT can enhance educational practices by (a) enhancing access and understanding, (b) developing curricula, lesson plans, and assessments, (c) fostering personalized, engaging, and efficient learning, and (d) supporting English learners and teachers.

4.2.1 Enhancing Access and Understanding

The most discussed idea in the literature (n = 14) is that ChatGPT effectively summarizes and presents complex topics so that students can more readily access and understand new material (Aktay et al., 2023; Çelik et al., 2024; Chan & Hu, 2023; Cooper, 2023; Karakose et al., 2023; Kılınç, 2023; Kortemeyer, 2023; Ngo, 2023; Nikolic et al., 2023; Rahman & Watanobe, 2023; Stojanov, 2023; Tlili et al., 2023; Tülübaş et al., 2023; Zhu et a., 2023). Aktay et al. (2023), for example, incorporated ChatGPT into a fourth-grade Turkish science class. Subsequent interviews with teachers and students revealed that ChatGPT answered questions and summarized topics quickly and coherently.

4.2.2 Developing Curricula, Lesson Plans, and Assessments

Eleven studies discussed ChatGPT's ability to generate and assist in the development of curricula, lesson plans, and assessments (Barrett & Pack, 2023; Jauhiainen & Guerra, 2023; Jeon & Lee, 2023; Kılınç, 2023; Lee & Zhai, 2024; Mena et al., 2024; Mohamed, 2024; Rahman & Watanobe, 2023; Rakap, 2024; Tlili et al., 2023; Zhu et al., 2023). For example, Jeon and Lee (2023) found that Korean elementary school teachers valued ChatGPT's capacity to serve as a teaching assistant by developing content and assessing students' work. In another study, Lee and Zhai (2024) found that ChatGPT can help pre-service teachers design well-differentiated instruction and formative assessments.

4.2.3 Fostering Personalized, Engaging, and Efficient Learning

Nine studies highlighted how ChatGPT fosters personalized learning experiences by generating customized guidance and feedback for students with diverse learning needs (Chan & Hu, 2023; Jacob et al., 2024; Jauhiainen & Guerra, 2023; Kakhki et al., 2024; Kılınç, 2023; Ngo, 2023; Stojanov, 2023; Sullivan et al., 2023; Taktak et al., 2024). Undergraduate students in Chan and Hu's (2023) study, for example, appreciated ChatGPT's differentiated and immediate learning guidance and feedback. In another study, Jauhiainen and Guerra (2023) found that ChatGPT can provide customized history lesson plans and tests based on fourth through sixth-grade students' individual learning gaps. ChatGPT's ability to foster personalized and adaptive learning experiences may be particularly pertinent for students with autism (Rakap, 2023) and other disabilities (Rahman & Watanobe, 2023).

Building upon personalization, nine studies discussed how ChatGPT's interactive conversational features make learning more engaging (Aktay et al., 2023; C. Liu et al., 2024a; Jauhiainen & Guerra, 2023; Jeon & Lee, 2023; Kakhki et al., 2024; Rahman & Watanobe, 2023; Stojanov, 2023; Taktak et al., 2024; Tlili et al., 2023). C. Liu et al. (2024a), for example, found that reading interest and engagement were higher among fifth-grade Taiwanese students when ChatGPT was incorporated into their reading program.

Six studies also discussed how ChatGPT makes learning more efficient (Chan & Hu, 2023; Kılınç, 2023; Mogavi et al., 2024; Ngo, 2023; Rakap, 2023; Taktak et al., 2024). For example, students in Ngo's (2023) study reported that ChatGPT saved them time on various learning tasks, since it provided real-time interactions and accessible and coherent information. Social media users also emphasized efficiency as a major benefit (Mogavi et al., 2024).

The interplay of personalization, engagement, and efficiency contributes to the promotion of students' learning autonomy. Five studies explored how ChatGPT builds learning autonomy and encourages self-directed learning (Chan & Hu, 2023; Jauhiainen & Guerra, 2023; Kakhki et al., 2024; Kılınç, 2023; Stojanov, 2023). For example, in panel sessions with postsecondary faculty members, administrators, and students in the United States, participants highlighted ChatGPT's ability to create personalized and self-directed learning paths according to students' interests and backgrounds (Kakhki et al., 2024). Stojanov (2023) similarly found that ChatGPT can serve as a helpful tutoring tool, because it allows students to ask questions without fear of judgment, providing a solid foundation for self-directed learning. Additionally, Kılınç (2023) found that ChatGPT promotes students' autonomy by providing real-time, personalized feedback that caters to their academic needs and pace.

4.2.4 Supporting English Learners and Teachers

Notably, 10 studies revealed that ChatGPT supports English teachers and learners by providing targeted guidance and personalized feedback on various English writing tasks (Escalante et al., 2023; Evmenova et al., 2024; Guo & Wang, 2024; Jeon & Lee, 2023; M. Liu et al., 2024b; Mahapatra, 2024; Punar & Yangin, 2024; Shabara et al., 2024; Tseng & Lin, 2024; Yan, 2023). Tseng and Lin (2024), for example, found that ChatGPT provided targeted and informational feedback to EFL students about the organization, coherence, and grammar of their writing. For English teachers, ChatGPT serves as an efficient English teaching assistant, especially in developing learning materials (e.g., discussion topics, vocabulary lists) and activities and auto-grading essays (Jeon & Lee, 2023; Mena et al., 2024; Mizumoto & Eguchi, 2023; Mohamed, 2024; Yang, 2024). For instance, Mizumoto and Eguchi (2023) found that ChatGPT consistently graded students' essays in alignment with the given linguistic criteria, offered timely feedback, and reduced teachers' grading time.

Four studies also compared ChatGPT's feedback on students' writing to a human tutor's feedback. These studies primarily found that ChatGPT needs improved consistency on both language and content evaluation, and students tended to make more progress with human tutors' individualized feedback (Escalante et al., 2023; Evmenova et al., 2024; Shabara et al., 2024). Guo and Wang (2024), however, noted that ChatGPT provides more extensive, directive feedback which is useful in supplementing human tutors' feedback.

4.3 The Challenges of ChatGPT

To address the third research question of this review, we investigated various challenges and concerns created by the incorporation of ChatGPT, including (a) harming academic integrity, (b) producing inaccurate, biased, or made-up information, (c) providing inaccurate feedback, and (d) undermining students' critical thinking skills.

4.3.1 Harming Academic Integrity

Potential AI plagiarism and academic integrity concerns are one major challenge of ChatGPT, as evidenced by 14 studies (Barrett & Pack, 2023; Chan & Hu, 2023; Chaudhry et al., 2023; Cong-Lem et al., 2024; De Winter, 2024; Kakhki & Gendron, 2024; Lee et al., 2024; Nikolic et al., 2023; Rahman & Watanobe, 2023; Sullivan et al., 2023; Tlili et al., 2023; Yan,

2023; Yeadon et al., 2023; Zhu et al., 2023). Students could over-rely on ChatGPT because it can quickly and articulately generate texts on almost any topic, potentially leading to plagiarism. This risk is further compounded by evidence that current detection programs and humans struggle to identify AI-produced texts (Alexander et al., 2023; Cong-Lem et al., 2024; Lee & Zhai, 2024; Sullivan et al., 2023; Yan, 2023; Yeadon et al., 2023). Several studies also questioned the relevance of traditional assessments given this new technology (Chaudhry et al., 2023; Nikolic et al., 2023; Yeadon et al., 2023). Notably, Perkins et al. (2024) found that AI-detection software Turnitin was generally effective, but there was still a discrepancy between the software and human detection.

Not all researchers agree, however, that the risk of plagiarism to traditional assessment is significant. For example, Gregorcic and Pendrill (2023) noted that ChatGPT's responses contained incorrect and contradictory information, reducing students' ability to use ChatGPT for "cheating." Similarly, Kortemeyer (2023) found that ChatGPT can barely pass an introductory physics course, and Dahlkemper et al. (2023) found that students tended to give higher ratings to the accuracy of physics responses provided by humans rather than those from ChatGPT.

4.3.2 Producing Inaccurate, Biased, or Made-Up Information

Another frequent challenge of ChatGPT is its tendency to sometimes produce inaccurate information, as discussed in 13 studies (C. Liu et al., 2024a; Chan & Hu, 2023; Gregorcic & Pendrill, 2023; Kılınç, 2023; Kortemeyer, 2023; Lee & Zhai, 2024; M. Liu et al. 2024b; Rahman & Watanobe, 2023; Stojanov, 2023; Sullivan et al., 2023; Taktak et al., 2024; Tlili et al., 2023; Zhu et al., 2023). In physics education, Gregorcic and Pendrill (2023) describe how the chatbot often produces linguistically coherent but incorrect responses. Furthermore, ChatGPT tends to present biased content, favoring some perspectives and identities over others (Aktay et al., 2023; Cooper, 2023; Rahman & Watanobe, 2023; Tlili et al., 2023; Zhu et al., 2023). Kılınç (2023) also noted that ChatGPT lacks cultural sensitivity.

These inaccuracies are compounded by the difficulty of assessing the quality of ChatGPT's output, since it does not always provide references and supporting evidence for its output (Cooper, 2023; M. Liu et al., 2024b; Ngo, 2023; Tülübaş et al., 2023). One positive development is that ChatGPT-4 and Bing AI (Microsoft AI Chatbot) provide references and resources related to their output, making these tools more reliable (M. Liu et al., 2024b).

4.3.3 Providing Inaccurate Feedback

Six studies revealed ChatGPT's potential for providing inaccurate feedback on students' assessments and teachers' curriculum planning (Escalante et al., 2023; Jacob et al., 2024; Kılınç, 2023; Shabara et al., 2024; Stojanov, 2023; Tseng & Lin, 2024). Researchers thus suggest that ChatGPT be used as a tutoring tool rather than for assessment (Jacob et al., 2024; Stojanov, 2023; Tseng & Lin, 2024). Lee and Zhai (2024) similarly propose that ChatGPT should not be used as a single information source but as a starting point for exploration. Users should also critically examine the source of ChatGPT's information (Lee & Zhai, 2024).

Additionally, there is a debate on ChatGPT's ability to provide effective and concise feedback for English learners. Multiple researchers found that ChatGPT's feedback to EFL

learners tended to be unreliable and inconsistent and did not result in significant linguistic progress compared to those who received feedback from human tutors (e.g., Escalante et al., 2023; Jacob et al., 2024). Guo and Wang (2024), however, found that ChatGPT was efficient in generating a large amount of directive writing feedback pertaining to content, organization and language usage. Punar and Yangin (2024) also found that ChatGPT's timely feedback effectively assists EFL learners in self-editing their writing assignments and potentially developing their writing abilities.

4.3.4 Undermining Students' Critical Thinking Skills

While some researchers have suggested that ChatGPT has the potential to foster critical thinking (Bitzenbauer, 2023; Zhu et al., 2023), others are concerned about ChatGPT's limited critical thinking (Zhu et al., 2023) and how it might undermine students' critical thinking skills (Chan & Hu, 2023; Rahman & Watanobe, 2023; Tlili et al., 2023). For example, Rahman and Watanobe (2023) suggested that if students over-rely on ChatGPT and simply utilize the output without critically examining it, then ChatGPT can be a great barrier for developing students' critical thinking and problem-solving skills.

4.4 Research Methods in ChatGPT Empirical Studies

To address the fourth research question, we examined five aspects of the empirical studies: (a) the population, (b) geographic contexts, (c) the study design, (d) data collection, and (d) the research methods.

4.4.1 Research Population

As shown in Table 1, the empirical studies encompass a broad range of populations, reflecting diverse and global educational settings and participant groups, including (a) university students, (b) university faculty; (c) K-12 students; (d) K-12 experienced or pre-service teachers and (e) public groups. Some studies spanned multiple populations.

4.4.1.1 University Students, Faculty, and Administrators Overall, studies (n = 27) focused on postsecondary populations, including students, faculty members, and/or administrators, were most discussed in the empirical literature. The highest number of these studies, 15, focused on university students only (both undergraduate and postgraduate) across various majors in higher education settings (Çelik et al., 2024; Chan & Hu, 2023; Dahlkemper et al., 2023; Escalante et al., 2023; Huallpa et al., 2023; Jacob et al., 2024; Lee et al., 2024; M. Liu et al., 2024b; Mahapatra, 2024; Ngo, 2023; Punar Özçelik & Yangin Eksi, 2024; Romero Rodríguez et al., 2023; Tseng & Lin, 2024; X. Wang et al., 2024b; Yan, 2023). These studies explored students' attitudes regarding incorporating ChatGPT into education (e.g., Chan & Hu, 2023; Romero Rodríguez et al., 2023) and ChatGPT's benefits and threats (e.g., Ngo, 2023).

While studies focused on postsecondary students were most prominent, we also identified six studies focused on university faculty populations (Alexander et al., 2023; Cong-Lem et al., 2024; Cooper, 2023; Mohamed, 2024; Perkins et al., 2024; Stojanov, 2023). These

Table 1 Research population distribution (N = 52)

Population	Studies	Number of studies	Percent of studies
University students (only)	Çelik et al. (2024); Chan and Hu (2023); Dahlkemper et al. (2023); Escalante et al. (2023); Huallpa et al. (2023); Jacob et al. (2024); Lee et al. (2024); M. Liu et al. (2024b); Mahapatra (2024); Ngo (2023); Punar Özçelik and Yangin Eksi (2024); Romero Rodríguez et al. (2023); Tseng and Lin (2024); X. Wang et al. (2024b); Yan (2023)	15	29%
University faculty (only)	Alexander et al. (2023); Cong-Lem et al. (2024); Cooper (2023); Mohamed (2024); Perkins et al. (2024); Stojanov (2023)	6	12%
University students and faculty (only)	Barrett and Pack (2023); Guo and Wang (2024); Rahman and Watanobe, 2023; Shabara et al. (2024); Yang (2024)	5	10%
University students, faculty, and administrators	Kakhki et al. (2024)	1	2%
K-12 students	Aktay et al. (2023); Bitzenbauer (2023); C. Liu et al. (2024a); Evmenova et al. (2024); Jauhiainen and Guerra (2023); Urrutia and Araya (2024)	6	12%
Pre-service K-12 teachers	Lee and Zhai (2024); Rakap (2024)	2	4%
Experienced K-12 teachers (only)	Jeon and Lee (2023); Mena Octavio et al. (2024)	2	4%
Experienced K-12 teachers and school principals	Taktak et al. (2024)	1	2%
Public groups	Mogavi et al. (2024); Sullivan et al. (2023); Tlili et al. (2023)	3	6%
Not applicable—The researchers are focused on ChatGPT output only.	Chaudhry et al. (2023); De Winter (2024); Gregorcic and Pendrill (2023); Karakose et al. (2023); Kılınç (2023); Kortemeyer (2023); Mizumoto and Eguchi (2023); Nikolic et al. (2023); Tülübaş et al. (2023); Yeadon et al. (2023); Zhu et al. (2023)	11	21%

studies explore faculty members' perspectives towards ChatGPT's strength and weakness (e.g., Cooper, 2023; Mohamed, 2024; Stojanov, 2023), ethical issues on AI-based plagiarism (e.g., Alexander et al., 2023; Cong-Lem et al., 2024), and the effectiveness of using various AI detection tools such as Turnitin (Perkins et al., 2024).

