Journal of Science Education and Technology
https://doi.org/10.1007/s10956-025-10260-x

Chatbots in Science Education: A Scoping Review of Early Empirical Evidence

Mario Calvo-Utrilla[1,2,3] · **Esther Paños**[1,3] · **José-Reyes Ruíz-Gallardo**[1,3]

Received: 27 January 2025 / Accepted: 10 September 2025

Abstract
Chatbots are making a strong entry into education, supporting both students and teachers. This study aims to deepen the understanding of the educational use of chatbots in science education, including their advantages and limitations. A scoping review of the articles published up to January 1, 2025, was conducted following PRISMA guidelines across the Web of Science, Scopus, and ERIC databases, using search terms related to science education and chatbots. From an initial pool of 608 articles, 40 met all inclusion criteria. Most of the selected studies were exploratory (32.5%), with fewer intervention-based designs. Chatbots in science education are promising tools but are still in their early stages, and this could explain the large number of exploratory studies found. ChatGPT is the most studied and has demonstrated excellent linguistic capabilities but needs to improve its scientific accuracy and analytical skills. Tutoring students is the most commonly found application of chatbots. They also have the potential to support teachers by reducing their workload, although empirical data are needed to confirm this. Integrating Artificial Intelligence literacy and critical thinking skills into curricula, alongside comprehensive teacher professional development, is crucial for the effective and responsible use of chatbots in science education.

Keywords Chatbots · ChatGPT · Scoping review · Science teaching · Artificial intelligence

Introduction

The emergence of certain technological advances has historically generated great expectations, skepticism, and even fears (De Bruyckere et al., 2015). Throughout history, there are examples in the educational field, such as the advent of the phonograph and later cinema, about which Thomas Alva Edison predicted radical changes in teaching methods, even replacing teachers (Kirschner & Hendrick, 2020). Will the same happen with Artificial Intelligence (AI)? This significant technological advance threatens to revolutionize many everyday aspects, including education.

In reality, AI has existed since the emergence of ELIZA in the mid-twentieth century (Dahlkemper et al., 2023), but it was thanks to the appearance of ChatGPT (Generative Pre-trained Transformer) by the company OpenAI on November 30, 2022, that public opinion has begun to pay special attention (Bitzenbauer, 2023; Humphry & Fuller, 2023; Wang, 2023).

As part of the broader set of AI tools, chatbots promise to tackle several long-standing hurdles in science teaching. First, their dialogic format can help surface and confront students' deeply rooted misconceptions, such as those documented in physics and biology, by posing refutational questions and offering targeted analogies (Holmes et al., 2022; Labadze et al., 2023). Second, they can deliver step-by-step formative feedback on multistage quantitative problems, reducing the delay that typically accompanies paper-based assignments, a critical need in mathematically dense topics like stoichiometry or kinematics (Park & Martin, 2024). Third, by segmenting explanations and regulating information flow in real time, chatbots can lower

✉ Mario Calvo-Utrilla
Mario.Calvo4@alu.uclm.es

Esther Paños
Esther.Panos@uclm.es

José-Reyes Ruíz-Gallardo
JoseReyes.Ruiz@uclm.es

[1] Department of Pedagogy, Science Education, School of Education, University of Castilla-La Mancha, Plaza de La Universidad, 3, 02071 Albacete, Spain

[2] Duque de Alarcón Secondary School, Valera de Abajo, Cuenca, Spain

[3] Botany, Ethnobiology, and Education Research Group, Botanic Institute, Albacete, Spain

Published online: 10 October 2025

extraneous cognitive load in abstract domains where students often struggle to link representations, precisely the challenge highlighted by Ekici (2016) and Erinosho (2013) for physics. Finally, teachers can redeploy the time saved on routine grading or rubric design towards higher-order instructional tasks, thereby addressing workload concerns flagged in numerous STEM settings (Hwang & Chang, 2021; Kuhail et al., 2023).

Despite these compelling affordances, several concerns have also been raised regarding the use of chatbots, which question their validity as educational tools. For instance, some authors have warned that systems such as ChatGPT may produce "hallucinations" (Feldman-Maggor et al., 2025; Wang, 2023) and present incorrect information as if it were true (Park & Martin, 2024). In addition, it has been noted that these tools could exacerbate educational inequalities by reproducing or even amplifying existing biases (Avraamidou, 2024; Feldman-Maggor et al., 2025). Furthermore, some studies question their effectiveness for specific tasks, such as tutoring students (Dahlkemper et al., 2023; Ding et al., 2023) or addressing numerical problems (Sperling & Lincoln, 2024; Wan & Chen, 2024).

Regarding the systematic reviews on chatbots already published, some were conducted before the emergence of ChatGPT, such as those by Kuhail et al. (2023) and Hwang and Chang (2021), while others explicitly excluded it due to its novelty (Debets et al., 2025). Some reviews that included ChatGPT considered only a small number of studies, such as Park and Martin's (2024) analysis of 14 articles. Moreover, some reviews also highlight the need to improve methodological rigor in assessing how chatbots are evaluated for their educational effectiveness (Debets et al., 2025). Finally, several reviews addressing chatbots in science education also included research from mathematics, health sciences, or computer programming (Alneyadi & Wardat, 2023; Debets et al., 2025), which may limit the transferability of their conclusions to science education in the disciplines of physics, chemistry, biology, and geology.

Considering the above, this study aims to explore the impact of AI tools, particularly chatbots, on science education, focusing on the uses proposed for them in the literature. Additionally, the review seeks to analyze the benefits and limitations of their use, in order to determine whether they are consistently identified across studies or merely highlighted by a few authors. Furthermore, the analysis addresses the strengths and weaknesses of the included studies, to determine the degree of consolidation of this field of research and to evaluate the extent to which the reported findings can be generalized. In summary, the review synthesizes current research on chatbots in science education, mapping what is known and highlighting the gaps that must be addressed before these tools can be deployed with confidence at scale.

Literature Review

Artificial Intelligence in Education (AIED)

The definition of AI is a subject of some controversy (Sheikh et al., 2023). When applying AI to education, Holmes et al. (2022) define the term AI&ED (Artificial Intelligence and Education) as a generic term that encompasses all possible relationships between education and Artificial Intelligence. They further divide AI&ED based on its educational purpose: a) preparing for AI (a form of AI literacy, focused on how to use AI) (Laupichler et al., 2022), also known as the human dimension within AI literacy (Holmes et al., 2022); b) learning about AI (theoretical knowledge or curriculum content focused on what AI is); c) learning with AI. Although the last two categories may appear similar, they represent distinct dimensions (Holmes, 2023; Holmes et al., 2022). Learning about AI refers to developing students' and teachers' knowledge and skills regarding how AI systems function (i.e., understanding how to use them and their underlying principles). This is also known as the technological dimension within AI literacy. Learning with AI refers to employing AI-driven tools in teaching and learning processes. In this case, the tools serve as a means rather than an educational end in themselves.

The focus of this study is on Learning with AI. This term encompasses all its uses in the educational process, such as supporting: a) administrative systems (enrollment management, scholarships, schedules, etc.) (Holmes et al., 2022); b) teachers (content creation, assessments, class control, or workload reduction) (Chang et al., 2023; Cooper, 2023); c) students (instant and individualized instruction, 24/7, etc.) (Lin & Ye, 2023).

Although numerous benefits of AI in education have been highlighted, such as its potential to transform educational practices or reduce teachers' workload (Bitzenbauer, 2023; Deveci-Topal et al., 2021), there is also a more pessimistic view that predicts certain dangers in its implementation, such as the risk that AI may reduce teachers to mere student monitors (Holmes & Tuomi, 2022). However, many authors remain skeptical of this claim, arguing that without the guidance of the teacher, students are unable to achieve the benefits of learning (Chiu et al., 2023; Humphry & Fuller, 2023; Labadze et al., 2023).

In this regard, although many AI tools are presented as resources that guide students (Chang et al., 2023; Kılınç, 2023), and there might be a temptation to use them without teachers' guidance, the results could primarily benefit students with a high level of prior knowledge and harm beginners or those with intermediate knowledge levels, who greatly benefit from guided instruction (Chiu et al.,

2023; Dahlkemper et al., 2023; De Bruyckere et al., 2015). The reality is that AI applications lack the depth and capacity of a human being (Holmes & Tuomi, 2022; Labadze et al., 2023), especially in higher-order skills such as metacognition (Bitzenbauer, 2023; Dahlkemper et al., 2023; Wang, 2023).