Five studies also integrated data from both university students and faculty (Barrett & Pack, 2023; Guo & Wang, 2024; Rahman & Watanobe, 2023; Shabara et al., 2024; Yang, 2024). Two of these studies used the student and faculty data to provide multiple perspectives on generative AI in education (Barrett & Park, 2023; Rahman & Watanobe, 2023). In the remaining three studies, the faculty members were the primary participants and the students' compositions were used to explore differences in AI-generated and teacher-generated feedback (Guo & Wang, 2024; Shabara et al., 2024; Yang, 2024).

Finally, Kakhki et al. (2024) collected data from university students, faculty members, and administrators to identify affordances and risks of ChatGPT in higher education.

4.4.1.2 K-12 Students, K-12 Pre-Service and Experienced Teachers, and K-12 School Principals In total, 11 studies focused on elementary and secondary school education, includ-

ing students, teachers, and principals and spanning a variety of domains including language, science, social studies, and special education. Among these 11 studies, six studies focused on K-12 students (Aktay et al., 2023; Bitzenbauer, 2023; C. Liu et al., 2024a; Evmenova et al., 2024; Jauhiainen & Guerra, 2023; Urrutia & Araya, 2024). These studies examined students' perceptions about using ChatGPT in the classroom (e.g., Aktay et al., 2023; Bitzenbauer, 2023); whether ChatGPT can create personalized learning experiences and how students perceive these experiences (e.g. Jauhiainen & Guerra, 2023); ChatGPT's writing feedback for students with and without disabilities and/or English language learners (Evmenova et al., 2024); and whether chatbots impact students' reading engagement and interest (C. Liu et al., 2024a). Overall, four studies focused on K-12 teachers, including two focused on pre-service teachers (Lee & Zhai, 2024; Rakap, 2023) and two on experienced teachers (Jeon & Lee, 2023; Mena Octavio et al., 2024). These studies explored ChatGPT's potential to support teachers with lesson planning (Lee & Zhai, 2024), Individualized Education Program (IEP) design (Rakap, 2023), and general teaching assistance (Jeon & Lee, 2023; Mena Octavio et al., 2024).

In addition, Taktak et al. (2024) collected data from teachers and principals in Turkey to explore the strengths, weaknesses, opportunities, and threats (SWOT) associated with ChatGPT in K-12 education.

4.4.1.3 Other Populations Three studies collected data from public groups (Mogavi et al., 2024; Sullivan et al., 2023; Tlili et al., 2023). Specifically, researchers collected social media posts or news articles relevant to the implications of ChatGPT in the general educational context, exploring the general public's or early adopters' perspectives on using ChatGPT for learning.

Additionally, researchers focused on ChatGPT output only in 11 of the 52 studies, and thus, no participant population was specified. See Table 1 for details.

4.4.2 Geographic Context

Table 2 displays the geographic context of each study. Notably, the geographic context was not clearly specified in 13 out of 52 studies.

Table 2 indicates that the research studies were conducted in a diverse range of geographic contexts with the highest numbers coming from China (n=4; Guo & Wang, 2024; X. Wang et al., 2024b; Yan, 2023; Yang, 2024) and Turkey (n=4; Aktay et al., 2023; Celik et al., 2024; Punar & Yangin, 2024; Taktak et al., 2024), followed by Taiwan (n=3; C. Liu et al., 2024a; Lee et al., 2024; Tseng & Lin, 2024) and the United States (n=3; Barrett & Pack, 2023; Jacob et al., 2024; Kakhki et al., 2024). Additionally, three studies that incorporated data from social media and news articles were designated as "global" with respect to their geographic contexts, because their research spanned multiple geographic locations (Mogavi et al., 2024; Sullivan et al., 2023; Tlili et al., 2023). For example, Sullivan et al. (2023) utilized news articles from New Zealand, Australia, the United Kingdom, and the United States.

Table 2 Geographic context of the studies (N = 52)

Geographic context	Studies	Number of studies	Percent of studies
China	Guo and Wang (2024); X. Wang et al., (2024b); Yan (2023); Yang (2024)	4	8%
Turkey	Aktay et al. (2023); Celik et al. (2024); Punar and Yangin (2024); Taktak et al. (2024)	4	8%
Taiwan	C. Liu et al. (2024a); Lee et al. (2024); Tseng and Lin (2024)	3	6%
United States	Barrett and Pack (2023); Jacob et al. (2024); Kakhki et al. (2024)	3	6%
Germany	Bitzenbauer (2023); Dahlkemper et al. (2023)	2	4%
Spain	Mena et al. (2024); Romero Rodríguez et al. (2023)	2	4%
South Korea	Jeon and Lee (2023); Lee and Zhai (2024)	2	4%
Vietnam	Cong-Lem et al. (2024); Ngo (2023)	2	4%
Asian-Pacific region	Escalante et al. (2023)	1	2%
Australia	Nikolic et al. (2023)	1	2%
Cyprus	Alexander et al. (2023)	1	2%
Egypt	Shabara et al. (2024)	1	2%
Hong Kong	Chan and Hu (2023)	1	2%
India	Mahapatra (2024)	1	2%
Latin America	Huallpa et al. (2023)	1	2%
Netherlands	De Winter (2024)	1	2%
New Zealand	M. Liu et al. (2024b)	1	2%
Saudi Arabia	Mohamed (2024)	1	2%
Southeast Asia	Perkins et al. (2024)	1	2%
United Arab Emirates	Chaudhry et al. (2023)	1	2%
United Kingdom	Yeadon et al. (2023)	1	2%
Uruguay	Jauhiainen and Guerra (2023)	1	2%
Global	Mogavi et al. (2024); Sullivan et al. (2023); Tlili et al. (2023)	3	6%
Not clearly specified	Cooper (2023); Evmenova et al. (2024); Gregorcic and Pendrill (2023); Karakose et al. (2023); Kılınç (2023); Kortemeyer (2023); Mizumoto and Eguchi (2023); Rahman and Watanobe (2023); Rakap (2024); Stojanov (2023); Tülübaş et al. (2023); Urrutia and Araya (2024); Zhu et al. (2023)	13	25%

There were observed differences in the geographic context for postsecondary and K-12 studies. For instance, all studies focused on China (n = 4) and the United States (n = 3) were represented among the 27 postsecondary education studies. Of the remaining 20 postsecondary education studies, two studies each collected data from Taiwan (Lee et al., 2024; Tseng & Lin, 2024), Turkey (Celik et al., 2024; Punar Özçelik & Yangin Eksi, 2024), and Vietnam (Cong-Lem et al., 2024; Ngo, 2023). One study each was conducted in the following geographic locations: Asian-Pacific region (Escalante et al. et al., 2023), Cyprus (Alexander et al., 2023), Egypt (Shabara et al., 2024), Germany (Dahlkemper et al., 2023), Hong Kong (Chan & Hu, 2023), India (Mahapatra, 2024), Latin America (Huallpa et al., 2023), New Zealand (M. Liu et al., 2024b), Saudi Arabia (Mohamed, 2024), Southeast Asia (Perkins et

al., 2024), and Spain (Romero Rodríguez et al., 2023). Three postsecondary studies had an unclear geographic context (Cooper, 2023; Rahman & Watanobe, 2023; Stojanov, 2023).

In terms of the studies conducted in K-12 educational settings, two studies each focused on South Korea and Turkey. Both Korean studies focused on either pre-service (Lee & Zhai, 2024) or experienced (Jeon & Lee, 2023) teachers, while one Turkey study focused on students (Aktay et al., 2023) and the other focused on experienced teachers and school principals (Taktak et al., 2024). Of the seven remaining K-12 education studies, one study each focused on K-12 students in Germany (Bitzenbauer, 2023), Taiwan (C. Liu et al., 2024a), and Uruguay (Jauhiainen & Guerra, 2023), while one focused on experienced K-12 teachers in Spain (Mena Octavio et al., 2024). The three remaining K-12 studies had an unclear geographic location (Evmenova et al., 2024; Rakap, 2024; Urrutia & Araya, 2024).

4.4.3 The Study Design

Researchers employed diverse study designs to explore the intersection of ChatGPT and education. As displayed in Fig. 3, in nearly half of the studies (n = 25 or 48%), researchers focused on human interactions with ChatGPT to explore its usage patterns in education. Researchers also designed AI-enhanced learning (n = 16) and teaching (n = 11) interventions. Although these intervention studies rarely utilized a "true" experimental design with random assignment, several adopted some form of treatment and control groups (e.g., C. Liu et al., 2024a; Lee et al., 2024; Mahapatra, 2024; Rakap, 2024; X. Wang et al., 2024b). These intervention studies purposefully integrated ChatGPT into learning or teaching procedures and then evaluated ChatGPT's effectiveness by comparing some learning outcome before and after the intervention.

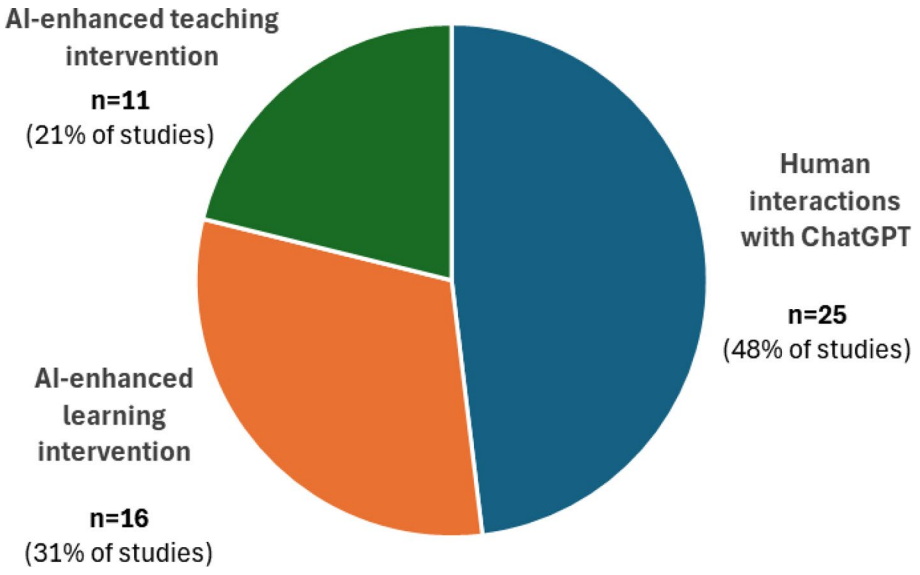

Fig. 3 The study design

4.4.3.1 Human Interactions with ChatGPT In studies examining human interactions with ChatGPT, many researchers integrated ChatGPT into various educational settings (e.g., student interactions with ChatGPT in science education) to gain a better understanding of teachers' and students' perceptions of its implications, without implementing structured interventions (Aktay et al., 2023; Barrett & Pack, 2023; Chan & Hu, 2023; Cong-Lem et al., 2024; Guo & Wang, 2024; Huallpa et al., 2023; Kakhki et al., 2024; Mogavi et al., 2024; Ngo, 2023; Romero Rodríguez et al., 2023; Stojanov, 2023; Sullivan et al., 2023; Tlili et al., 2023). Other researchers adopted a more evaluative approach by critically assessing ChatGPT's performance on specific tasks such as automated grading or answering exam questions (Alexander et al., 2023; Chaudhry et al., 2023; Cooper, 2023; De Winter, 2024; Gregorcic & Pendrill, 2023; Kortemeyer, 2023; Perkins et al., 2024; Taktak et al., 2024; Tülübaş et al., 2023; Urrutia & Araya, 2024; Yeadon et al., 2023; Zhu et al., 2023).

4.4.3.2 AI-Enhanced Learning Interventions The AI-enhanced learning interventions include guiding students to interact with ChatGPT, such as by asking questions and then using ChatGPT's responses to better comprehend learning materials (Bitzenbauer, 2023; Dahlkemper et al., 2023; Jacob et al., 2024; Lee et al., 2024; Mahapatra, 2024). Another common intervention involved using ChatGPT to provide feedback for students' writing tasks (Escalante et al., 2023; Evmenova et al., 2024; Guo & Wang, 2024; M. Liu et al., 2024b; Punar & Yangın, 2024; Tseng & Lin, 2024; Yan, 2023), to assist students' reading tasks (Çelik et al., 2024; C. Liu et al., 2024a ; X. Wang et al., 2024b), and to help students with college-level assessments (Nikolic et al., 2023).

4.4.3.3 AI-Enhanced Teaching Interventions The AI-enhanced teaching interventions include using ChatGPT to design personalized learning goals for students with disabilities (Rakap, 2024), to create personalized learning experience (Jauhiainen & Guerra, 2023; Mohamed, 2024; Kılınç, 2023; Rahman & Watanobe, 2023), and to assist teachers in their instructional design (Jeon & Lee, 2023; Lee & Zhai, 2024; Mena et al., 2024) and the grading process (Mizumoto & Eguchi, 2023; Shabara et al., 2024; Yang, 2024).