Due to the novelty of AI, there are currently exaggerations about its uses and capabilities (Holmes et al., 2022). In this regard, like all technological advances, it can pose risks and challenges (Miao et al., 2021). Nevertheless, some AI tools have already achieved significant milestones, and their use has expanded to ordinary people, such as Automatic Writing Evaluation, which provides feedback on written content submitted to the system; chatbots, which automatically answer human questions; dialogue-based tutoring systems, which engage students in conversations about specific topics; machine learning, which uses algorithms to analyze big data, identify patterns, and make inferences to learn (Holmes et al., 2022). Among these tools, chatbots occupy a unique position: they can engage learners in natural-language dialogue, give just-in-time explanations, and, in principle, emulate aspects of one-to-one tutoring. For that reason, the remainder of this review narrows its scope to empirical studies on chatbots in science education, examining both their reported benefits and their current limitations.

Chatbots in Education

One of the AI tools that is having a significant impact on education is chatbots (Cooper, 2023). Chatbots are designed to automatically respond to messages through the processing of natural language (NLP), from vast amounts of data (Holmes & Tuomi, 2022), to build models based on statistical inferences (Gregorcic & Pendrill, 2023). And although there have been examples of these tools since the mid-twentieth century (Dahlkemper et al., 2023), it is with the appearance of ChatGPT that their usage has become widespread. It is free, easy to use, and provides high-quality responses, even in highly specialized fields (Dahlkemper et al., 2023), allowing for meaningful conversations (Gregorcic & Pendrill, 2023).

The use of chatbots has increasingly expanded in educational contexts to support teachers (Bitzenbauer, 2023; Li et al., 2024), monitor the teaching–learning process (Chang et al., 2023; Lin & Ye, 2023), or create educational materials (Almasri, 2024; Park & Martin, 2024), saving teachers time (Labadze et al., 2023). It has also shown effectiveness in supporting students by providing instant and continuous guidance (24/7) (Alneyadi & Wardat, 2023; Vasconcelos & dos Santos, 2023), motivating students (Lee et al., 2023; Lin & Ye, 2023), or developing diverse skills (Labadze et al., 2023), among others.

Despite these affordances, several challenges and limitations have also been pointed out regarding the use of chatbots in educational settings. For example, some authors have argued that the learning benefits attributed to chatbots may be due to the "novelty effect," with students' interest fading over time (Kuhail et al., 2023). Another issue commonly associated with the use of chatbots is teachers' fear that students might attempt to deceive them by presenting outputs from chatbots as their own work (Wang, 2023). Some authors have also pointed out ethical issues related to the use of AI tools (Almasri, 2024). For example, Avraamidou (2024) argues that AI can dehumanize learning by standardizing it and promoting instant learning rather than a more reflective approach based on social relationships.

In the specific case of ChatGPT, issues have also been identified. ChatGPT can generate a wide variety of responses due to the data it has been trained on (worldwide internet data up to 2021), which can include all sorts of errors and inaccuracies. GPT-3.5 sometimes provides false information as if it were true (Dahlkemper et al., 2023), creating an illusion of understanding (Park & Martin, 2024). For instance, when asked physics questions, it does not acknowledge its limitations in knowledge (Gregorcic & Pendrill, 2023), and answers with great confidence. For these and other reasons, several authors emphasize the importance of providing specific training for both teachers and students in order to respond effectively to the challenges associated with these tools (Almasri, 2024; Labadze et al., 2023; Park & Martin, 2024).

Considering both their advantages and limitations, and given that the emergence of ChatGPT has attracted most of the public's attention, it is important to note that a variety of other chatbot types exist. Therefore, it may be useful to provide a classification that accounts for these different types. In the review by Debets et al. (2025), which analyzed studies published up to October 2023 (excluding those based on ChatGPT due to the lack of consolidated empirical evidence at that time), a classification into four categories was proposed, based on a previous typology by Agarwal and Wadhwa (2020):

a) Rule-based chatbots (R-BC), which use predefined responses based on pattern matching or rigid dialog structures, such as those created using AIML technology.
b) AI-based chatbots (AI-BC), which use machine learning techniques, language models, neural networks, or reinforcement learning, among others. They analyze large amounts of data, identifying patterns to generate responses. One commonly used architecture is the transformer.
c) Hybrid chatbots (H-BC), which combine techniques from the two previous categories (AI or rule-based).

d) Platform-based chatbots (P-BC), which also use one of the two previous technologies but are developed through creation platforms or tools such as DialogFlow. When the technology used was not clearly specified by the authors, Debets et al. (2025) classified them into this category. These were the most common in their review (59%), possibly because they required less programming knowledge, making them accessible to teachers without a technical background (Kuhail et al., 2023).

Whichever type of chatbot is used for educational purposes, it should be adapted to the specific context in which it is to be applied (Almasri, 2024), since the lack of a theoretical basis in selecting the type of chatbot may limit the learning outcomes achieved (Debets et al., 2025). Banihashem and Macfadyen (2021) argued that designing educational tools with a solid theoretical foundation could have several important functions, such as supporting the selection and development of the tool, guiding its use, interpreting the resulting learning outcomes, and preventing technological determinism.

In this regard, Zhang et al. (2024), in their review, suggest that chatbot-assisted learning can be grounded in up to eight theoretical frameworks, such as constructivist theories, situated/contextualized learning theories, cognitive theories of multimedia learning, self-regulated learning theories, output hypotheses, flow theory, collaborative learning theories, and motivation theories. In addition, other concepts that may support this type of learning include scaffolding, as chatbots could offer expert temporary support aligned with each learner's pace (Wood et al., 1976). Also, the cognitive load theory (Sweller et al., 1998), already named in Debets' (2025) review, offers a useful framework to understand how chatbots might reduce the burden on working memory, for example, through step-by-step clarifications or worked examples in areas where students face the greatest difficulties (Paas & Van Merriënboer, 1994). We may also refer to Vygotsky's concept of the zone of proximal development (ZPD), as chatbots have the potential to support learning by identifying what students already know and offering appropriately challenging content just beyond their current level of understanding. In their review, Debets et al. (2025) found that although some chatbot interventions were effective even without a theoretical foundation, all those that were ineffective lacked a solid theoretical basis. This suggests that grounding educational chatbots in theory can significantly improve their design and implementation.

Chatbots Applied to Science Education

Although the benefits of implementing technology in science education had already been highlighted by other authors (Cajas, 2001; Deveci-Topal et al., 2021; Linn, 2003), previous reviews showed that the use of chatbots was not as widespread in scientific disciplines as in language learning (Hwang & Chang, 2021; Zhang et al., 2024) or computer science education (Kuhail et al., 2023). Some authors argued that there were still few studies specifically focused on chatbots in science education or STEM contexts (Cooper, 2023), and even fewer that included empirical findings (Deveci-Topal et al., 2021). This limited implementation of chatbots in science education may have been related to some challenges. Hwang and Chang (2023) suggested that it could be more difficult to integrate chatbots into subjects that required strong skills in calculation or problem-solving, such as mathematics and science. Other authors have also pointed out the potential issues chatbots face when dealing with complex concepts (Gregorcic & Pendrill, 2023; Sperling & Lincoln, 2024; Wang, 2023).

For example, the review by Park and Martin (2024), which covered studies published between January and September 2023 on ChatGPT and science education (earth sciences, physics, chemistry, and biology), found that ChatGPT performed poorly in mathematics tasks and in the understanding of scientific concepts, standing out mainly in writing-related activities. However, they believed this limitation could become an opportunity, enabling students to learn from ChatGPT's mistakes and thus develop critical thinking skills. Their review identified 14 articles in which ChatGPT had been applied mostly in higher education (64%), in case study designs (71%), and mainly in physics (50%) and chemistry (36%).

Similar results were reported by Almasri (2024), who conducted a review covering studies published from 2014 to November 2023, using search terms related to AI and science education (also including computer science, mathematics, and engineering). He found 74 relevant articles, most of which were focused on the university level (48%), and on subjects such as general science (20%) and physics (14%). Almasri (2024) concluded that although AI tools could support the understanding of complex scientific topics, the development of problem-solving skills, and foster motivation and interest in science, these tools also showed limitations in dealing with some content areas and adapting to specific educational contexts.

Moreover, while some of the limitations of incorporating chatbots into science education have already been noted, recent authors have identified a potential shift since 2021, with a growing number of chatbots being developed for STEM disciplines (Debets et al., 2025; Hwang and Chang 2023; Park & Martin, 2024). This trend may have been accelerated by the disruptive emergence of ChatGPT, which has significantly influenced current educational practices. However, this impact is not yet fully captured by prior reviews, as many of them concluded their searches by late 2023 or, as in the case of the review by Debets et al. (2025),

pointed to ChatGPT merely as a potential direction for future research.