4.4.4 Data Collection Methods

Figure 4 displays the data collection methods used across all studies. Since these categories are not mutually exclusive, some studies are represented multiple times. Of the 52 studies, 26 reported one data source; 14 reported two data sources; 10 reported three data sources; one reported four data sources; and one study did not clearly specify any data collection techniques. The most frequently collected data source was ChatGPT's responses (n = 24), followed by surveys/questionnaires (n = 17), students' assignments or test results (n = 14), and individual (usually semi-structured) interviews (n = 13). Other data sources included teachers' qualitative feedback on student work (n = 5), group interviews (e.g., panel session, focus group) (n = 4), social media posts or news articles (n = 3), educational artifacts (e.g., lesson plan) (n = 3), observations (n = 3), expert feedback as part of an autoethnography (n = 1), and think aloud interviews (n = 1).

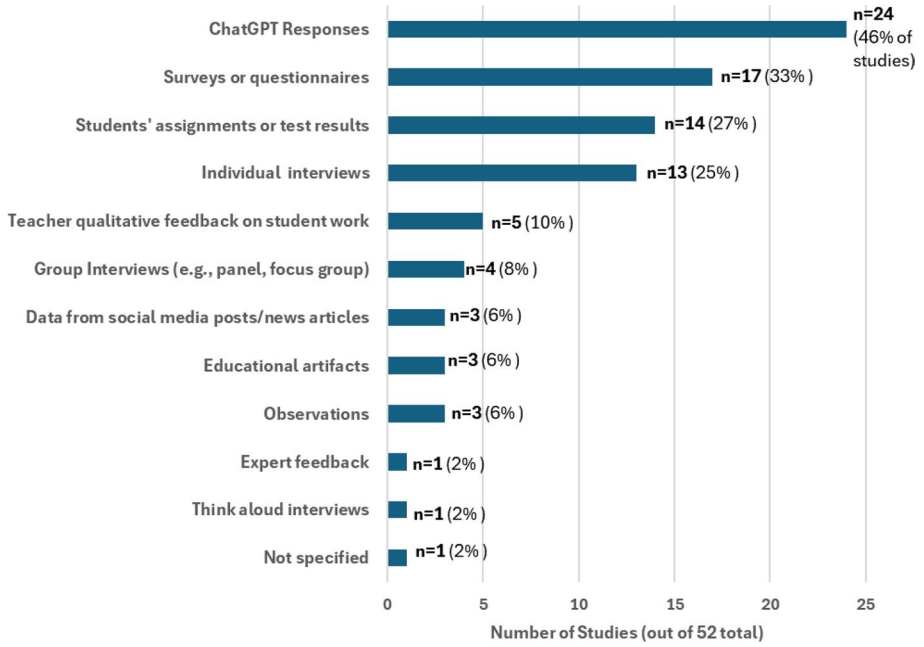

Fig. 4 Data collection methods. The data collection categories were not mutually exclusive and thus the percentages do not sum to 100%

4.4.5 The Research Method

As shown in Fig. 5, the reviewed empirical studies employed a variety of research approaches, mainly including (a) qualitative methods (n = 21), (b) quantitative methods (n = 14), and (c) mixed methods (n = 15). Two studies were also classified as miscellaneous (Nikolic et al., 2023; Zhu et al., 2023), because they used a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis as their primary analytical technique.

4.4.5.1 Qualitative Methods The most frequently used method was qualitative research methods which was used in 21 studies (Aktay et al., 2023; Alexander et al., 2023; Cong-Lem et al., 2024; Cooper, 2023; Evmenova et al., 2024; Gregorcic & Pendrill, 2023; Jacob et al., 2024; Jeon & Lee, 2023; Kakhki et al., 2024; Kılınç, 2023; M. Liu et al., 2024b; Mena Octavio et al., 2024; Mogavi et al., 2024; Mohamed, 2024; Punar Özçelik & Yangin Eksi, 2024; Stojanov, 2023; Sullivan et al., 2023; Taktak et al., 2024; Tlili et al., 2023; Tseng & Lin, 2024; Yan, 2023). The specific research design (e.g., case study, ethnography) was explicitly mentioned in only nine of the 21 published qualitative research manuscripts. Five of the studies utilized case study methods (Gregorcic & Pendrill, 2023; Jacob et al., 2024; Mena Octavio et al., 2024; Punar & Yangin, 2024; Tlili et al., 2023). Two studies positioned the researcher as the focal participant, through autoethnography (Stojanov, 2023) and a "self-study" approach (Cooper, 2023). Additionally, Yan (2023) utilized a multi-method

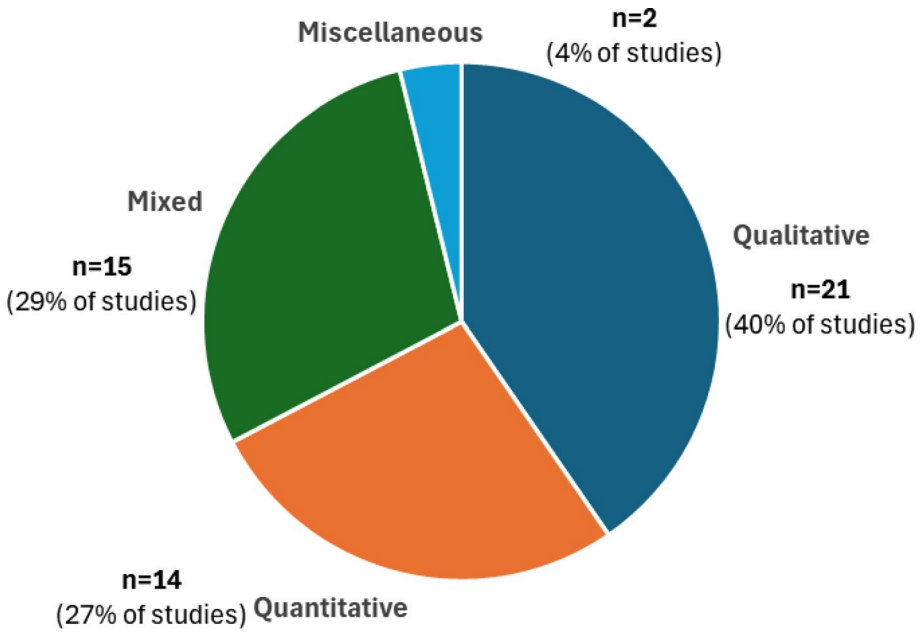

Fig. 5 Types of research methods (N = 52)

qualitative approach and Evmenova et al. (2024) described their research design as secondary data analysis.

Next, the data analysis technique was mentioned in 20 of the 21 published qualitative research manuscripts. Commonly utilized techniques were thematic analysis (e.g., Cong-Lem et al., 2024; Evmenova et al., 2024; Mogavi et al., 2024) and content analysis (e.g., Aktay et al., 2023; Kılınç, 2023; Mohamed, 2024). Other techniques mentioned were document analysis (e.g., Yan, 2023), sentiment analysis (e.g., Sullivan et al., 2023), grounded theory (e.g., Kakhki et al., 2024), and SWOT analysis (e.g., Taktak et al., 2024). Some researchers combined multiple data analysis techniques, such as Yan (2023) who used both thematic and document analysis and Sullivan et al. (2023) who used thematic and sentiment analysis. Additionally, some researchers conducted an exploratory analysis of the qualitative data (e.g., Cooper, 2023; Tseng & Lin, 2024), in which a formal analytical approach such as grounded theory or thematic analysis was not specified.

These qualitative studies used various data sources including individual interviews (e.g., Jacob et al., 2024), group interviews (e.g., Kakhki et al., 2024), ChatGPT responses (e.g., Cooper, 2023), surveys (e.g., Cong-Lem et al., 2024), student assignments (e.g., Tseng & Lin, 2024), social media and news articles (e.g., Sullivan et al., 2023), educational artifacts (e.g., Yan, 2023), and observations (e.g., Stojanov, 2023). For example, Aktay (2023) examined students' perceptions of integrating ChatGPT in their learning via content analysis of interview transcripts, and then generated common themes such as ChatGPT's benefits and threats and students' recommendations on its appropriate usage.

4.4.5.2 Quantitative Methods Quantitative methods were utilized in 14 studies (Barrett & Pack, 2023; Bitzenbauer, 2023; Çelik et al., 2024; Chaudhry et al., 2023; Dahlkemper et al., 2023; De Winter, 2024; Huallpa et al., 2023; Lee et al., 2024; Mizumoto & Eguchi, 2023; Rakap, 2024; Romero Rodríguez et al., 2023; Shabara et al., 2024; Urrutia & Araya, 2024; Yang, 2024). The specific research design (e.g., experimental) was explicitly mentioned in eight of the 14 published quantitative research manuscripts. Three studies used experimental designs (Celik et al., 2024; Lee et al., 2024; Rakap, 2024) with one described as a Randomized Controlled Trial (RCT) (Lee et al., 2024). Additionally, three studies used cross-sectional designs (Barrett & Pack, 2023; Huallpa et al., 2023; Romero Rodríguez et al., 2023) and one used a correlational non-experimental design (Shabara et al., 2024). One study was described as a case study that utilizes a quasi-experimental design (Chaudhry et al., 2023).

These quantitative studies commonly used data from surveys/questionnaires (e.g., Barrett & Pack, 2023) and test results (e.g., Urrutia & Araya, 2024), but some also used ChatGPT output as a data source (e.g., De Winter, 2024). Researchers employed various forms of quantitative analysis to address their research questions, including descriptive statistics (e.g., Bitzenbauer, 2023), inferential statistics (e.g., De Winter, 2024), and psychometric techniques (e.g., Rakap, 2024), with several studies incorporating all three approaches (e.g., Romero Rodriguez et al., 2023). The inferential statistical techniques included bivariate and multivariate statistics. The bivariate analyses included chi-square tests (e.g., Rakap, 2024), independent samples t-tests (Escalante et al., 2023), Mann–Whitney U tests (e.g., Barrett & Pack, 2023), paired samples t-tests (e.g., Shabara et al., 2024), Pearson correlations (e.g., Rakap, 2024), Wilcoxon Signed-Rank Tests (e.g., Celik et al., 2024), and ANOVA (e.g., Dahlkemper et al., 2023). The multivariate techniques include ANCOVA (e.g., Lee et al., 2024), regression analysis (e.g., Huallpa et al., 2023), and structural equation modeling (e.g., Romero Rodriguez et al., 2023). The psychometric analyses included those focused on inter-rater (e.g., Rakap, 2024) and intra-rater (e.g., Yang, 2024) reliability with respect to scoring. For example, Yang (2024) focused on differences in writing composition scores generated between ChatGPT and university writing instructors, and thus, examined both the intra-rater reliability of ChatGPT scores and the inter-rater reliability between ChatGPT and human-generated scores.

4.4.5.3 Mixed Methods Fifteen studies utilized mixed methods by combining qualitative and quantitative approaches (C. Liu et al., 2024a; Chan & Hu, 2023; Escalante et al., 2023; Guo & Wang, 2024; Jauhiainen & Guerra, 2023; Karakose et al., 2023; Kortemeyer, 2023; Lee & Zhai, 2024; Mahapatra, 2024; Ngo, 2023; Perkins et al., 2024; Rahman & Watanobe, 2023; Tülübaş et al., 2023; X. Wang et al., 2024b; Yeadon et al., 2023). Six of these studies specified their research design, with two described as case studies (Jauhiainen & Guerra, 2023; Kortemeyer, 2023) and four as quasi-experimental studies (C. Liu et al., 2024a; Escalante et al., 2023; Mahapatra, 2024; X. Wang et al., 2024b). These mixed methods studies combined data from ChatGPT output (e.g., Yeadon et al., 2023), surveys/questionnaires (e.g., Guo & Wang, 2024), student assignments or test results (e.g., Perkins et al., 2024), teacher feedback on assignments (e.g., Escalante et al., 2023), individual semi-structured interviews (e.g., Ngo, 2023), group interviews (e.g., Tülübaş et al., 2023), and

educational artifacts (e.g., Lee & Zhai, 2024). The researchers used a variety of data analysis techniques. Five studies utilized descriptive statistics (e.g., frequencies, means) to analyze the quantitative data and then analyzed the qualitative data with either content analysis (Kortemeyer, 2023; Rahman & Watanobe, 2023; Yeadon et al., 2023), thematic analysis (Chan & Hu, 2023), or grounded theory (Lee & Zhai, 2024). Next, five studies combined both descriptive and inferential (e.g., t-test) statistics with thematic analysis (Escalante et al., 2023; Guo & Wang, 2024; Ngo, 2023) or a qualitative exploratory approach (Jauhiainen & Guerra, 2023; Perkins et al., 2024). Furthermore, two studies combined content analysis with both descriptive statistics and psychometric techniques (Karakose et al., 2023; Tülübaş et al., 2023), while one combined a phronetic iterative qualitative approach and inferential statistics (Mahapatra, 2024). C. Liu et al. (2024a) and X. Wang et al. (2024b) also combined a qualitative exploratory approach with inferential statistics and psychometric techniques, with the latter study also incorporating descriptive statistics.

4.4.5.4 Miscellaneous Method Two studies were classified as adopting a "miscellaneous" method (Nikolic et al., 2023; Zhu et al., 2023), because their research did not clearly fall under quantitative, qualitative, or mixed methods. Both studies used Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis. Nikolic et al. (2023) explored how ChatGPT might support engineering education, particularly concerning assessment, while Zhu et al. (2023) explored ChatGPT's use in teaching and learning. While Nikolic et al. (2023) described utilizing ChatGPT output to inform their analysis, it was unclear whether Zhu et al. (2023) collected any data or relied solely on previously published studies.

Across the findings, the themes connect directly to our guiding frameworks. For example, evidence of ChatGPT's role in scaffolding and tutoring aligns with sociocultural theory, while patterns of increased autonomy, motivation, and competence reflect self-determination theory. Similarly, the development of differentiated lesson plans and accessible materials corresponds with principles of accessibility and personalized learning. Finally, concerns about plagiarism, misinformation, and bias underscore the importance of fostering critical AI literacy. These connections show how the frameworks shaped not only our research questions but also our interpretation of the findings.

5 Discussion

Through our analysis of 52 selected empirical studies, this review contributes to the ongoing conversation about the role of ChatGPT in education. These studies represent diverse global perspectives from both K-12 and postsecondary educational institutions. Importantly, these studies provide an overview of initial reactions to generative AI technology, underscoring the rapidly developing significance of ChatGPT and its impact on education, including related opportunities and concerns.

As we highlight the benefits of using ChatGPT in education such as creating accessible, personalized learning experiences and serving as a teaching or learning assistant, we also acknowledge concerns and threats to educational practice in terms of accuracy and AI plagiarism. If we intend to implement generative AI effectively, we need to explore effec-

tive techniques to maximize benefits and to mitigate concerns. Fortunately, multiple studies provide valuable insights regarding how these concerns might be managed, allowing ChatGPT's benefits to significantly enhance students' learning outcomes.

In this section, we summarize four key insights on the integration of generative AI in education: (a) promoting personalized learning and learner autonomy; (b) promoting equitable learning and accessibility; (c) fostering pedagogical innovation through creative curriculum, lesson planning, and assessment design; and (d) promoting critical AI literacy and academic integrity.

Table 3 presents a synthesis of our main findings in relation to the three guiding theoretical frameworks, illustrating how sociocultural, motivational, and accessibility/critical literacy perspectives collectively explain the educational implications of ChatGPT.

5.1 Promoting Personalized Learning and Learner Autonomy

Students' sense of personal autonomy as learners contributes to their academic motivation (Deci & Ryan, 2012). When learners feel autonomous, they regard themselves as acting willfully rather than as being controlled by others (Ryan & Deci, 2000). To feel autonomous, students' academic tasks must also align with their competency levels, so they do

Table 3 Integration of main findings with theoretical frameworks

Main finding/theme	Sociocultural theory	Self-determination theory	Accessibility/personalized learning and critical AI literacy
1. Promoting personalized learning and learner autonomy	ChatGPT acts as a *mediating tool* that scaffolds students within their Zone of Proximal Development, supporting independent exploration and collaboration	Enhances *autonomy* and *competence* through self-paced, feedback-driven learning that increases intrinsic motivation	Supports *differentiated* and *personalized* instruction for diverse learners, fostering inclusion and equitable access
2. Promoting equitable learning and accessibility	Functions as a *cultural and cognitive tool* that bridges learning barriers and supports collaborative meaning-making	Fulfills *relatedness* by connecting learners with peers, teachers, and broader learning communities	Aligns with *Universal Design for Learning (UDL)* principles to remove barriers for multilingual and disabled learners while advancing *AI literacy* for equity
3. Fostering pedagogical innovation through curriculum, lesson, and assessment design	Repositions teachers as *co-constructors of knowledge*, using AI as a collaborative partner to design and refine learning tasks	Strengthens *competence* for both teachers and students by automating routine tasks, allowing focus on creative, higher-order goals	Promotes *adaptive design* and *multimodal learning materials* responsive to diverse learner needs and contexts
4. Promoting critical AI literacy and academic integrity	Encourages *socially mediated reflection* and dialogue about technology use in learning communities	Cultivates *autonomous ethical judgment* and responsibility for AI-assisted learning and assessment	Advances *critical AI literacy* by teaching students to question bias, reliability, and authorship in AI-generated content

This table summarizes how the review's main findings align with the three guiding theoretical frameworks—sociocultural theory, self-determination theory, and accessibility/personalized learning with critical AI literacy—showing how each framework contributes to understanding the educational implications of ChatGPT.

not feel overwhelmed (Deci & Dyan, 2012). Furthermore, from a sociocultural perspective (Vygotsky, 1978), generative AI tools such as ChatGPT can be used as personalized learning tools to support students' learning (Ormrod, 2011; Thorne, 2024), allowing students to accomplish tasks within their "zone of proximal development."

The reviewed studies highlight how generative AI has fostered a significant pedagogical shift towards personalized learning (e.g., Jauhiainen & Guerra, 2023; Kakhki et al., 2024; Kılınç, 2023; X. Wang et al., 2024b). Personalized learning allows modifications in pacing and methods across students and encourages teachers to tailor instruction to students' interests (Borja et al., 2015; U.S. Department of Education, 2010). These personalized experiences could help students in advancing their knowledge on topics that they have mastered and further support students in understanding topics that they find challenging. In one study, Jauhiainen and Guerra (2023) found that generative AI could successfully create personalized, enjoyable social studies and history lessons for fourth through sixth grade students and that students' self-reported learning was correlated with their interest in the lesson.

Furthermore, university students have reported that ChatGPT-generated personalized feedback has increased their autonomy (Mahapatra, 2024). AI-enhanced personalized learning positions students as active learners, as they take more control of their learning (Jeon & Lee, 2023). For instance, students can proactively use ChatGPT to develop their academic writing by using the tool to brainstorm topics; to organize their ideas into words; and to improve their grammar, structure, and coherence (C. Wang et al., 2024a).

5.2 Promoting Equitable Learning and Accessibility

An important distinction has been made in recent years between equality and equity in learning. Equality, on the one hand, focuses on providing equal learning resources to every student regardless of their individual needs and circumstances, whereas equity focuses on personalizing the learning experience for each student (Minow, 2021). Accordingly, an equitable learning environment might be one where every student moves at their own pace with lessons tailored to their interests.

The Universal Design for Learning (UDL) framework underscores that courses can be designed in ways that promote or thwart equitable and accessible education (Meyer et al., 2014). The key UDL design principles are (1) providing multiple opportunities for student engagement, (2) utilizing multiple instructional tools to engage students, and (3) providing multiple opportunities for students to demonstrate their learning (Meyer et al., 2014). UDL thus promotes personalized learning, which is a strength of generative AI.

Given these principles, generative AI could be integrated as an alternative instructional tool to engage students. Courses and lessons could be designed to integrate generative AI to promote accessible and equitable learning, given the responsiveness of tools like ChatGPT to students' needs (e.g., Evmenova et al., 2024; Kılınç, 2023; X. Wang et al., 2024b) and by offering increased access to learning resources (e.g., Cooper, 2023; Karakose et al., 2023; Kılınç, 2023). More specifically, by alleviating barriers, such as those related to language proficiency or disabilities, ChatGPT can support teachers in expanding opportunities for customized learning supports (e.g., through translation or simplification of academic con-

tent) (e.g., Escalante et al., 2023; Evmenova et al., 2024; Jeon & Lee, 2023) and in reducing any observed achievement gaps between marginalized groups and their peers. For example, ChatGPT can break down complex concepts (e.g., recycling process in science) (Aktay et al., 2023) and simplify texts (Celik et al., 2024).

In another study, M. Liu et al. (2024b) demonstrated how generative AI could be used to support multimodal writing. Multimodal writing is consistent with the third principal of UDL—i.e., providing multiple avenues for students to demonstrate their learning (Meyer et al., 2014). Chinese EFL international undergraduates who wrote a multimodal essay, incorporating generative AI for both written text and image generation, tended to provide more examples to support their arguments compared to students who used generative AI to write a traditional essay (M. Liu et al., 2024b).

However, although ChatGPT promotes accessibility and equitable learning in many respects, it is crucial that that it does not unintentionally exacerbate existing inequities. For example, users must pay for the unlimited-prompt version of the tool, in addition to paying for internet connectivity and an electronic device. These economic barriers may limit some students' access to the tool, further exacerbating equity concerns. One solution might be increased technology and AI-centered funding for under-resourced schools.

5.3 Fostering Pedagogical Innovation Through Creative Curriculum, Lesson Planning, and Assessment Design

Pedagogical innovation is the development of creative or new ideas that enhance teaching and learning and/or foster greater student engagement and better learning outcomes (Rachmad, 2022). Generative AI tools can support transformative shifts toward pedagogical innovation (1) through the development of learning resources and (2) by facilitating a redistribution of teachers' workloads toward creative and higher-order pedagogical tasks.

5.3.1 Development of Learning Resources

Teachers can use generative AI tools such as ChatGPT to design formative assessments (Lee & Zhai, 2024), to create customized learning materials (Jeon & Lee, 2023), to develop interactive activities (Jeon & Lee, 2023), and to automatically grade students' writing and provide targeted, actionable insights (Mizumoto & Eguchi, 2023). Furthermore, there is considerable potential for teachers to utilize ChatGPT to design innovative multimodal activities. ChatGPT's multimodal functions, like DALL-E for image creation and SORA for video generation, allow users to convert text to image, video, or audio (OpenAI, 2025). These tools support teachers in creating authentic learning scenarios, engaging students in innovative learning with real-world contexts (He et al., 2024).

5.3.2 Fostering a Redistribution of Teachers' Workload Toward Creative and Higher-Order Pedagogical Tasks

ChatGPT can foster a redistribution of teachers' workload by completing routine tasks, freeing time for designing creative activities and higher order pedagogical tasks. For example, teachers could use ChatGPT to design lesson plans for routine topics, such as weekly vocab-

ulary lists (Mena Octavio et al., 2024), and then, allocate more time for creating fun group activities or designing out-of-classroom experiences.

In addition to supporting the development of the learning resources described above, ChatGPT can support teachers in providing personalized feedback to their students (e.g., Evemenova et al., 2024; Guo & Wang, 2024; Nikolic et al., 2023; Punar Özçelik & Yangin Eksi, 2024). First, it is typically not possible for teachers to offer instant feedback to every student, but ChatGPT can provide real-time feedback (Nikolic et al., 2023). Second, Guo and Wang (2024) found that ChatGPT provided more extensive and more directive feedback to EFL students than teacher-generated feedback.

There are some limitations, however, of AI-generated feedback compared to human feedback. Guo and Wang (2024), for example, found that teachers offered more informative feedback and incorporated clarifying questions, which may further support students' understanding. Other limitations of AI-generated feedback compared to human-generated feedback include ChatGPT's relative lack of empathy (Kılınç, 2023) and cultural insensitivity (Mohamed, 2024).

Given professional development opportunities and other supports, teachers could learn to balance AI-generated feedback with their own feedback, saving time in their workloads when possible while not sacrificing the human connection that students may value from their teachers' feedback. Indeed, Escalante et al. (2023) found that university students were evenly split in their preferences for AI-generated versus human tutors. According to self-determination theory (Ryan & Deci, 2000), one's sense of relatedness or feelings of connection to others is also a key component of students' motivation, so it is essential that the human connection is not missing when incorporating generative AI into the classroom. Thus, teachers could determine which assignments are best suited for feedback from ChatGPT and which assignments would most benefit from human feedback, such as students' essays on their cultural backgrounds. This balanced approach allows teachers to direct their energy to areas where it might be most appreciated by students.

5.4 Promoting Critical AI Literacy and Academic Integrity

While there are many positive developments concerning ChatGPT, this section describes several concerns regarding ChatGPT in education as well as potential pathways to address these issues.

5.4.1 Accuracy, Bias, and Use of AI

There are concerns about the quality of information provided by ChatGPT. First, research has shown that ChatGPT may produce inaccurate, misleading, and/or contradictory information (e.g., C. Liu et al., 2024a; Gregorcic & Pendrill, 2023; Kortemeyer, 2023; Lee & Zhai, 2024; Punar Özçelik & Yangin Eksi, 2024; Stojanov, 2023; Zhu et al., 2023). Second, ChatGPT's responses may be biased, based on training data used for the large language model (e.g., Huallpa et al., 2023; Taktak et al., 2024). Consequently, ChatGPT may summarize a particular topic based on dominant perspectives, neglecting voices from marginalized groups.

In addition to concerns about quality, educators are concerned that students may overrely on generative AI in learning, resulting in plagiarism and major threats to academic

integrity (e.g., Chaudhry et al., 2023; Cong-Lem et al., 2024; Kakhki & Gendron, 2024) and harming students' overall academic motivation (Mahapatra, 2024). Students might, for example, misuse ChatGPT to plagiarize an entire essay assignment. While AI detection tools, such as Grammarly, Quillbot, ZeroGPT, Scribbr, and Turnitin, may provide some assistance in identifying inappropriate use of AI content in students' work, plagiarism detection tools are not always able to reliably identify ChatGPT-generated text (Chaudhry et al., 2023). Perkins et al. (2024) found that Turnitin was generally able to detect AI-generated content but showed limitations in detecting paraphrased content.

5.4.2 Addressing Concerns About AI's Accuracy, Bias, and Use

The integration of generative AI into education provides an opportunity for educators to develop students' "twenty-first century skills" (National Research Council, 2012). Specifically, developing students' critical thinking (Thomas & Lok, 2015) and critical AI literacy (Mills et al., 2023) are integral to ethical use of generative AI in the classroom. This ensures that the integration of generative AI in education can be navigated appropriately to maximize its benefits. Developing critical AI literacy involves cultivating students' critical thinking skills—i.e., their knowledge, skills, and dispositions (Thomas & Lok, 2015)—but with particular focus on the AI domain (Mills et al., 2023), ultimately allowing students to recognize the strengths and limitations of AI (Long & Magerko, 2020).

A critical AI literacy curriculum would provide students with the AI-related knowledge and skills to critically examine the reliability and accuracy of AI output and to interrogate AI's role in current power structures and whether and how AI could be causing harm (Mills et al., 2023). First, rather than using ChatGPT directly for output adoption, researchers suggest that students view ChatGPT as an advisory tool, or an initial resource for information gathering and idea brainstorming (Barrett & Pack, 2023; Ngo, 2023; Stojanov, 2023). For example, students can use generative AI tools like ChatGPT to outline their ideas, to gain inspiration, and to search for relevant information for their first draft, using ChatGPT to strengthen their original work at the beginning stage of writing. Relatedly, practical tools have been developed to enhance students' critical thinking skills when interacting with generative AI. For example, Lee et al. (2024) utilized a new guidance-based ChatGPT-assisted Learning Aid (GCLA) tool, which was designed to engage students in critical thinking prior to receiving ChatGPT's responses. GCLA promotes independent problem solving by providing students with hints to steer their continued engagement. This iterative process encourages students to view generative AI as an advisor that can support their agency as learners.