Chatbots offer clear potential to address some of the unique challenges in science education, such as personalized tutoring (Kılınç, 2023) or support in improving subject comprehension and increasing student motivation (Almasri, 2024). These tools can reduce cognitive load by providing scaffolding, rephrasing, or revisiting misunderstood content, adapting to individual learning paces, and even engaging in Socratic dialogue (Gregorcic & Pendrill, 2023). Furthermore, some of the numerical inaccuracies observed in earlier versions of chatbots (Ding et al., 2023; Sperling & Lincoln, 2024) appear to be improving with newer models (Bewersdorff et al., 2023; Ramkorun, 2024; Tassoti, 2024). Taken together, these aspects suggest that chatbots may be a promising tool for science education.

Additionally, some authors have also examined the use of chatbots within established theoretical frameworks. For example, Peikos and Stavrou (2025) showed that Shulman's (1986) framework of Pedagogical Content Knowledge (PCK) and its subsequent developments can serve as a basis for analyzing how teachers guide, evaluate, and refine ChatGPT outputs in science lesson planning, always under the teacher's critical supervision. Feldman-Maggor et al. (2025) employed the technological PCK (TPACK) framework to analyze how technological, pedagogical, and disciplinary knowledge interact when integrating generative AI in science classrooms. For these authors, responsible use of ChatGPT in chemistry education is essential, both in the formulation of prompts and in the evaluation of chatbot responses, where teachers' disciplinary and pedagogical expertise remains indispensable. At the same time, they argue that new teacher competencies are required when working with chatbots such as ChatGPT, since the traditional TPACK model may not be sufficient for their effective implementation in AI-based educational tools. Both authors consider it essential to attend to these educational frameworks when interacting with chatbots to achieve effective educational outcomes.

Furthermore, some authors think that technological advances like AI are unstoppable, and there is no choice but to implement them in the classroom (Chang et al., 2023). Considering the close relationship between Science and Technology and the overwhelming irruption of AI in the educational sphere, it could be interesting to provide an overview of the existing literature. This review could help inform educational practice and policy, as well as identify areas that require further research or intervention.

Therefore, it is necessary to pose the following research questions (RQ):

RQ1: What is the educational use of chatbots in science education?

RQ2: What educational benefits and limitations are reported when chatbots are used in science education?

RQ3: What methodological strengths and weaknesses do the included studies exhibit?

Method

A Scoping Review was conducted following the methodological principles of PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2020 (Page et al., 2021). The scoping review is particularly valuable when the topic has not been extensively reviewed before, as in the present case. Scoping reviews "aim to provide an overview or map of the evidence" (Munn et al., 2018, p. 3).

Search Strategies

To conduct the search, the following terms were used, separated by Boolean operators AND and OR: ("Science Education" OR "Science Teaching" OR "Science Instruction" OR "Physics teaching" OR "Chemistry teaching" OR "Biology teaching" OR "Geology teaching" OR "Physics education" OR "Chemistry education" OR "Biology education" OR "Geology education" OR "STEM" OR "STEAM") AND ("Chatbot" OR "ChatGPT"). The same Boolean string was used in all databases.

To search for articles in the field of science education, it was decided to include search terms related to the four subjects that commonly constitute this field: physics, chemistry, biology, and geology. To further broaden the search range, it was also decided to include STEM and STEAM.

To select eligible articles, the following admission criteria were also established: 1) all selected articles had to be peer-reviewed but not systematic reviews; 2) they had to be written in English or Spanish; 3) they must have proposed some use for chatbots applied to science education; 4) they had to be retrieved from the Web of Science, Scopus, or ERIC databases without an initial time limit, and up to January 1, 2025. The article screening process, as well as the articles excluded for each of these reasons, is documented in Fig. 1.

Studies Included in the Scoping Review

A total of 608 articles were found, with 190 in the Web of Science database, 256 in ERIC, and 162 in Scopus. In a first step, an automatic screening of duplicates was carried out using Zotero software, followed by a human review using Microsoft Excel as a management tool. Figure 1 outlines and summarizes the entire process. Two researchers were responsible for screening all articles. Inter-reviewer agreement on study inclusion was assessed using Cohen's

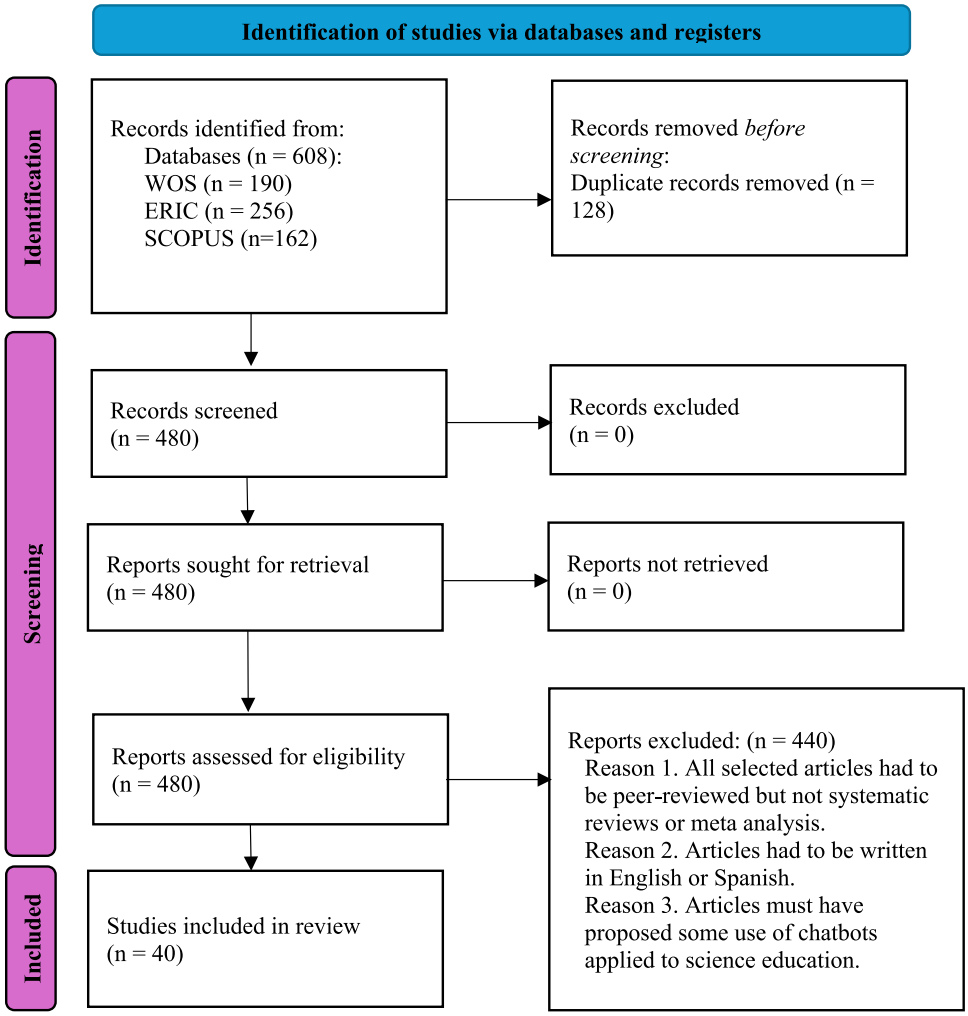

Fig. 1 Flowchart. Source: PRISMA

kappa coefficient, which yielded a value of 0.91 ($p < 0.001$), indicating a high level of agreement. Discrepancies were resolved through discussion with a third researcher, who confirmed the final selection of studies, which resulted in full agreement among all reviewers. In total, 40 papers were included in the final analysis.

In our review, in order to retrieve the highest possible number of articles related to science education and chatbots, the keywords used for the search were intentionally broad. This approach allowed us to include a wide range of studies, which were later screened manually to exclude those that did not meet the inclusion criteria. For example, the search term "science education" yielded many articles from fields such as "computer science education" or "healthcare," which were not eligible for inclusion in this review. This may explain why many articles were left out. For instance, in the review by Debets et al. (2025), most of the articles found (25%) belonged to these two disciplines.

Finally, since its release, ChatGPT has gained massive popularity (Labadze et al., 2023), thus it would have been reasonable to expect a larger number of chatbot-related studies to appear in our review as well. However, many of these publications turned out to be dissertations or papers evaluating ChatGPT's performance in specific disciplines, such as passing exams (Fergus et al., 2023; Revalde et al., 2025).

Literature Characterization and Data Coding

In order to address the first RQ, the selected papers were first characterized: an ad hoc table was created with the following data extracted from the included studies: (1) reference (authors and year of publication); (2) study design (exploratory [E]; intervention [I]; or case studies [C]); (3) country in which the study data were collected; (4) subject (science subject or STEM); (5) type of chatbot (own (a custom chatbot based on rule-based systems or one built on a large language model (LLM) different from ChatGPT); GPT-3.0 (ChatGPT); GPT-3.5; GPT-3.5 Turbo; GPT-4; or when the authors of the article did not specify the exact version used in their study, it was coded as GPT without version); (6) educational stage (primary; secondary; university); (7) output

variable (the variable reported by the authors in their studies, such as teachers' or students' opinions, students' academic performance, among others).

Secondly, to complete the response to RQ1, researchers determined the target group (teachers, preservice teachers, or students), their contributions to the science education field, and the main result of their application. A tree diagram was created for this purpose.

Finally, to address the second RQ, the main benefits and limitations of using chatbots in science education, as reported in the results of the included studies, were compiled. This process followed an inductive approach, as no predefined categories were established (Bryman, 2016). Two researchers independently conducted an in-depth reading of all the papers and identified the most representative categories through an iterative process (Biasutti, 2015). This inductive approach was based on the principles of significance and recurrence (Cebrián et al., 2022). Subsequently, the categories were cross-checked, reaching agreements for their standardization, and those appearing in at least three articles were selected. Following Ojala's (2021) recommendation, selected quotes are presented to illustrate the meaning and content of each category. For instance, one category was "students' motivation," exemplified by the following quotes: a) "…, our AI system aims to enhance student engagement and motivation, fostering a positive learning environment …" (Bewersdorff et al., 2023, p. 8); b) "…, which tends to increase students' learning motivation." (Liang et al., 2023, p. 14); c) "…, chatbots are an effective way to motivate students to learn, …" (Lin & Ye, 2023, p. 279).

Another emergent category was "reduction of teachers' workload," illustrated by a) "The GPA serves as an assistant to help instructors organize the entire learning process, thereby reducing their burden and pressure" (Wei et al., 2024); b) "The findings emphasized the benefits of ChatGPT in developing an implementable course plan, delivering adaptable information, and time-saving" (Okulu & Muslu, 2024, p. 7450); c) "AI can eliminate the need to write basic physics problems from scratch, allowing instructors to shift their time and attention to more creative endeavors and student support" (Sperling & Lincoln, 2024, p. 314).

After this inductive process, the categories determined were a) benefits: (1) potential for improving learning, (2) reduction of teachers' workload, (3) students' motivation, (4) high linguistic quality in writing, (5) personal tutor, (6) 24/7 support, (7) support for teachers in one-on-one tutoring, (8) learning monitoring; b) limitations: (1) general and even erroneous information, (2) require critical thinking, (3) not replace human teachers, (4) answers of GPT need refinement, (5) requires topic-specific knowledge, (6) issues with information sources/authorship rights, (7) technical issues, (8) ongoing updating of teachers, (9) issues with numerical analysis.

Analysis of Strengths and Weaknesses of the Included Studies

To complement this review and to address the third RQ, an assessment of the quality of the included studies was conducted through an analysis of their main strengths and weaknesses. The process was inspired by common procedures used in systematic reviews within the field of Health Sciences, where the risk of bias is typically assessed using standardized tools such as the "*Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies"* from the United States National Institutes of Health (NIH, 2022). These tools generally evaluate the quality of the studies across various domains and categorize them as "good" if most criteria are met and the risk of bias is low, "fair" if some criteria are met and the risk of bias is moderate, or "poor" if few criteria are met and the risk of bias is high.

Given that this review is situated within the field of education, many of the categories included in the NIH tool were not directly applicable. Nevertheless, a simplified quality assessment was conducted to provide readers with an overview of the methodological rigor of the included studies. In the case of exploratory studies, the evaluation criteria were adjusted to account for the specific dynamics of chatbot-based interactions. For instance, ChatGPT (and similar AI-based chatbots) may not produce identical responses to the same prompt, as noted by Dahlkemper et al. (2023).

For studies that involved interventions or case studies, the following aspects were considered: random sample, presence of a control group, sample characterization, pre- and post-test design, sample size, duration of the intervention, and the source of outcome measures. For exploratory studies, the variables identified were: definition of objective or research question; the research questions were investigated through authors' opinions, experts, or rubrics/evaluation standards; total/partial answers were offered to the conversation with the chatbot; if several attempts were carried out on the same prompt, and if they offered a percentage of success/error, and if the studies clearly acknowledged their methodological limitations.

Results

General Aspects of the Research (RQ1)

Table 1 summarizes the main general aspects of the papers analyzed. Regarding the country of origin, there is no clear trend, with a diversity of nationalities. However, many articles have been published in China and the USA. As can also be observed in Fig. 2, concerning subjects, physics is the most analyzed. In terms of educational stage, secondary education is the most researched, and ChatGPT is the most

Table 1 General aspects of the research included in the review

ID	Study	Design[1]	Country	Subject	Type of chatbot	Educational stage	Output variable[2]
1	Cooper, 2023	E	Australia	Sciences	GPT	N.S.[3]	Op author
2	Wang, 2023	E	Australia	Microbiology	GPT-3.5	University	Op author
3	Deveci-Topal et al., 2021	I	Turkey	Physics	DialogFlow (Google)	Secondary	Academic performance and op students
4	Dahlkemper et al., 2023	C	Germany	Physics	GPT-3.5	University	Linguistic and scientific quality scores
5	Gregorcic & Pendrill, 2023	E	Sweden	Physics	GPT	N.S	Op authors
6	Chang et al., 2023	C	Korea	Sciences	Inquirybot (RBC[5] and NLP[6])	Primary	Op teachers and students and usage statics
7	Humphry & Fuller, 2023	E	USA	Chemistry	GPT	N.S	Op authors
8	Lin & Ye, 2023	I	Taiwan	Biology	Own (LINE)	Secondary	Academic performance
9	Bitzenbauer, 2023	C	Germany	Physics	GPT-3	Secondary	Op author and students
10	Liang et al., 2023	E	China	Physics	GPT-3	N.S	Op authors
11	Kılınç, 2023	E	Turkey	Chemistry	GPT-4	N.S	Op author
12	Bewersdorff et al., 2023	E	Germany	Biology	GPT-3.5–4	Secondary	Op authors
13	Nguyen, 2023	I	USA	Sciences	Kibot (RBC[5] and NLP[6])	Secondary	Students' systems thinking
14	Ding et al., 2023	C	USA	Physics	GPT-3	University	ChatGPT's accuracy and students' trust in Chat-GPT
15	Li et al., 2024	I	China	STEM	GPT	University (p.s.[4])	Critical thinking, academic performance and cognitive load
16	Küchemann et al., 2023	I	Germany	Physics	GPT-3.5	University (p.s.[4])	Linguistic and scientific quality scores
17	Alneyadi & Wardat, 2023	I	UAE	Physics	GPT	Secondary	Academic performance and op students
18	Lee et al., 2023	I	Korea	Physics	Danbee.Ai (RBC[5])	Primary	Academic performance and op students
19	Vasconcelos & dos Santos, 2023	E	Brazil	STEM	GPT-4	Secondary	Op authors
20	Latif & Zhai, 2023	E	Greece	Sciences	GPT-3.5 Turbo	Secondary	BERT and ChatGPT's accuracy
21	Alan & Yurt, 2024	E	Turkey	Physics	GPT-3.5	Secondary	Op author
22	Chen & Chang, 2024	I	Taiwan	Physics	GPT-3.5 Turbo	Secondary	Academic performance and op students
23	Cheung et al., 2024	I	China	Sciences	ChatGPT	Secondary	Academic performance and op students
24	Duy et al., 2024	I	Vietnam	Physics	Messnow (RBC[5])	Secondary	Academic performance and op students
25	Güldal & Dinçer, 2024	C	Turkey	Physics	Hermes (RBC[5])	University	Op students
26	Guo & Lee, 2023	C	USA	Chemistry	GPT	University	Academic performance and op students
27	Huang et al., 2024	I	China	Chemistry	Sider (GPT-3.5 Turbo)	University (p.s.[4])	Academic performance and op students
28	Ng et al., 2024	I	China	Physics	Nemobot (RBC[5]) and SRLbot (ChatGPT)	Secondary	Academic performance and op students
29	Ramkorun, 2024	E	USA	Physics	GPT-3.5	University	Op author
30	Ruff et al., 2024	C	USA	Chemistry	GPT-3.5	University	Critical thinking and op students
31	Sperling & Lincoln, 2024	E	USA	Physics	GPT (4.0), Almanack and Flint (LLMs)	Secondary	Op author
32	Tassoti, 2024	C	Austria	Chemistry	GPT-3.5	University	Op students
33	Uğraş & Uğraş, 2024	C	Turkey	STEM	GPT	University (p.s.[4])	Op teachers

Table 1 (continued)