Second, students should also critically examine ChatGPT's output (Bitzenbauer, 2023; Siegle, 2023; Zhu et al., 2023), recognizing that no technology is infallible. Various activities can spark students' engagement in critical analysis including (1) fact-checking exercises in which students compare AI output with the original source material (Huang et al., 2023), (2) activities in which students identify gaps in AI content (Huang et al., 2023), and (3) reflective activities in which students purposefully question the "how" and "why" of AI content (Wu, 2024).

Researchers and other educators have also developed guidelines for ChatGPT best practices. For example, Pack and Maloney (2023) suggest following a four-step approach that begins with assigning a role to ChatGPT, such as a fifth-grade science teacher. The second step is assigning a purpose, such as assisting the student with their knowledge of the gravitational force of Earth on objects. The third step is specifying the bounds, such as specifying whether ChatGPT should provide suggestions or make direct modifications. The final step is refining ChatGPT's output, so the student (or teacher) receives the most helpful information. Additionally, websites such as "AI for Education" provide several detailed examples and language frames for writing high-quality prompts and obtaining more reliable AI responses (AI for education, 2023).

6 Conclusion

Our review highlighted the strengths and promises of ChatGPT. Notably, ChatGPT fosters students' autonomy and competence (Deci & Ryan, 2012) through personalized learning experiences that promote equitable learning and accessibility. We also addressed pertinent concerns like academic integrity, accuracy, and bias and provided suggestions for beginning to address these issues.

Laurillard (2008) provides a technology-enhanced pedagogy framework for balancing these promises and concerns related to ChatGPT. Specifically, generative AI technologies should not replace traditional teaching methods but rather enhance educators' creativity and innovation toward creating meaningful learning experiences for their students. Educators might therefore consider incorporating ChatGPT to the extent that it frees their time to engage in higher-order thinking, to design creative activities, and to facilitate positive student–teacher interactions.

Teachers should be supported in these endeavors through professional training that includes such topics as (1) the usage and functions of ChatGPT and (2) effective adoption of ChatGPT into their classrooms (Selwyn, 2022). These skills, coupled with increased emphasis on critical AI literacy (Mills et al., 2023), will better ensure that AI tools like ChatGPT achieve their promise for fostering innovative teaching (Holmes et al., 2019).

Appendix A

See Table 4.

Table 4 Overview of the Reviewed Studies (N = 52)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Aktay et al. (2023)	To reveal students' thoughts on using ChatGPT in the classroom	N = 15 Academic level: 4th grade Location: Turkey	Method: Qualitative Data Collection: Semi-structured interviews Data Analysis: Content analysis	Students' thoughts on ChatGPT were organized into five themes: (1) positive opinions on ChatGPT in teaching, (2) advantages of teaching with ChatGPT, (3) concerns with ChatGPT in classrooms, (4) recommendations to use ChatGPT in science and mathematics, and (5) other applications of ChatGPT in education
Alexander et al. (2023)	To explore 1) the effectiveness of plagiarism detectors in distinguishing human-generated from AI-generated academic text, 2) ESL instructors' writing assessment criteria, and 3) the effectiveness of these criteria in distinguishing between human- and AI-generated writing	N = 6 English as a Second Language lecturers Academic level: Undergraduate and graduate higher education Location: Cyprus	Method: Qualitative Data Collection: Participants provide feedback on four academic essays (1 fully human-written and 3 with varying amounts of AI-generated text) Data Analysis: Analysis of participants' review of the essays	While AI detectors could accurately identify fully AI-generated or human-written texts, their performance fell drastically in accuracy with texts written both by AI and humans. ESL teachers were less successful at detecting AI-generated texts than the AI detectors
Barrett & Pack (2023)	To explore and compare undergraduate students' and teachers' perspectives about using GenAI for various writing tasks (e.g., brainstorming, outlining, providing feedback)	N = 226 Students: 158 Teachers: 68 Academic level: Undergraduate higher education Location: United States	Method: Quantitative Design: Cross-sectional Data Collection: Questionnaire Data Analysis: Mann–Whitney U test, principal components analysis	Undergraduate students and teachers generally agreed that using GenAI to brainstorm ideas or model answers is acceptable. Conversely, using GenAI to complete writing assignments was viewed as unacceptable, regardless of whether such use was disclosed. Finally, teachers tended to have a more positive outlook than students on teacher use of AI
Bitzenbauer (2023)	To explore how an intervention impacts students' opinions about the importance of AI, in general, and ChatGPT in particular	N: 53 Academic level: 12th grade Language: German Location: Germany	Method: Quantitative Data Collection: Questionnaires Data Analysis: Descriptive statistics	Physics students' attitudes toward ChatGPT became more positive in most areas following the intervention

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Çelik et al. (2024)	To evaluate ChatGPT-simplified text's effectiveness in enhancing university-level EFL learners' reading comprehension and inferencing skills and alleviating their reading anxiety levels	N = 105 Academic Level: Undergraduate higher education Language Level: EFL, pre-intermediate/intermediate level (B1) Location: Turkey	Method: Quantitative Design: Within-subjects experimental design Treatment Condition: ChatGPT-simplified text Control Condition: Original text Data Collection: Surveys, reading comprehension test Data Analysis: Wilcoxon Signed-Rank Test	Students demonstrated statistically significant improvements in both their reading comprehension and inferencing abilities with ChatGPT-simplified texts, but no significant change in their reading anxiety
Chan & Hu (2023)	To explore 1) university students' familiarity with GenAI technologies like ChatGPT, 2) university students' perceptions of potential benefits and challenges associated with using GenAI in teaching and learning, and 3) how can GenAI be effectively integrated into higher education to enhance teaching and learning outcomes	N = 399 Academic level: Undergraduate and postgraduate Location: Hong Kong	Method: Mixed methods Data Collection: Closed-ended and open-ended online surveys and questionnaires Data Analysis: Descriptive statistical analysis, thematic analysis	University students have a generally good understanding of and positive attitude toward GenAI technologies. While students identified both benefits and concerns associated with using GenAI, they would like to integrate GenAI technologies in their learning practices and future careers
Chaudhry et al. (2023)	To use an experimental approach to assess whether ChatGPT can provide complete solutions to student assessments in important domains for professional development (e.g., critical thinking, problem-solving, communication, ethics)	ChatGPT responses Academic Level: Undergraduate higher education Subject: Bachelor of Business Administration in Human Resource Management (BBA-HRM) program Location: Abu Dhabi	Method: Quantitative Design: Case study, Quasi-experimental design Control Condition: Student course assignments/ assessments Treatment Condition: ChatGPT responses Data Collection: Randomly selected student course assignments/assessments; ChatGPT-generated responses that were graded by instructors Data Analysis: Descriptive statistics	ChatGPT demonstrated the ability to generate high-quality responses across various assessments, often matching or exceeding the performance of top students. However, limitations were observed in handling assignments requiring empirical research or extensive word counts. Also, plagiarism detection tools were not always able to identify ChatGPT-generated text

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Cong-Lem et al. (2024)	To explore 1) how Vietnamese EFL teachers perceive students' engagement in dishonest academic practices using AI in terms of their understanding of academic dishonesty enabled by AI, the underlying causes, and potential consequences for students and 2) ELL teachers' beliefs about effective strategies for preventing academic dishonesty	N = 31 EFL teachers Academic level: Higher education Location: Vietnam	Method: Qualitative Data Collection: Open-ended surveys Data Analysis: Thematic analysis	Vietnamese EFL teachers perceived the most common causes of AI-enabled dishonesty as poor student motivation, academic pressure, and a lack of linguistic skills. Further, teachers called for stricter AI-related regulations and increased student awareness
Cooper (2023)	To explore 1) how ChatGPT answers science education-related questions, 2) pedagogical strategies for utilizing ChatGPT in science education, and 3) the researcher's reflections on the use of ChatGPT in the study and as a research tool	N = 1 Faculty member Academic Level: Higher education Location: Australia	Method: Qualitative Design: Self-study, where the researcher compares ChatGPT responses to own knowledge and experience Data Collection: ChatGPT dialogue Data Analysis: Critical analysis of ChatGPT responses	ChatGPT's responses to science education-related questions are well-aligned with empirical research but fall short in the use of appropriate citations. Science educators may find ChatGPT useful when developing quizzes, rubrics, and science units, while researchers might find ChatGPT's editing capacities helpful

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Dahlkemper (2023)	To explore 1) how introductory physics students rate ChatGPT responses to phenomenological mechanics comprehension questions with respect to scientific accuracy and linguistic quality compared to a masked sample solution, 2) the impact of students' self-estimated content-related expertise on their ratings, and 3) whether discrepancies concerning the rating of scientific accuracy remain after accounting for the rated linguistic quality of the responses	N = 102 Academic Level: First- and second-year physics undergraduate students Language: German Location: Germany	Method: Quantitative Data Collection: Introductory mechanics sample test questions and 3 ChatGPT-generated and one human-written solution Participants were told that all were ChatGPT-generated responses Participants rated responses' scientific accuracy and linguistic quality using a five-point Likert scale Physics faculty also rated the responses and their ratings were used as an expert comparison Data Analysis: Exploratory factor analysis, ANOVA, ANCOVA	While university students rated the scientific accuracy of the human responses higher than the ChatGPT responses, they rated the linguistic quality similarly. There were lower gaps in ratings of scientific accuracy for students who rated their science expertise lower, and there was an association between ratings of linguistic quality and ratings of scientific accuracy
De Winter (2024)	To determine ChatGPT 3.5 and 4's capabilities on a high-stakes Dutch national high school English reading comprehension exam, both overall and in relation to human students	ChatGPT 3.5 and 4 Academic Level: High school Location: Netherlands	Method: Quantitative Data Collection: ChatGPT generated responses on three English reading comprehensive examples, using versions 3.5 and 4 Data Analysis: Bootstrapping self-consistency analysis	GPT-3.5 produced a mean score of 7.3 and GPT-4 produced a mean score of 8.3 and 8.1, outperforming average students. GPT-4 displayed fewer inconsistencies in answers compared to GPT-3.5. GPT-3.5 required re-prompting for some questions. GPT-4 performed better on multiple-choice questions but struggled occasionally with open-ended items. Higher "temperature" settings introduced randomness, revealing areas of uncertainty in GPT-4's responses

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Escalante et al. (2023)	To examine, in Study 1, whether the application of AI-generated feedback results in superior linguistic progress among ENL students compared to students who receive feedback from human tutors and, in Study 2, whether ENL students' preference for AI-generated feedback surpasses that for human tutor-generated feedback	Study 1: N = 48 Study 2: N = 43 Academic Level: Higher education Language Level: CEFR B1 English proficiency Location: Asian-Pacific Region	Study 1: Method: Quantitative methods Design: Longitudinal repeated measures quasi experimental design Treatment Condition: ChatGPT-generated feedback over six-week period Control Condition: Human feedback Data Collection: Pre- (week 1) and post-writing (week 8) assessment Analytic rubric used to score writing on coherence, content, language use, and integration of sources and evidence Data Analysis: Repeated measures ANOVA, Independent samples t-tests Study 2: Method: Mixed methods Data Collection: Participants received human-generated and ChatGPT-generated feedback over six-week period Weekly survey over six-week period to assess feedback preferences Data Analysis: Descriptive statistics, Thematic analysis	ENL students who received AI-generated feedback did not have superior linguistic progress compared to those who received feedback from a human tutor. Students' preferences for AI-generated versus human tutors were split evenly

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Evmenova et al. (2024)	To explore 1) differences between responses generated by GPT-3.5 and GPT-4 given prompts that provide varying specificity about students' essays, 2) the nature of ChatGPT's writing advice for students with and without disabilities and/or ELLs, and 3) how GPT-3.5 and GPT-4-generated formative feedback compares to the teacher-generated feedback when given the same rubric	N = 34 41% had high-incidence disabilities 24% English language learners 35% identified as struggling writers by their teachers Academic Level: Grades 3–7	Method: Qualitative Design: Secondary data analysis Data Collection: ChatGPT 3.5 and 4-generated feedback on students' essays Teacher feedback used for comparison Data Analysis: Inductive thematic analysis	Both versions of ChatGPT generated extensive feedback for each essay. GPT-4 provided more detailed instructional suggestions, often recommending specific activities. Further, both versions of ChatGPT feedback partially matched teacher feedback, but teachers provided more individualized feedback
Gregorcic & Pendrill (2023)	To explore how ChatGPT answers physics questions and the relevance of ChatGPT in physics education	ChatGPT	Method: Qualitative Design: Case study Data Collection: Dialogue between researchers and ChatGPT	ChatGPT often provides inaccurate and contradictory physics information, and therefore, it would not be useful for student cheating. Since ChatGPT can reliably formulate incorrect responses, these responses could be used to train science educators (e.g., to identify problematic scientific argumentation)
Guo & Wang (2024)	To explore 1) differences in ChatGPT- and teacher-generated writing feedback for EFL students and 2) EFL teachers' perceptions of ChatGPT-generated feedback	N = 5 EFL teachers N = 50 Students in an academic English course Academic level: Undergraduate higher education Location: China	Method: Mixed methods Data Collection: ChatGPT and the five teachers provided feedback on argumentative essays written by 50 students in an academic English course Questionnaire soliciting teachers' ratings of ChatGPT feedback quality Data Analysis: Coding feedback segments for quantitative analysis Mann–Whitney test Descriptive statistics Thematic analysis	There were differences in ChatGPT- and teacher-generated writing feedback for EFL students, such that ChatGPT generated a larger amount of feedback and more directive feedback and praise. In comparison, teachers tended to give informative feedback or including clarifying questions in their feedback. Further, teachers had mixed perceptions of the feedback provided by ChatGPT