ID	Study	Design[1]	Country	Subject	Type of chatbot	Educational stage	Output variable[2]
34	Wan & Chen, 2024	C	USA	Physics	GPT-3.5 turbo	University	Op students and experts
35	Yin et al., 2024a	I	China	Biology	RBC[5] (Flow.ai)	University	Op students
36	Yin et al., 2024b	I	China	Biology	RBC[5] (Flow.ai)	University	Academic performance and op students
37	Okulu & Muslu, 2024	E	Turkey	Sciences	GPT-3.5	University (p.s.[4])	Op authors
38	Lee & Zhai, 2024	C	Korea	STEM	ChatGPT	University (p.s.[4])	Op teachers
39	Wei et al., 2024	I	China	Sciences	Dong Dong (GPA[7])	Secondary	Academic performance and op students
40	Archila et al., 2024	I	Colombia	Biology	ChatGPT	University	Critical thinking and op students

1 *E*, exploratory; *I*, intervention; *C*, case study. **2** *Op*, opinion. **3** No sample. **4** Preservice teachers. **5** Rule-based chatbot. **6** Natural language processing. **7** Generative artificial intelligence pedagogical agent

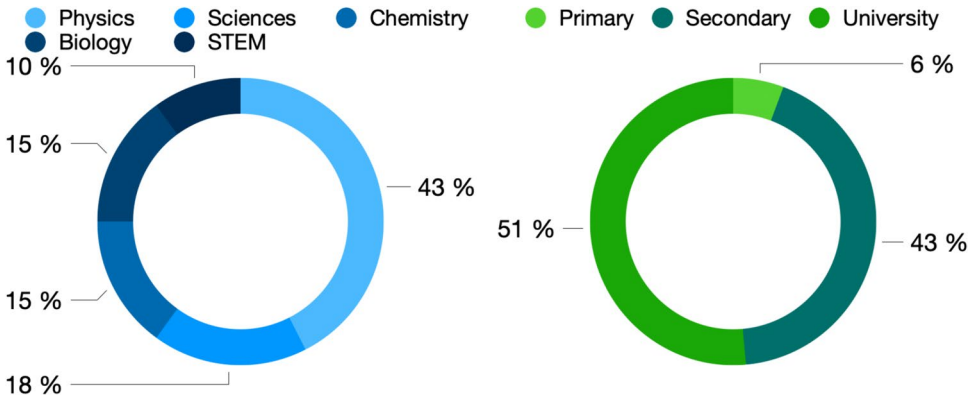

Fig. 2 Distribution of articles by subject (1a) and education stage (1b)

used chatbot (30/40) in the reviewed articles. Moreover, most articles are, as expected, very recent and have been published in 2023 and 2024. Finally, many of the articles found are intervention articles (16/40), followed by exploratory studies (13/40), and case studies (11/40).

Uses of Chatbots (RQ1)

Most research investigates the chatbots' use by students, although in some cases, authors suggest various uses within the same article. Moreover, the majority use of chatbots is as a tutor. The reported results by the authors vary, with some being useful and others disappointing, or considering that chatbots, up to now, are not effective tools.

Figure 3 shows the main uses of chatbots according to the target group, and the number of studies in each group is indicated in square brackets. The bullet points are the main highlights from study results when using the chatbot for educational purposes (the number of the study is identified in brackets).

As can be seen in Fig. 3, although many authors point out challenges in using chatbots for educational purposes, several recognize their potential to enhance education, especially in three key areas: creating resources for teachers, tutoring or conducting activities for students, and learning from mistakes.

Benefits and Limitations of Chatbots in Education (RQ2)

As can be observed in Fig. 4 (Appendix 1 contains a detailed table with all the categories identified), authors mostly agree that chatbots have the potential to enhance learning (35/40), since, for example, students can ask them to solve specific doubts, and provide contextualized examples, etc. This fact is related to their capacity as a personal tutor (24/40). Another important benefit is their ability to reduce teachers' workload (15/40), as they can assist in creating materials such as rubrics, teaching units, or assessments.

One of the most significant limitations identified is that many authors emphasize the importance of critical thinking (19/40) when using chatbots. This highlights that these tools may not be suitable for students who lack prior understanding of the subject. This observation aligns with findings indicating that ChatGPT may provide answers with incorrect or overly general information (20/40). All of this aligns with

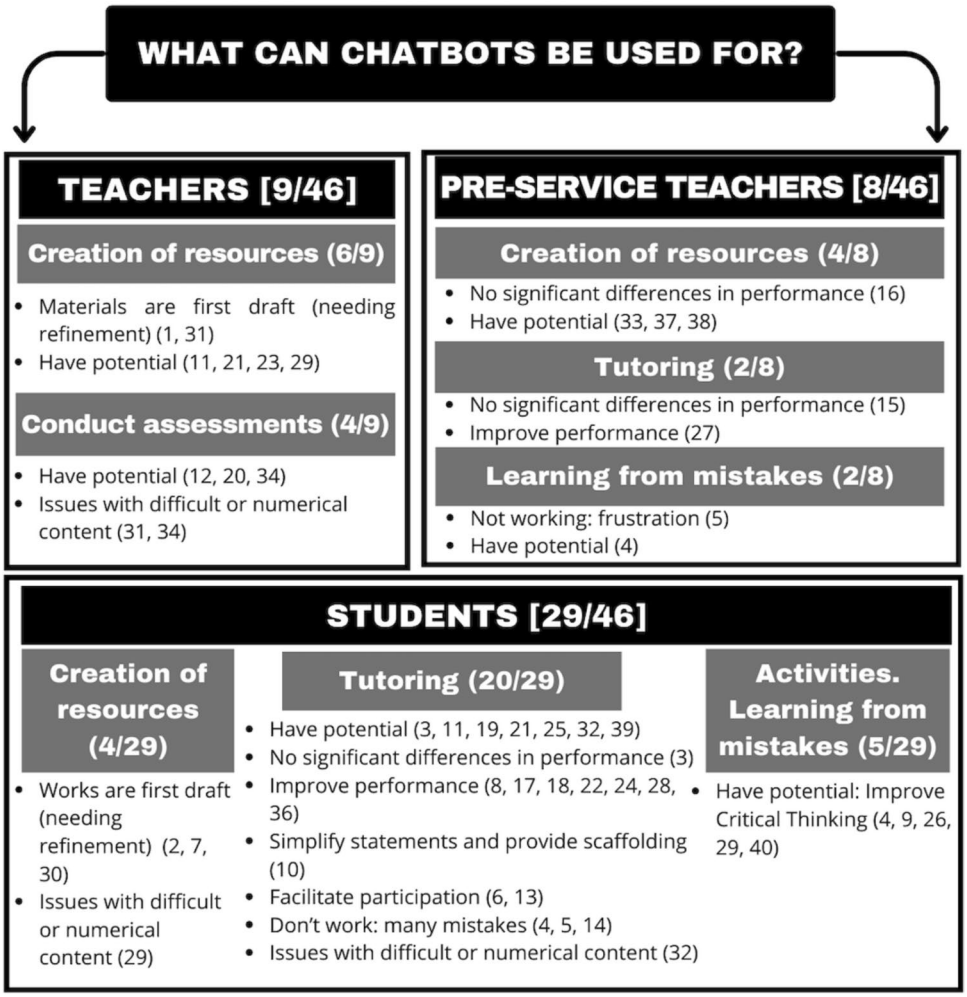

Fig. 3 Uses of chatbots according to user groups

the most frequently mentioned limitation in the articles: the fact that it does not replace human teachers (21/40).

Strengths and Weaknesses of the Included Studies (RQ3)

Tables 2 and 3 outline the characteristics of interventions and case studies, respectively. In general, the samples were small (only 6 out of 40 had a sample larger than 100 subjects), and the duration of interventions seems short. It is also noteworthy that many of these types of studies only gather opinions, not empirical data. Additionally, only three fully characterize the sample with data, such as whether they have support at home with science tasks or the education level of the parents.

Continuing with exploratory studies (Table 4), most of them use ChatGPT, relegating other chatbots to intervention or case studies. It is noteworthy that the majority evaluate responses through mere authors' opinions, and only three included experts for analysis. It is also noticeable that no study explicitly outlines criteria or standards in advance to evaluate ChatGPT's outputs. Only two studies make multiple attempts at the same questions to analyze the variability of ChatGPT answers.

Discussion

This review set out to examine the emerging role of chatbots in science education. It analyzed how chatbots are being used in educational contexts, what benefits and limitations are reported across studies, and what methodological strengths and weaknesses characterize the current body of research. In doing so, the study aimed to synthesize early empirical evidence, determine the degree of consolidation of this field, and identify gaps that must be addressed before these tools can be confidently implemented at scale.