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Huallpa (2023)	To explore university students' perspectives and opinions concerning how ChatGPA is integrated in education	N = 220 Academic level: Undergraduate Location: Latino-American universities	Method: Quantitative Design: Cross-sectional Data Collection: Questionnaires Data Analysis: Descriptive statistics, regression analysis	Individuals' reliance on AI is influenced by various demographic factors, attitudes, and experiences. Participants had a relatively favorable impression and experience with ChatGPT but were neutral regarding the general accessibility of AI. Participants also believed that institutions should outline rules concerning ethical applications for ChatGPT
Jacob et al. (2024)	To explore 1) how ChatGPT supports each stage of the academic writing process for a graduate student and second language writer, 2) to what degree the student's authentic voice comes through when writing with ChatGPT, and 3) the student's perceptions of ChatGPT as an academic writing tool over time	N = 1 Participant: "Kailing," female Chinese international student Academic Level: PhD student Language Level: Moderately proficient in English, developing academic English skills Location: United States	Method: Qualitative Design: Case study Data Collection: Four semi-structured interviews, written artifacts (e.g., research papers, proposals, ChatGPT logs) Data Analysis: Inductive, thematic qualitative coding	Kailing utilized ChatGPT extensively for brainstorming and text production. Further, Kailing took measures to ensure that her voice aligned with ChatGPT. While she was initially enthusiastic about ChatGPT's idea-generation capabilities, Kailing eventually recognized its limitations
Jauhiainen & Guerra (2023)	To explore 1) whether generative AI can provide personified learning material for students with varied knowledge about the topic in a educational lesson, 2) potential differences among students from diverse backgrounds in the amount of time spent learning generative AI-modified material, and 3) and students' perceptions of learning with generative AI-modified material	N = 110 Ages: 8–14 Academic Level: Grades 4–6 Language: Spanish Location: Montevideo, Uruguay	Method: Mixed methods Design: Case study Data Collection: ChatGPT-generated test lessons that incorporated learning content and data collection (e.g., background questions, closed- and open-response knowledge-based questions) Data Analysis: Descriptive statistics, cross-tabulations *Note: The details of the qualitative analysis were not directly specified*	Generative AI can be utilized to create personalized (social studies/history) lesson material for students in accordance with their prior knowledge and interest in the content. The majority of students enjoyed learning the generative AI-modified material and reported learning "much" or "very much." Participants' interest in the topic and enjoyment of their educational experience were also both positively correlated with students' reported learning

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Jeon & Lee (2023)	To understand 1) the pedagogical value of an LLM-powered chatbot in language education and 2) how English language teachers thought their roles might change if chatbots were used	N = 11 English language teachers Ages: 27–38 Teaching experience: 2–11 years Female: 9 Male: 2 Location: South Korea	Method: Qualitative Data Collection: Semi-structured interviews, interaction logs Data Analysis: Qualitative data analysis (i.e., developing coding scheme, identifying themes and subcategories)	Findings identified four roles for ChatGPT in language education: interlocutor, content provider, teaching assistant, and evaluator. Teachers recognized ChatGPT's potential to enhance teaching but stressed the importance of integrating it thoughtfully. Teachers also identified three primary ways that they could utilize ChatGPT to facilitate learning: using their pedagogical expertise to arrange educational resources, fostering active student investigation, and raising ethical awareness about AI
Kakhki, Oguz & Gendron (2024)	To identify the affordances of ChatGPT for higher education	N = 28 Panel members, including faculty members, administrators, and students from various U.S. universities, with expertise on ChatGPT in higher education Location: United States	Method: Qualitative Data Collection: Panel session conversations Data Analysis: Grounded theory methodology	ChatGPT has many affordances including reducing assessment challenges, supporting deep learning, fostering creativity, enhancing writing abilities, and encouraging lifelong learning, innovative teaching methods, and flexible approaches. On the other hand, ChatGPT risks undermining academic integrity, stifling creativity, and reinforcing biases

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Karakose et al. (2023)	To compare the accuracy, clarity, conciseness, and breadth of ChatGPT 3.5 and 4, which has implications for future use of using AI-based tools in educational research,	ChatGPT 3.5 and ChatGPT 4	Method: Mixed methods Data Collection: ChatGPT 3.5 and ChatGPT 4's responses to questions about the COVID-19 pandemic Ratings of ChatGPT's responses on a three-point scale Data Analysis: Cohen's kappa (inter-rater agreement), descriptive statistical analysis Analysis of ChatGPT responses in terms of their content	Both ChatGPT 3.5 and 4 show promise as a research tool, given the accuracy, clarity, conciseness, and breadth of information provided about the COVID-19 pandemic. ChatGPT 4 outperformed ChatGPT 3.5 in terms of providing synthesized, categorized, and comprehensive information
Kılınç (2023)	To explore 1) the potential roles and applications of ChatGPT in distance science education, 2) how ChatGPT can effectively create personalized and adaptive learning experiences in distance science education, and 3) the benefits and limitations of using ChatGPT in distance science education and pathways for addressing these issues	ChatGPT	Method: Qualitative Data Collection: Dialogue with ChatGPT Data Analysis: Qualitative content analysis	ChatGPT can 1) analyze students' responses to science assignments and provide individualized feedback, 2) adapt the science content in response to learners' needs and abilities, 3) support curriculum design by suggesting topics, resources, and learning activities, and 4) provide an efficient means for communication between teachers and students. However, for each of these applications, there are affordances and limitations. For example, while ChatGPT can increase students' autonomy by providing personalized feedback, ChatGPT lacks the empathy and understanding of human feedback

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Kortemeyer (2023)	To compare ChatGPT's responses to introductory physics questions compared to human-generated responses	ChatGPT	Method: Mixed methods Design: Case study Data Collection: ChatGPT responses to various assessment materials (e.g., exams, homework problems, programming exercises) Data Analysis: Descriptive statistics. Analysis of the content of the ChatGPT responses	In terms of performance, ChatGPT scored 53.05 percent overall and earned a 1.5 GPA. While this performance was sufficient to pass the course, the score reduced the overall GPA needed to graduate. The mistakes made were similar to those made by novice human physics students and relate to how ChatGPT gathers information
Lee et al. (2024)	To evaluate the effectiveness of guidance-based ChatGPT-assisted learning aid (GCLA) in enhancing self-regulated learning, higher-order thinking skills, and knowledge construction compared to traditional ChatGPT use in higher education	N = 61 Treatment group: 31 Control group: 30 Academic level: First-year undergraduate Location: Southern Taiwan	Method: Quantitative Design: Randomized controlled trial Treatment Condition: Guidance-based ChatGPT-assisted learning aid (GCLA) Control Condition: Traditional ChatGPT Data Collection: Pre-, post-, and delayed chemistry knowledge tests. Survey Data Analysis: Descriptive statistics, ANCOVA	The GCLA significantly improved students' cognitive engagement, behavioral engagement, and self-efficacy, as well as their critical thinking, problem-solving, and creativity compared to traditional ChatGPT. It also enhanced knowledge construction and retention, measured by a post-test, delivered immediately after the intervention, and a delayed test, delivered two weeks following the intervention
Lee & Zhai (2024)	To investigate 1) how widely pre-service science teachers integrate ChatGPT, 2) their proficiency in planning and integrating ChatGPT, 3) their perceptions about the usefulness of ChatGPT, and 4) their concerns about integrating ChatGPT into science teaching and learning	N = 29 Pre-service elementary science teachers (undergraduate) Location: South Korea	Method: Mixed methods Data Collection: Lesson plans collected from pre-service teachers, open-ended survey questions, data from teachers' conversations with ChatGPT for triangulation, scoring rubric to assess lesson plans Data Analysis: Descriptive statistics, constant comparative method to analyze qualitative data	Pre-service teachers integrated ChatGPT into various science domains and teaching methods. Teachers demonstrated varying levels of proficiency in selecting the most appropriate ChatGPT functions. While teachers scored high on aligning instructional strategies with ChatGPT, they demonstrated limitations in fully utilizing the technology's potential. Further, the teachers identified some uses of ChatGPT, but they also expressed concerns about the reliability and accuracy of ChatGPT

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
C. Liu et al. (2024a)	To explore 1) students' participation in chatbot training activity, in general, and its relation to students' interest and engagement in reading and 2) students' interest and engagement in reading based on both their chatbot participation and English proficiency	N = 95 Treatment group: 47 Control group: 48 Academic level: 5th Location: Taiwan	Method: Mixed methods Design: Quasi-experimental design Treatment Condition: Standard reading program in Weeks 1 and 2 and Chatbots incorporated in Weeks 3 through 9 Control Group: Standard reading program for all nine weeks Data Collection: Surveys, Chatbot dialogue, Interviews Data Analysis: Coding Chatbot dialogue, Inter-rater reliability, ANOVA, Pearson correlation analysis, cluster analysis	Students in the treatment group reported greater engagement and interest while reading, but some question types in the training field were negatively correlated with engagement and interest. Four clusters were identified and labelled as "low social connection readers," "high interest active challengers," "low proficiency moderate trainers," and "low interest active trainers" (p. 1141)
M. Liu et al. (2024b)	To explore 1) whether, and in what ways, EFL students create texts differently in generative AI-assisted digital multimodal composition and traditional writing and 2) how students produce images in generative AI-assisted digital multimodal composition	N = 8 Multimodal essay group: N = 4 Traditional essay group: N = 4 Chinese EFL international undergraduates Location: New Zealand	Method: Qualitative Multimodal Essay Condition: Participants used AI for written text and image generation Traditional Essay Condition: Participants used AI for written text only Data Collection: Screen recordings with think-aloud, semi-structured interview, multimodal text Data Analysis: Inductive qualitative coding	The multimodal group used more "bridge text" (i.e., text connecting ideas), provided more examples to support their arguments, and incorporated more summarized AI text. Students evaluated the AI-generated images based on their prior knowledge about the topic and style preference; whether the image seemed realistic and relevant; and how the intended audience might perceive the image

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Mahapatra (2024)	To explore 1) whether feedback from ChatGPT can improve undergraduate ESL students' academic writing skills and 2) students' perceptions regarding ChatGPT's impact on their academic writing	N = 134 Treatment group: N = 78 Control group: N = 56 Academic Level: Undergraduate science and engineering ESL students Location: India Ages: 18–19	Method: Mixed Methods Design: Quasi-experimental design Treatment Condition: ChatGPT as a writing feedback tool Control Condition: No writing feedback from ChatGPT Data Collection: Writing pre- and two post-tests (immediate and delayed), focus group discussions Data Analysis: Repeated measures ANOVA, Independent samples t-test, phronetic iterative qualitative analysis	The gains in academic writing were higher for students in the treatment group. Additionally, students viewed ChatGPT positively overall, noting such positive features as improved grammar, ability to promote autonomy and peer collaboration. Some concerns included decreased motivation to think and dependency on AI
Mena Octavio et al. (2024)	To explore an EFL teacher's process integrating ChatGPT into her classroom, including strategies used to obtain coherent and qualitative responses from ChatGPT and validate ChatGPT's output, and to determine whether ChatGPT was effective in supporting the teacher's instructional goals	Primary Participant: N = 1 One experienced female EFL teacher with over ten years of teaching experience; actively using ChatGPT for lesson planning and classroom tasks Secondary participants: Unspecified sample size; Students aged 4 to adults (levels A1–C1 according to CEFR) Location: Spain	Method: Qualitative Design: Single instrumental case study Data Collection: ChatGPT interaction logs, teacher-designed lesson plans, semi-structured post-study interviews Data Analysis: Thematic analysis	The teacher used ChatGPT for lesson planning, in-class activities, and assessments. ChatGPT significantly reduced workload and generated level-appropriate lesson plans that engaged students. The teacher also developed a five-step strategy for effective prompt crafting. Teachers validated ChatGPT's output by cross-referencing the information with other sources, reviewing the output with colleagues and students, and monitoring the output for potential biases

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Mizumoto & Eguchi (2023)	To determine the reliability of GPT-3.5 for Automated Essay Scoring (AES) tasks and to assess how linguistic features contribute to enhancing GPT's scoring accuracy	Essays from 12,100 English learners across 11 native language groups No human participants; data were derived from the TOEFL11 corpus	Method: Quantitative Data Collection: Text-davinci-003 model with IELTS Writing Band Descriptors as the rubric Data Analysis: Descriptive statistics, Regression, Cross-validation, Scoring agreement	The study showed that AES scores by GPT-3.5 are somewhat reliable. GPT-3.5 AES performed moderately with fair-to-moderate agreement. The inclusion of linguistic features showed some improvement
Mogavi et al. (2024)	To explore how early adopters predominantly use and perceive ChatGPT in education, including ChatGPT's influence on the educational process	N = 6,000 text samples N = 1,500 text samples each from Twitter (recently X), Reddit, YouTube, and LinkedIn N = 300 text samples *per platform* from researchers, educators, students, parents, and general users	Method: Qualitative Data Collection: Using APIs and data scraping to gather social media data on the use of ChatGPT in education Stratified random sampling of text samples by social media platform and early adopter role Data Analysis: Thematic inductive analysis	ChatGPT is most frequently utilized in K-12 education, higher education, and skills training. Social media discussions revealed that productivity, efficiency, and ethics are the predominant topics related to the integration of ChatGPT into education
Mohamed (2024)	To investigate faculty members' perceptions of the advantages and disadvantages of utilizing ChatGPT for EFL teaching and learning	N = 10 EFL university teachers with MA or Ph.D Academic level: Higher education Location: Saudi Arabia	Method: Qualitative Data Collection: Interviews (in-person and email) Data Analysis: Qualitative content analysis	Faculty identified the following advantages of ChatGPT in language learning: real-time feedback, personalized instruction, extension information base, cost effectiveness, and human-like interaction. On the other hand, the disadvantages include limitations in human connections, contextual inaccuracies, inadequate phonetic or intonation support, cultural insensitivity, and ethical concerns. ChatGPT can be integrated into EFL education by complementing traditional methods and providing efficient assessment, skills development, and teacher training