Research involving the use of chatbots in educational practice for science education is very recent and predominantly employs ChatGPT. In this regard, its popularity among the public has also been noted by other authors such

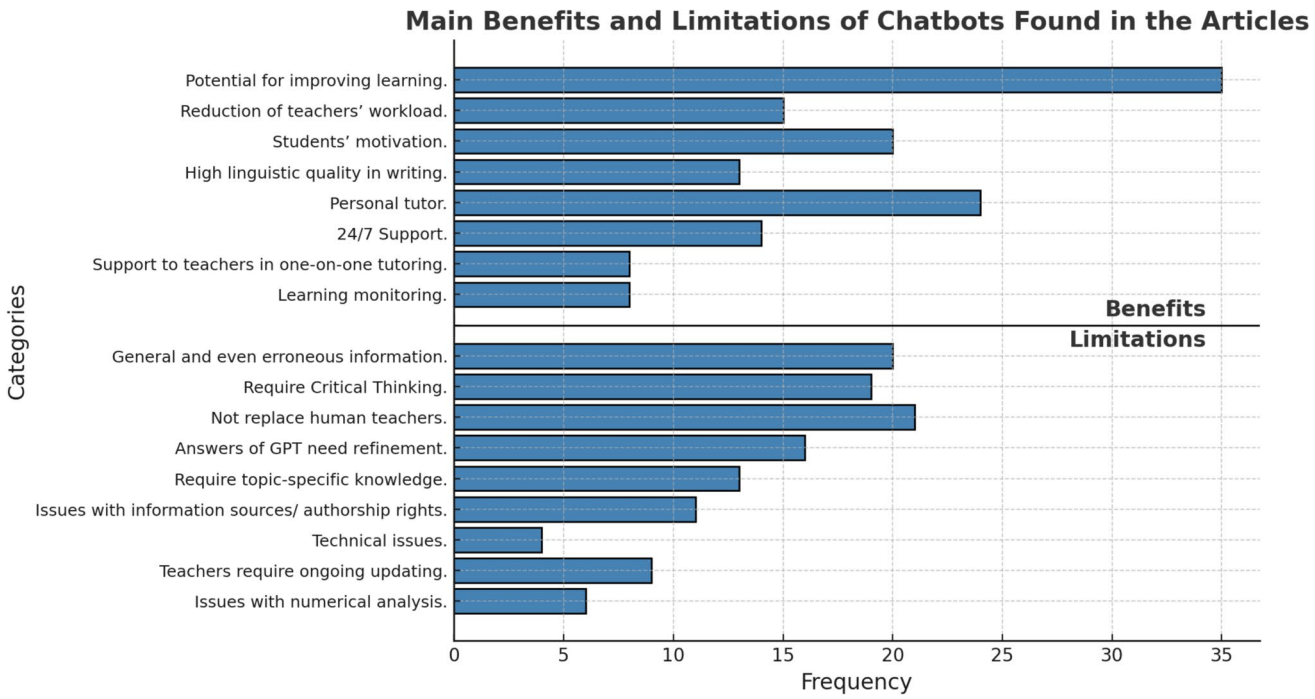

Fig. 4 Main benefits and limitations of chatbots identified in the reviewed articles

as Dahlkemper et al. (2023), Kılınç (2023), and Liang et al. (2023).

Perhaps the reasons for this growth are motivated by some of ChatGPT's features: (1) the free availability of some of its versions (Humphry & Fuller, 2023); in fact, in this study, the majority of the investigations used the free version; (2) its accessibility (Liang et al., 2023), easy installation, and customization, even for non-expert users; (3) ease of use, very similar to other widely used instant messaging applications. However, the fact that information input is in a written chat format poses a limitation for some authors (Wang, 2023). Nevertheless, the incorporation of voice facilities in the latest versions could address these issues. Additionally, (4) it can be used in a broad range of fields, even for daily life (Chang et al., 2023; Dahlkemper et al., 2023), increasing its usefulness.

It is striking that a considerable amount of research has investigated the use of chatbots in physics (Fig. 2), which is consistent with the findings reported in the reviews by Debets et al. (2025) and Park and Martin (2024). This could be attributed to the fact that students often perceive physics as the most challenging subject during their high school years, due to the prevalence of abstract concepts and the complex numerical calculations that hinder their learning (Ekici, 2016; Erinosho, 2013). The increased attention directed towards Physics may stem from educators' and researchers' desire to make the subject more engaging and accessible for students through the integration of chatbots.

However, the results obtained reveal a controversy surrounding the effectiveness of chatbots in physics education, particularly in explaining complex phenomena and supporting numerical calculations. On the one hand, Alneyadi and Wardat (2023) and Chen and Chang (2024) found that using chatbots as tutors leads to improved academic performance among students. Similarly, Vasconcelos and dos Santos (2023) consider ChatGPT-4 to be an effective tutor capable of elucidating difficult physics concepts. On the other hand, in Küchemann et al. (2023)'s experience, students who did not use ChatGPT achieved higher-quality results. Moreover, Gregorcic and Pendrill (2023) deem ChatGPT to be inadequate as a tutor for problem-solving. Ding et al. (2023) and Ramkorun (2024) found that chatbots may encounter challenges with the scientific accuracy of their responses, which could lead to conceptual errors. Furthermore, Liang et al. (2023) and Sperling and Lincoln (2024) confirm that GPT sometimes struggles with complex arithmetic operations.

Similar difficulties have also been encountered in other disciplines, such as Chemistry. Kılınç (2023) identifies numerical issues with ChatGPT, while Humphry and Fuller (2023) question its efficacy in this regard, and Tassoti (2024) highlights the need for a critical evaluation of chemistry-related answers generated by ChatGPT. Despite the predominant use of ChatGPT due to its advantages, criticism regarding the quality of its responses on specific scientific content persists (Cooper, 2023; Wei et al., 2024).

Table 2 Parameters of interventions

ID of study	3	8	13	15	16	17	18	22	23	24	27	28	35	36	39	40
1. Random sample			X	X	X	X				X			X	X	X	
2. Has a control group	X	X	X	X	X	X	X	X		X		X	X	X	X	
3. Characterizing the sample in advance (socio-economic level, mother's education level, etc.)	X		X							X						
4. Carry out pre/post-test	X	X	X	X		X	X	X	X		X	X	X	X	X	X
5. Sample size[1]	41 (20c)	34 (17c)	172 (58c)	23 (12c)	26 (13c)	122 (64c)	192 (111c)	202 (70c)	117	100 (50c)	59	74 (36c)	62 (31c)	62 (31c)	60 (30c)	55
6. Intervention length[2]	26 h	2 w	8 s	8×3 h	1 s	12 w	7 s	2 w	6 s×40 min	90 min	4 s×90 min	10 s	N.S	1 s	14 w×120 min	16 w×160 min
7. Type of outcome[3]	Rto; Op	Rto	Th	Rto; Op; Th	Rto	Rto; Op	Rto; Op	Rto; Op	Rto	Rto; Op	Rto; Op	Rto; Op	Op	Rto; Op	Rto; Op	Op; Th

1 *c*, Sample amount in control group; **2** *H*, hours; *w*, weeks; *s*, sessions; *min*, minutes. **3** *Rto*, academic performance; *op*, opinions; *Th*, students' systems thinking

Table 3 Parameters of case studies

ID of study	4	6	9	14	25	26	30	32	33	34	38
1. Random sample											
2. Has a control group											
3. Characterizing the sample in advance (socioeconomic level, mother's education level, etc.)						X			X		
4. Carry out pre/post-test			X			X		X			
5. Sample size	102	18	53	40	22	29	13	23	43	85	29
6. Intervention length[1]	20 min	3×40 min	90 min	16 w	3 w	1 s	n.n	8 s×90 min	n.n	n.n	4 s×2 h
7. Type of outcome[2]	Op	Op	Op	Op	Op	Op; Th	Op	Op	Op	Op	Op

n.n., It is not necessary. **1** *H*, hours; *w*, weeks; *s*, sessions; *min*, minutes. **2** *Rto*, academic performance; *Op*, opinions; *Th*, students' systems thinking

A possible explanation for these disparities may lie in the different language model versions employed in the studies. Newer models have greatly improved their ability to solve numerical problems and perform advanced reasoning, even in free versions. This point was also raised by Vasconcelos and dos Santos (2023), who suggested that several of the limitations identified could be resolved with the newer versions of ChatGPT.

Nevertheless, it is also relevant to consider the context from which ChatGPT is evaluated. Some studies analyze its potential as a tutor through tasks designed by teachers or expert researchers, with high expectations regarding the accuracy and depth of the responses generated. In contrast, other works observe its direct use by students in real learning contexts, where expectations may differ. Moreover, prompts developed by specialists tend to be more complex and structured than those an average student might formulate, which can influence the chatbot's performance and the perception of its usefulness. This methodological heterogeneity hinders direct comparison and underscores the need for research that clarifies when ChatGPT can effectively tutor in abstract disciplines like physics.

In contrast to physics or chemistry, where issues have been identified, it is thought-provoking to note the absence of studies concerning geology. This trend aligns with its historical underrepresentation within scientific research in the educational field (Saçkes, 2015), a concerning pattern that may hinder the development and implementation of innovative teaching strategies in this discipline.