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Ngo (2023)	To assess students' perceptions of using ChatGPT in learning, including their overall perception, perceived benefits and barriers, and possible solutions for effectively utilizing ChatGPT	N = 200 Academic level: Higher education Location: Vietnam	Method: Mixed methods Data Collection: Questionnaires, semi-structured interviews Data Analysis: Descriptive statistics, one-sample t-tests, thematic analysis	Students had an overall favorable opinion of ChatGPT's application. Students perceived the tool's simplicity as a major benefit, but they were also particularly aware of ChatGPT's barriers (e.g., concerns about the credibility of the information provided). Finally, students identified possible solutions for utilizing ChatGPT including cross-referencing ChatGPT information with other sources. Students also recommended that academic guidelines be developed regarding ChatGPT use
Nikolic et al. (2023)	To explore how ChatGPT might affect assessment methods and be used to support learning in engineering education	ChatGPT responses Academic level: Higher education Location: Australia	Method: Miscellaneous Data Collection: 9 faculty with engineering expertise conducted iterative cycling through of ChatGPT responses to varying assessment prompts on ten subjects and different types of assessments (e.g., online quizzes, written research-based assessments) in engineering education Data analysis: SWOT Analysis	ChatGPT could pass some subjects and perform well on certain types of assessments, but it struggled with tasks involving detailed responses or specific academic formatting. ChatGPT's utility is also related to the skill level of the user, in that some users could potentially obtain "passable" responses that reflect critical reflection. ChatGPT can be useful for helping students in the classroom by providing instant feedback, for example
Perkins et al. (2024)	To explore the capability of academic staff assisted by the Turnitin Artificial Intelligence (AI) detection tool to identify the use of AI-generated content in university assessments	N = 15 University staff members (faculty) Location: Southeast Asia	Method: Mixed methods Data Collection: Scores on assessment, Participant reports of potential AI-generated material, Feedback on assessment Data Analysis: Descriptive statistics, (Unspecified) Statistical significance test, Qualitative analysis of participants' feedback on assessment	Turnitin's AI tool was generally effective at detecting AI-generated content but showed limitations when handling prompt-engineered or paraphrased content. Participants provided similar scores, on average, to genuine and AI-generated content

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Punar Özçelik & Yangin Eksi (2024)	To explore 1) ChatGPT's potential for helping students improve their writing through self-editing and 2) students' perceptions about ChatGPT's utility as a learning assistant	N = 11 Academic level: Undergraduate Language level: Above B1 Location: Turkey	Method: Qualitative Design: One-shot case study, Pre-experimental design Treatment: Two 50-min online writing lessons for two weeks Data Collection: Observations with fieldnotes, Unstructured open-ended interviews Data Analysis: Thematic qualitative data analysis	While students found ChatGPT beneficial for writing (e.g., providing feedback, improving formality of tone), they also reported limitations (e.g., technical issues, incorrect corrections). Students also had mixed opinions about using ChatGPT for formal and informal texts. Finally, students suggested being more cautious with ChatGPT's proposed changes, since some changes can modify the overall meaning
Rahman & Watanobe (2023)	To 1) investigate the opportunities afforded by ChatGPT in education, 2) identify possible threats and related solutions concerning ChatGPT in education, 3) illustrate how ChatGPT could support programming education, and 4) understand postsecondary students' and teachers' perspectives on how ChatGPT supports programming learning and teaching	N = Unspecified Higher education students and teachers Academic Level: Undergraduate and graduate higher education	Method: Mixed methods Data Collection: Questionnaires, ChatGPT coding-related experiments Data Analysis: Descriptive statistics; Analysis of the content generated by ChatGPT	ChatGPT provides opportunities for learners, educators, and researchers, including personalized learning support and grammar support. ChatGPT can support programming education by providing opportunities for routine practice and personalized learning support. Students and teachers were generally satisfied with the programming support provided by ChatGPT. At the same time, ChatGPT presents some challenges to educators including concerns about academic integrity and over-reliance on AI. The authors recommend using plagiarism detection tools to identify AI-generated text, among other strategies

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Rakap (2024)	To investigate 1) how the use of ChatGPT correlates to the quality, content, and time spent on IEP goal development for novice special education teachers and 2) the relationship between previous training on IEP and goal development and use of ChatGPT on IEP goal development	N = 22 Novice special education teachers Treatment Group: N = 11 Control Group: N = 11 Ages: 22–27	Method: Quantitative Design: Experimental study Treatment Group: Tasked with using ChatGPT to generate IEP goals Control Group: No ChatGPT assistance Data Collection: Demographic information form, IEP goals developed by participants, Time spent developing IEP goals, Ratings of IEP quality Data Analysis: Inter-rater reliability, two-sample t-test, Chi-square analysis, one-way ANOVA with post-hoc tests, Pearson product-moment correlation	Mean scores on goal quality were higher for the ChatGPT group, even among novice teachers without prior training. Novice teachers in the ChatGPT group also spent less time on their goals. These findings highlight the potential for ChatGPT as an effective tool in IEP goal development
Romero Rodríguez et al. (2023)	To explore ChatGPT's acceptance by undergraduate higher education students	N = 400 Academic level: Undergraduate higher education Location: University of Granada, Spain	Method: Quantitative Design: Cross-sectional design Data Collection: Survey (demographics, technology acceptance and use) Data Analysis: Scale reliability, descriptive statistics, Independent samples t-test, ANOVA, convergent and divergent validity, structural equation modeling	User experience conditioned higher scores on all seven main constructs (i.e., factors hypothesized as related to the adoption and use of ChatGPT), but gender was not a determining characteristic for any. Several factors were influential in students' behavioral intentions to use ChatGPT including user experience, performance expectancy (i.e., academic performance), hedonic motivation (i.e., pleasure), price value, and habit (i.e., automatically learned behaviors). Finally, the conditioning factors in user behavior were facilitating conditions (i.e., support for technology use), habit, and behavioral intention

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Shabara et al. (2024)	To compare the accuracy and consistency of ChatGPT's L2 writing scores and teacher-moderated scores	N = 11 Trained and experienced teachers (Median = 12.9 years experience) N = 100 Student essays Academic Level: Undergraduate higher education (international branch university) Language Level: CEFR B1-level undergraduate L2 learners Location: Egypt	Method: Quantitative Design: Quantitative correlational non-experimental design Data Collection: 100 randomly selected essays out of a pool of 599 Scores on analytic rubric with three writing dimensions Two (or three, in cases of disagreement) sets of scores from human-raters and two sets of ChatGPT scores Data Analysis: Intraclass correlation coefficients (ICC), Descriptive statistics, Paired and independent samples t-tests	ChatGPT's inter-rater reliability was moderate. When compared to teachers' scores, ChatGPT's results showed poor-to-moderate agreement. Of the three writing dimensions, the agreement was weakest on organization, content, and relevance and highest on communication quality, academic vocabulary, and style. ChatGPT gave higher scores than teachers in all areas except language use for which there were no significant differences
Stojanov (2023)	To explore the process of learning how ChatGPT works using autoethnography	N = 1 Higher education faculty member Self-described as having a "rudimentary" knowledge of computer programming Location: New Zealand	Method: Qualitative Design: Autoethnography Data Collection: About 7 h of conversation with ChatGPT 3.5; watching YouTube videos about ChatGPT; observations and reflections; Feedback on summary of ChatGPT's functionality from ChatGPT and external expert Data Analysis: Using observations, reflections, and feedback to write a report about how ChatGPT works	ChatGPT provided helpful feedback and scaffolding, but it sometimes also provided contradictory information. ChatGPT should thus be used with caution. Educators should document their experience using ChatGPT to better understand its affordances and constraints
Sullivan et al. (2023)	To investigate how ChatGPT is disrupting higher education and consider the implications of AI tools in universities	N = 100 News articles published in New Zealand, Australia, UK, and the US	Method: Qualitative Data Collection: Systemic search for English-language news articles about ChatGPT in higher education Data Analysis: Thematic analysis, sentiment analysis	The most common themes were academic integrity and avoidance of ChatGPT. The three additional themes concerned ChatGPT policies, embracing ChatGPT, and voice (i.e., who is talking about ChatGPT?). Additionally, sentiment was roughly evenly distributed between positive and negative comments

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Taktak et al. (2024)	To assess the strengths and weaknesses of ChatGPT in K-12 education, the opportunities it creates for innovative teaching and learning, and the potential threats it poses to the educational process	N = 21 17 teachers, 4 principals Academic level: K-12 Location: Turkey	Method: Qualitative Data Collection: Semi-structured face-to-face interviews Data Analysis: Content analysis categorized into SWOT (Strengths, Weaknesses, Opportunities, Threats)	The SWOT analysis revealed that the strengths of ChatGPT in K-12 education include personalized learning, efficiency, accessible feedback, and content richness. The weaknesses include misinformation, ethical concerns, lack of higher-order thinking, and dependence on AI. ChatGPT provided opportunities such as smart classrooms, lesson planning, skill development, and collaborative learning. ChatGPT's threats to education included reduced teacher-student interaction, biases and discrimination, privacy and security, and virtual addiction
Tlili et al. (2023)	To identify concerns about the utilization of chatbots in education, focusing on ChatGPT	N = 2330 Tweets from 1530 Twitter users N = 19 stakeholders (e.g., students, educators, developers) who posted their experiences with ChatGPT publicly through blogs N = 3 experienced educators	Method: Qualitative Design: Instrumental case study design Data Collection: Tweets, Interviews, Stakeholder ratings of their ChatGPT experience, Daily meetings with the 3 experienced educators regarding their ChatGPT user experience Data Analysis: Social network analysis (sentiment analysis) of tweets, content analysis, summary of user experiences	The public holds diverse views on the use of chatbots. Most users were optimistic about the use of AI, but they also raised concerns about its use in educational contexts. The interviews centered around five themes: educational transformation, response quality, personality and emotions, usefulness, and ethics

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Tseng & Lin (2024)	To evaluate the effectiveness of integrating ChatGPT into EFL writing instruction using the Technological Pedagogical Content Knowledge (TPACK) and Analysis, Design, Development, Implementation, and Evaluation (ADDIE) models	N = 15 Academic level: EFL, junior or senior undergraduate higher education Ages: 20–22 Location: Taiwan	Method: Qualitative Data Collection: Student essays, reflective writings Data Analysis: Qualitative analysis consistent with ADDIE framework	The study revealed that TPACK and ADDIE can integrate ChatGPT for generating, organizing, and revising essays. It also showed that systematic scaffolding helped students transition from relying on AI to critically engaging with the outputs. Students appreciated AI's efficiency but expressed concerns about originality and over-reliance. Students' reflections also highlighted the balance between leveraging AI and maintaining a unique authorial voice
Tülübaş et al. (2023)	To evaluate ChatGPT's potential to produce accurate, clear, concise, and unbiased information on a specified research topic	ChatGPT 3.5 and 4	Method: Mixed methods Data Collection: ChatGPT dialogue, Ratings of ChatGPT responses, focus group among raters Data Analysis: Inter-rater reliability (Cohen's kappa), Descriptive statistics, Comparative content analysis of ChatGPT 3.5 and 4	Both ChatGPT 3.5 and 4 produced accurate information about emergency remote teaching, but ChatGPT 4 performed better on queries requiring more judgment in terms of the level of clarity, detail, and synthesis provided. While ChatGPT is a promising technology, humans are still needed to ensure that high quality output is produced
Urrutia & Araya (2024)	To explore 1) how Large Language Models (LLMs) perform in detecting incoherent responses to fourth-grade math word problems and 2) the comparative performance of LLMs and Machine Learning (ML) classifiers	N = 677 open-ended test responses Academic Level: Fourth grade	Method: Quantitative Data Collection: Fourth-grade students' responses to open-ended math test items; the responses were fed through detection software; the performance of several AI tools were assessed using several prompting approaches Data Analysis: Descriptive statistics of various indicators of accuracy – e.g., precision, recall, F1 (a joint measure of precision and recall); Statistical significance test	LLMs were less successful in detecting incoherent fourth graders' responses to math problems than the ML models, the latter of which had been trained on thousands of examples. Recursive questions – i.e., questions for which students must discuss the accuracy of an answer – presented the greatest challenge for the LLMs

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
X. Wang et al. (2024b)	To explore 1) how a chat-based reading comprehension support (ChatPRCS) system quantifies reading comprehension skills to generate personalized questions for students; 2) the quality of these questions and whether they address students' learning needs; and 3) whether ChatPRCS impacts students' learning achievement, motivation, and cognitive load	N = 42 Education technology majors Treatment Group: N = 22 Control Group: N = 20 Academic level: Undergraduate higher education Location: China	Method: Mixed methods Design: Quasi-experimental study Treatment Group: ChatPRCS reading comprehension support Control Group: Standardized questions for reading comprehension practice Data Collection: Pre- and post-study tests, questionnaires, qualitative interviews Data Analysis: Reliability (Cronbach's alpha), Descriptive statistics, independent sample t-tests, ANCOVA, Mann-Whitney U Test, Qualitative analysis	First, ChatPRCS quantified students' reading comprehension skills using four metrics: academic knowledge level, effective learning duration, historical performance, and answer precision. Historical performance was especially critical for accurately identifying students' reading comprehension skills. Next, questions developed by experts and machine-generated questions were similar in quality and design. Finally, compared to the control group, the treatment group scored higher on learning achievement and learning motivation. However, the findings concerning cognitive load were split, such that there were no significant differences in terms of mental effort, but the treatment group scored higher on mental load
Yan (2023)	To investigate how undergraduate students respond to ChatGPT being applied in L2 writing classes, whether students can grasp its functions well, and how they perceive using AI in writing	N = 8 EFL majors Academic level: Undergraduate Location: China	Method: Qualitative Data Collection: Classroom observation, learning logs, in-depth interviews Data Analysis: Document analysis, thematic analysis	Students readily learned to leverage ChatGPT in their writing and collaborative activities improve their proficiency. While students recognized the strengths of ChatGPT, they expressed more concern about its potential application than satisfaction, especially for educational equity