Given these challenges of LLM-based chatbots such as ChatGPT, some researchers (Duy et al., 2024; Güldal & Dinçer, 2024) propose creating specific chatbots tailored with the content being taught in the classroom, such as rule-based chatbots or platform-based chatbots. This approach could improve the scientific quality of the responses and limit student dispersion (Chang et al., 2023), as ChatGPT can provide answers in a multitude of fields even if they are not directly related to the topic at hand. However, the authors have reported a decrease in the linguistic quality of responses when using these tailored chatbots (Chang et al., 2023; Ng et al., 2024). This outcome is not surprising, considering that ChatGPT has been "trained" with a vast amount of data, resulting in a remarkably high linguistic quality (Dahlkemper et al., 2023).

Moreover, this creates a kind of dilemma between linguistic quality and scientific quality depending on the type of chatbot. Therefore, it is necessary to carefully design and plan the learning objectives pursued with the chatbot (Liang et al., 2023). On the one hand, if high precision in scientific content is desired, and students should not be able to disperse or "play" by asking questions from other subjects, the best option is to design a specific chatbot, knowing that its linguistic capacity will be reduced. On the other hand, if the goal is for students to obtain preliminary versions of the content that they will later have to improve to enhance their scientific understanding personally, and if maximizing the options for exploring various topics and promoting research skills is desired, it is best to use a generalist chatbot.

It is also true that using a custom chatbot requires more work, and although some authors rely on other applications like Google's DialogFlow (Deveci-Topal et al., 2021), in general, they require certain programming knowledge. As a solution, Bewersdorff et al. (2023) and Latif and Zhai (2023) propose training a generalist chatbot, like GPT, with specific data depending on the intended use, thus achieving greater precision in its scientific responses. This approach could strike a balance between the ease of implementation and the desired level of accuracy in the chatbot's outputs, making it a more accessible option for educators who may not have extensive programming expertise.

In any case, when choosing the type of chatbot to use, it is also essential to consider the specific educational goal to be achieved. For example, one of the educational functions of chatbots is to serve as personalized tutors, which may lead to their use as tools for autonomous learning without teacher supervision. However, as noted in this review, teacher

Table 4 Parameters of exploratory studies

ID of study	1	2	5	7	10	11	12	19	20	21	30	32	38
1. Clearly define the objective/RQ of their study	X	-	-	-	X	X	X	X	-	X		X	X
2. Analyse the research questions through the author's opinions	X	X	X	X	X	X		X		X	X	X	X
3. Analyse research questions through expert opinions					X		X		X				
4. The rubrics or evaluation standards that have been followed in the analyses are provided							n.n						
5. Total and not partial responses to your conversation with the chatbot are offered	X	X	X	X	X	X	n.n	X			X		
6. Perform several tests of the same question and offer a % of success or error		-					X		X				
7. Show the limitations of the study					X		X	X	X				X

-, Only in some occasions; *n.n.*, It is not necessary

guidance is crucial for the appropriate use of chatbots by students. As highlighted by Debets et al. (2025), an educational chatbot based on Sweller et al. (1998)'s cognitive load theory could provide more effective support and feedback to students. Therefore, chatbots grounded in educational theories are more likely to succeed. However, as pointed out in the reviews by Heeg and Avraamidou (2023) and Debets et al. (2025), most studies involving chatbots lack a theoretical framework to support their educational proposals.

The observed benefits of chatbots in science education can be explained by well-established learning theories. Among others, the cognitive load theory (Sweller et al., 1998), chatbot responses can align with students' cognitive levels, providing structured and organized information through concise, scaffolded guidance to reduce extraneous cognitive load. This is critical in science fields where abstract content risks overwhelming working memory (Kirschner et al., 2006)—an issue that these tools could reduce or hopefully mitigate. Additionally, drawing on situated learning theory (Lave & Wenger, 1991), chatbots can narrow the gap between textbook concepts and lived experience by simulating authentic problem-solving contexts and anchoring interactions in real-world scenarios. This contextualization not only boosts engagement but is important for integrating scientific principles into durable cognitive schemata. Finally, Vygotsky's zone of proximal development (ZPD) further clarifies their potential: chatbots might function as dynamic scaffolds, offering calibrated support just beyond a learner's current competence, much like a human tutor would when guiding discovery. Although current LLM-based chatbots are not yet able to tailor their responses precisely to a learner's ZPD, the benefits reported in several studies may reflect partial forms of scaffolding, especially when chatbot interactions prompt students to reflect, pose questions, or receive elaborated feedback. From this perspective, the educational impact observed could be the result not only of access to information, but of a dialogue that approximates the cognitive support traditionally offered by human tutors. Thus, these theories not only help explain the results observed, but also suggest key principles to guide future chatbot design.

For chatbots to function as effective science learning tools, their design must be rooted in robust theoretical frameworks (Debets et al., 2025), addressing three core demands: the cognitive load inherent to complex concepts, the necessity of contextual anchoring for meaningful transfer, and adaptive scaffolding calibrated to learner progression. Crucially, then, developing or choosing educational chatbots transcends technical implementation; it is, above all, a pedagogical endeavor.

In a more general context, a common highlight of researchers is the ability of chatbots to reduce teachers' workloads. Following the American Federation of

Teachers (Johnson & Ricker, 2017), a stressful workload is a general tone among teachers. In fact, according to the same report, 61% of teachers find their job always or almost always stressful (twice as much as in other professions). They also complain about the lack of time to prepare their lessons and materials. Undoubtedly, the fact that chatbots may contribute to reducing this load, and the stress that it entails, is great news.

In this sense, Chang et al. (2023) claim that chatbots help by performing repetitive tasks, such as assessments and student monitoring, assisting in administrative tasks (Dahlkemper et al., 2023), and generating educational resources such as rubrics or teaching units (Alan & Yurt, 2024; Kılınç, 2023). Similarly, Wan and Chen (2024) highlight that even if ChatGPT's evaluation is only satisfactory in 70% of the cases, it can still save instructors considerable time and effort. So far, while ChatGPT has shown strong capabilities in generating educational resources in scientific disciplines, its ability to carry out educational assessments without supervision still falls short of an acceptable level of accuracy. However, further investigation is highly necessary to provide empirical data (not just opinions and perceptions) to determine whether these tools truly support teachers or, on the contrary, add to their workload and stress.

Chatbots can also contribute to the issue of high student-to-teacher ratios by addressing personalized student queries, and thus promoting a more individualized teaching (Deveci-Topal et al., 2021). Chatbots can offer significant support to students, providing instant tutoring, anywhere, anytime, and through a widely common means, such as mobile devices or smartphones (Güldal & Dinçer, 2024; Lin & Ye, 2023).

In general, students express positive opinions regarding the educational use of chatbots in science education, finding it motivating. In some cases, it seems that students feel more at ease asking questions than they do with a human teacher (Chang et al., 2023). This could lead to more effective learning, as it has been proven that motivation has a positive and reciprocal relationship with academic performance (Muijs & Reynolds, 2017).

For the proper use of chatbots, many authors also consider it essential to have developed higher-order skills, such as critical thinking (Cooper, 2023; Wei et al., 2024), and prior knowledge of specific content (Archila et al., 2024; Ruff et al., 2024). This may indicate that, although positive effects can be achieved with these tools, their maximum performance is reached when students already possess these skills, as is the case in university education. In any case, this critical thinking can only be developed with a deep prior knowledge (Willingham, 2019) and the corresponding subsequent analysis work (Deveci-Topal et al., 2021). This brings us back to the beginning: if students need good critical thinking to use AI, then students need education in content, culture, and values, along with guidance from someone who possesses all of these, such as an educator.

This is necessary because the most significant flaw detected by some authors is that program responses contain errors, overly general information, and contradictions, requiring scrutiny from an expert human filter (Ramkorun, 2024; Wan & Chen, 2024). However, the issue diminishes when chatbots are used to support teachers, as they are assumed to have sufficient knowledge and critical ability to evaluate and correct responses. In this regard, some authors highlight ChatGPT's potential to offer very good preliminary versions, even if they need correction (Cooper, 2023).

Alternatively, although the use of chatbots appears more promising as a tool to support teachers, most of the analyzed studies focus on how they can benefit students. However, a series of shortcomings have been detected in these studies: the few interventions conducted yield disparate results regarding academic performance improvement. While positive feedback from students is reported, research should involve larger samples, longer-duration interventions, and control groups, and should also measure learning outcomes rather than solely relying on participants' opinions, as is currently predominant. These limitations have been highlighted by authors such as Bitzenbauer (2023), Kılınç (2023), Lee and Zhai (2024), and Li et al. (2024), who emphasize the need for more rigorous and comprehensive studies to better understand the impact of chatbots on student learning.