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Yang (2024)	To assess 1) differences in average writing composition scores between ChatGPT and human writers, 2) the intra-rater reliability of ChatGPT scores, and 3) the inter-rater reliability between ChatGPT and human-generated scores	N = 3 university instructors with considerable writing instruction experience (human raters) N = 82 EFL writing compositions Academic Level: Undergraduate higher education Location: China	Method: Quantitative Data Collection: 82 expository essays sampled from the Written English Corpus of Chinese Learners (WECCL); two human raters and ChatGPT (two times) scored each essay on language, content, and organization Data Analysis: Descriptive statistics, inter-rater reliability (Pearson correlation), independent sample t-test, intra-rater reliability (Pearson correlation)	Despite ChatGPT's potential use for evaluating students' writing, current limitations must be addressed before it can be considered a reliability tool. For example, ChatGPT's intra-rater reliability falls below acceptable standards. The inter-rater reliability for the overall scores was moderate but deemed unreliable for assessing organizational features of writing. Also, while there were no significant difference in average scores between ChatGPT and human raters, ChatGPT displayed a wider range of scores
Yeadon et al. (2023)	To explore the performance of AI-generated responses to five short-answer physics questions	N = 10 AI-generated scripts with 5 short essay responses each (ChatGPT and davinci-003) Academic Level: Undergraduate higher education Subject: Physics Location: Durham University, UK	Method: Mixed methods Data Collection: 10 AI-generated scripts with 5 short essay responses each (ChatGPT and davinci-003); Independent scoring by five human raters on a scale of 0–100; Detection of potential AI-generated text using Grammarly, TurnitIn, OpenAI, and GPTZero Data Analysis: Descriptive statistics, qualitative analysis of essay content	DaVinci-003 and ChatGPT received scores on their short-form Physics essays that were comparable to second-year Physics students. Further, plagiarism detection tools were unreliable in detecting AI-generated work, even those tools designed specifically to detect AI-generated texts

Table 4 (continued)

Title	Purpose statement	Participants/ contexts/ demographic/ curriculum	Methodology/data collection and data analysis	Findings
Zhu et al. (2023)	To conduct a SWOT (strengths, weakness, opportunity, and threat) analysis concerning ChatGPT's use in teaching and learning	N/A	Method: Miscellaneous Data Collection: N/A Data Analysis: SWOT analysis	The SWOT analysis revealed that increased growth in online learning and students' need for personal support provides an opportunity to leverage ChatGPT in education, At the same time, threats to ChatGPT's integration include concerns about cheating and general resistance from educators. Some of ChatGPT's key strengths include providing feedback and engaging in human-like conversations, while some weaknesses include inaccurate and biased information

Declarations

Conflict of interest The authors declare that they have no competing interests. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

AEM Center. (2023). *What is accessibility?* Retrieved from https://aem.cast.org/get-started/defining-accessibility

AI for education. (2023). *GenAI chatbot prompt library for educators.* Retrieved from https://www.aiforeducation.io/prompt-library

Aktay, S., Seçkin, G. Ö. K., & Uzunoğlu, D. (2023). ChatGPT in education. *Türk Akademik Yayınlar Dergisi (TAY Journal), 7*(2), 378–406.

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology, 23*(2), 25–43.

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education, 20*(1), Article 59. https://doi.org/10.1186/s41239-023-00427-0

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology, 15*(3), Article ep430. https://doi.org/10.30935/cedtech/13176

Borja, L. A, Soto, S. T & Sanchez, T. X. (2015). Differentiating Instruction for EFL Learners. *International Journal of Humanities and Social Science 5 8*(1):30–36

Çelik, F., Yangın Ersanlı, C., & Arslanbay, G. (2024). Does AI simplification of authentic blog texts improve reading comprehension, inferencing, and anxiety? A one-shot intervention in Turkish EFL context. *The International Review of Research in Open and Distributed Learning, 25*(3), 287–303. https://doi.org/10.19173/irrodl.v25i3.7779

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT: A case study. *Cogent Education, 10*(1), 1–18. https://doi.org/10.1080/2331186X.2023.2210461

Cong-Lem, N., Tran, N. T., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and responses of Vietnamese EFL teachers. *Teaching English with Technology, 24*(1), 28–47.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 454–452. https://doi.org/10.1007/s10956-023-10039-y

Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research, 19*(1), Article 010142. https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

De Winter, J. C. F. (2024). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education, 34*(3), 915–930. https://doi.org/10.1007/s40593-023-00372-z

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 413–437). Sage.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), Article 57.

Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R. (2024). Improving writing feedback for struggling writers: Generative AI to the rescue? *TechTrends, 68*(4), 790–802. https://doi.org/10.1007/s11528-024-00965-y

Gregorcic, B., & Pendrill, A. M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), Article 035021.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

He, J., Li, L., Yao W., & Gao, H. (2024). Exploring future education: The innovative integration and practice of multimodal learning and ChatGPT. In: *2024 5th International Conference on Computer Science, Engineering, and Education (CSEE)* (pp. 18–23). https://doi.org/10.1109/CSEE63195.2024.00012.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education*. Center for Curriculum Redesign.

Huallpa, J. J. J., Arocutipa, P. F., Panduro, W. D., Huete, L. C., Flores Limo, F. A., Herrera, E. E., Callacna, R. A. A., Ariza Flores, C. A., Median Romero, M. A., Quispe, I. M., & Hernandez, F. A. (2023). Exploring the ethical considerations of using Chat GPT in university education. *Periodicals of Engineering and Natural Sciences, 11*(4), 105–115. https://doi.org/10.21533/pen.v11.i4.200

Huang, C., Coleman, M., Gachago, D., & Van Belle, J. (2023). Using ChatGPT to encourage critical AI literacy skills and for assessment in higher education. In: *Communications in computer and information science* (pp. 105–118). https://doi.org/10.1007/978-3-031-48536-7_8

Jacob, S. R., Tate, T., & Warschauer, M. (2024). Emergent AI-assisted discourse: A case study of a second language writer authoring with ChatGPT. *Journal of China Computer-Assisted Language Learning*. https://doi.org/10.1515/jccall-2024-0011

Jauhiainen, J. S., & Guerra, A. G. (2023). Generative AI and ChatGPT in school children's education: Evidence from a school lesson. *Sustainability, 15*(18), 14025. https://doi.org/10.3390/su151814025

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Kakhki, M., Oguz, A., & Gendron, M. (2024). Exploring the affordances of chatbots in higher education: A framework for understanding and utilizing ChatGPT. *Journal of Information Systems Education, 35*(3), 284–302. https://doi.org/10.62273/UIRX9922

Karakose, T., Demirkol, M., Aslan, N., Köse, H., & Yirci, R. (2023). A conversation with ChatGPT about the impact of the COVID-19 pandemic on education: Comparative review based on human–AI collaboration. *Educational Process International Journal, 12*(3), 7–25.

Kılınç, S. (2023). Embracing the future of distance science education: Opportunities and challenges of ChatGPT integration. *Asian Journal of Distance Education, 18*(1), 205–237.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research, 19*(1), Article 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Laurillard, D. (2008). Technology enhanced learning as a tool for pedagogical innovation. *Journal of Philosophy of Education, 42*(3–4), 521–533. https://doi.org/10.1111/j.1467-9752.2008.00658.x

Lee, G.-G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies, 17*, 1683–1700. https://doi.org/10.1109/TLT.2024.3401457

Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education, 21*(1), 16–28. https://doi.org/10.1186/s41239-024-00457-4

Liu, C.-C., Chen, W.-J., Lo, F., Chang, C.-H., & Lin, H.-M. (2024a). Teachable Q&A agent: The effect of chatbot training by students on reading interest and engagement. *Journal of Educational Computing Research, 62*(4), 1122–1154. https://doi.org/10.1177/07356331241236467

Liu, M., Zhang, L. J., & Biebricher, C. (2024b). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education, 211*, Article 104977. https://doi.org/10.1016/j.compedu.2023.104977

Long, D., & Magerko, B. (2020). *What is AI literacy? Competencies and design considerations.* In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Mena Octavio, M., González Argüello, M. V., & Pujolà, J.-T. (2024). ChatGPT as an AI L2 teaching support: A case study of an EFL teacher. *Technology in Language Teaching & Learning, 6*(1), Article 1142. https://doi.org/10.29140/tltl.v6n1.1142

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. Jossey-Bass.

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. CAST Professional Publishing.

Mills, A., Bali, M., & Eaton, L. (2023). How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process. *Journal of Applied Learning and Teaching, 6*(1), 16–30. https://doi.org/10.37074/jalt.2023.6.1.34

Minow, M. (2021). Equality vs. equity. *American Journal of Law and Equality, 1*, 167–168. https://doi.org/10.1162/ajle_a_00019

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., & Hui, P. (2024). ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100027. https://doi.org/10.1016/j.chbah.2023.100027

Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies, 29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies. https://doi.org/10.17226/13398

Ngo, T. T. A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning (Online), 18*(17), 4–19. https://doi.org/10.3991/ijet.v18i17.39019

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

OpenAI. (2025). Overview of ChatGPT and Sora. Retrieved from https://openai.com/chatgpt/overview/

Ormrod, J. (2011). Chapter 7: Knowledge construction. In *Educational psychology: Developing learners* (7th ed., pp. 217–236). Pearson.

Pack, A., & Maloney, J. (2023). Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. *Teaching English with Technology, 23*(2), 4–24.

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6

Punar Özçelik, N., & Yangin Eksi, G. (2024). Cultivating writing skills: The role of ChatGPT as a learning assistant—a case study. *Smart Learning Environments, 11*(1), 10–18. https://doi.org/10.1186/s40561-024-00296-8

Rachmad, Y. E. (2022). *Innovation in education theory.* Amiens Cathédrale Éditions Internationales, Édition Spéciale 2022. https://doi.org/10.17605/osf.io/rsqym

Rahman, M. M., & Watanabe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783.

Rakap, S. (2024). Chatting with GPT: Enhancing individualized education program goal development for novice special education teachers. *Journal of Special Education Technology*, *39*,(3), 339–348. https://doi.org/10.1177/01626434231211295

Recite me. (2024). *The importance of accessibility in education.* Retrieved from https://reciteme.com/us/news/accessibility-in-education/

Romero Rodríguez, J. M., Ramírez-Montoya, M. S., Buenestado Fernández, M., & Lara Lara, F. (2023). Use of ChatGPT at university as a tool for complex thinking: Students' perceived usefulness. *Journal of New Approaches in Educational Research, 12*(2), 323–339.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage.

Selwyn, N. (2022). *Education and technology: Key issues and debates*. Bloomsbury Publishing.

Shabara, R., ElEbyary, K., & Boraie, D. (2024). Teachers or ChatGPT: The issue of accuracy and consistency in L2 assessment. *Teaching English with Technology*. https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215

Siegle, D. (2023). A role for ChatGPT and AI in gifted education. *Gifted Child Today, 46*(3), 211–219. https://doi.org/10.1177/10762175231168443

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education, 20*(1), Article 35.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching, 6*(1), 1–10. https://doi.org/10.37074/jalt.2023.6.1.17

Taktak, M., Bellibaş, M. Ş, & Özgenel, M. (2024). Use of ChatGPT in education: Future strategic road map with SWOT analysis. *Educational Process International Journal*. https://doi.org/10.22521/edupij.2024.133.1

Tate, T., Doroudi, S., Ritchie, D., Xu, Y., & Warschauer, M. (2023). *Educational research and AI-generated writing: Confronting the coming tsunami*. EdArXiv. https://osf.io/preprints/edarxiv/4mec3_v1

Thomas, K., & Lok, B. (2015). Teaching critical thinking: An operational framework. In J. F. Lau & M. Davies (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 93–105). Palgrave Macmillan US.

Thorne, S. L. (2024). Generative artificial intelligence, co-evolution, and language education. *Modern Language Journal, 108*, 567–572. https://doi.org/10.1111/modl.12932

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), Article 15.

Tseng, Y.-C., & Lin, Y.-H. (2024). Enhancing English as a Foreign Language (EFL) learners' writing with ChatGPT: A university-level course design. *Electronic Journal of E-Learning, 22*(2), 78–97. https://doi.org/10.34190/ejel.21.5.3329

Tülübaş, T., Demirkol, M., Ozdemir, T. Y., Polat, H., Karakose, T., & Yirci, R. (2023). An interview with ChatGPT on emergency remote teaching: A comparative analysis based on human–AI collaboration. *Educational Process: International Journal, 12*(2), 93–110. https://doi.org/10.22521/edupij.2023.122.6

Urrutia, F., & Araya, R. (2024). Who's the best detective? Large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *Journal of Educational Computing Research, 61*(8), 187–218. https://doi.org/10.1177/07356331231191174

U.S. Department of Education. (2010). *Transforming American education: Learning powered by technology*. https://files.eric.ed.gov/fulltext/ED512681.pdf

Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes*. Harvard University Press.

Wang, C., Li, Z., & Bonk, C. (2024a). Understanding self-directed learning in AI-assisted writing: A mixed methods study of postsecondary learners. *Computers and Education Artificial Intelligence, 6*, Article 100247. https://doi.org/10.1016/j.caeai.2024.100247

Wang, X., Zhong, Y., Huang, C., & Huang, X. (2024b). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies, 17*, 1762–1776. https://doi.org/10.1109/tlt.2024.3405747

Williams, C. (2023). *Hype, or the future of learning and teaching? 3 Limits to AI's ability to write student essays*. London School of Economics Internet Blog. https://kar.kent.ac.uk/99505/

Wu, Y. (2024). Critical thinking pedagogics design in an era of ChatGPT and other AI tools—Shifting from teaching "what" to teaching "why" and "how." *Journal of Education and Development, 8*(1), Article 1. https://doi.org/10.20849/jed.v8i1.1404

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Yang, Y. (2024). The reliability of using ChatGPT in rating EFL writings. *Shanlax International Journal of Education, 12*(4), 49–59. https://doi.org/10.34293/education.v12i4.7855

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), Article 035027. https://doi.org/10.1088/1361-6552/acc5cf

Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning, 15*(2), 133–152. https://doi.org/10.34105/j.kmel.2023.15.008