Moreover, the high number of exploratory papers that have been found is logical since, as indicated by Babbie (2016), this type of research aims to familiarize oneself with a little-studied topic, identify relevant variables, and generate preliminary hypotheses. In this review, many educational proposals to support science teachers are evaluated in exploratory studies, mostly through the authors' opinions (with few consultations with experts), and without rubrics or standards for the chatbot's expected answers. In the same vein, authors conducting exploratory research should repeat the same questions several times to determine whether there is diversity in the responses, providing an accuracy percentage, given chatbots' ability to generate varied responses to the same question. Consequently, more robust educational interventions are needed to draw definitive conclusions.

It is important to note that the increasing presence of chatbots in science education demands their integration into curricula in a way that fosters AI literacy for students. This integration should promote critical thinking and technical skills. Although chatbots offer promising applications, they also pose ethical challenges, such as plagiarism and student over-reliance (Alan & Yurt, 2024).

Therefore, professional development programs for teachers are essential (Almasri, 2024; Feldman-Maggor et al., 2025; Labadze et al., 2023). These programs should not only focus on technical skills and best pedagogical practices, but

also thoroughly consider the underlying ethical implications. Without comprehensive training, teachers may struggle to keep up with students' chatbot usage, limiting their ability to facilitate appropriate, technology-supported learning. Equipping educators in these areas is key to maximizing the benefits and minimizing the risks of integrating chatbots into science education.

Some of these risks were noted by Avraamidou (2024), such as the perpetuation of Western monoculturalism, which reinforces the biases with which the models were trained. The exploitation of marginalized communities for the training of ChatGPT (Cooper, 2023; Kılınç, 2023) has also been the subject of criticism, as has the significant environmental impact—both in terms of energy consumption and the carbon footprint associated with these tools (Avraamidou, 2024)—as well as their high-water usage (Li et al., 2023).

We believe that educators should highlight the importance of using these tools in a responsible and ethical manner. Instead of focusing predominantly on potential misuse, they should inspire students to understand how AI, when critically used, can be a potent instrument for societal advancement. Moreover, both researchers and teachers should consider their students' privacy when using certain chatbots by encoding their work without including names (as, for example, recommended by the European Union's General Data Protection Regulation).

Finally, research in this field is still in its early stages. More controlled and randomized studies with empirical data are needed to examine the impact of chatbots across various scientific subjects, particularly in those areas that have received less attention. These studies would provide a more comprehensive understanding of the benefits and challenges associated with chatbot integration. Furthermore, these studies should also address ethical concerns to ensure their responsible and effective adoption.

Conclusions

Chatbots have shown promising potential in science education by offering clear opportunities to assist teachers and reduce their workload. In particular, they can support tasks such as lesson planning, rubric generation, activity suggestions, feedback, and evaluation, usages that remain relatively safe as long as teachers supervise and adapt the outputs with their disciplinary expertise.

When chatbots are used directly by students, however, greater caution is required. Despite their fluency, most models still lack scientific precision and sometimes generate incorrect or overly general answers. Consequently, teachers face both a challenge and an opportunity (indeed, an obligation) to instruct students in the responsible use of these tools, guiding them to scrutinize outputs, identify potential errors, and leverage chatbots as catalysts for deeper understanding rather than mere "copy-and-paste" generators of responses.

For this reason, integrating AI literacy and critical-thinking skills into the science curriculum becomes essential. Students must learn not only how to operate chatbots, but also how to question, verify, and refine their outputs. In turn, educators will require targeted professional development (focused on ethical, pedagogically sound, and evidence-based decision-making) to ensure effective and responsible use of artificial intelligence in the classroom.

Teachers are encouraged to pilot chatbots across diverse science contexts, particularly the markedly under-studied field of geology, as well as abstract topics in physics and chemistry, by designing specific, scaffolded tasks that require students to query the chatbot, critique its responses, and verify them against textbooks, datasets, or laboratory results. These studies should incorporate follow-up assessments over several weeks, months, or even years, enabling teachers to distinguish lasting learning gains from short-lived novelty effects. In this way, educators will build actionable evidence on accuracy, conceptual depth, and retention, and can iteratively refine chatbot integration.

In sum, chatbots should be regarded as valuable assistants, especially for teachers, but not yet as autonomous tutors for students. Their responsible implementation demands a blend of technical knowledge, pedagogical judgment, and ethical awareness.

Limitations of the Study and Future Research

Firstly, although efforts have been made to minimize research bias by searching for articles in different databases, only peer-reviewed articles have been considered. The review bias could be reduced by expanding the search to include book chapters, conference papers, dissertations, etc. It could also be improved if unpublished articles had been found. While these issues pertain to the generalization of results, they are typical challenges in many systematic reviews and meta-analyses.

Secondly, the scope of this review is limited by the characteristics of the included studies. Although the selection process was systematic, the final sample was relatively small, and the analysis of methodological quality revealed several common weaknesses, such as small participant samples, lack of control groups, and the absence of standardized instruments to assess learning outcomes. Furthermore, since the use of chatbots in science education is an emerging field, a large number of exploratory studies were found (32.5% in this review). Exploratory studies tend to have less methodological rigor than interventions or case studies, limiting

the generalization of this review. For all these reasons, the findings of this review should be interpreted with caution.

As a future task, conducting a meta-analysis remains pending until enough empirical studies on variables of interest, such as academic performance, workload in teachers and students, motivation, and attitude towards science using chatbots, become available.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10956-025-10260-x.

Author Contribution All authors conceptualized the study, wrote the first draft of the manuscript, contributed to subsequent drafts of the manuscript, and approved the final version for publication.

Funding This study has been partially supported by the UCLM Internal Grant Funds for research group activities (2022-GRIN-34471).

Data Availability All data generated or analyzed during this study are included in this published article.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing interests The authors declare no competing interests.

References

Agarwal, R., & Wadhwa, M. (2020). Review of state-of-the-art design techniques for chatbots. *SN Computer Science, 1*(5), 246. https://doi.org/10.1007/s42979-020-00255-3

Alan, S., & Yurt, E. (2024). Flipped learning: An innovative model for enhancing education through ChatGPT. *International Journal of Modern Education Studies, 8*(1), 124–148. https://doi.org/10.51383/ijonmes.2024.328

Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education, 54*(5), 977–997. https://doi.org/10.1007/s11165-024-10176-3

Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology, 15*(4), Article ep448. https://doi.org/10.30935/cedtech/13417

Archila, P. A., Ortiz, B. T., Truscott de Mejía, A.-M., & Molina, J. (2024). Thinking critically about scientific information generated by chatGPT. *Information and Learning Sciences, 125*(11/12), 1074–1106. https://doi.org/10.1108/ILS-04-2024-0040

Avraamidou, L. (2024). Can we disrupt the momentum of the AI colonization of science education? *Journal of Research in Science Teaching, 61*(10), 2570–2574. https://doi.org/10.1002/tea.21961

Babbie, E. (2016). *The basics of social research* (7th ed.). Cengage Learning.

Banihashem, K., & Macfadyen, L. P. (2021). Pedagogical design: Bridging learning theory and learning analytics. *Canadian Journal of Learning and Technology*, *47*(1). https://doi.org/10.21432/cjlt27959

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence, 5*, Article 100177. https://doi.org/10.1016/j.caeai.2023.100177

Biasutti, M. (2015). An intensive programme on education for sustainable development: The participants' experience. *Environmental Education Research, 21*(5), 734–752. https://doi.org/10.1080/13504622.2014.921805

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

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Cajas, F. (2001). The science/technology interaction: Implications for science literacy. *Journal of Research in Science Teaching, 38*(7), 715–729. https://doi.org/10.1002/tea.1028

Cebrián, G., Mogas, J., Palau, R., & Fuentes, M. (2022). Sustainability and the 2030 agenda within schools: A study of school principals' engagement and perceptions. *Environmental Education Research, 28*(6), 845–866. https://doi.org/10.1080/13504622.2022.2044017

Chang, J., Park, J., & Park, J. (2023). Using an artificial intelligence chatbot in scientific inquiry: Focusing on a guided-inquiry activity using Inquirybot. *Asia-Pacific Science Education, 9*(1), 44–74. https://doi.org/10.1163/23641177-bja10062

Chen, C.-H., & Chang, C.-L. (2024). Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies, 29*(14), 18621–18642. https://doi.org/10.1007/s10639-024-12553-x

Cheung, K. K. C., Pun, J. K. H., & Li, W. (2024). Students' holistic reading of socio-scientific texts on climate change in a ChatGPT scenario. *Research in Science Education, 54*(5), 957–976. https://doi.org/10.1007/s11165-024-10177-2

Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2172044

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444–452. 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 Bruyckere, P., Kirschner, P. A., & Hulshof, C. D. (2015). Urban myths about learning and education. *Elsevier*. https://doi.org/10.1016/C2013-0-18621-7

Debets, T., Banihashem, S. K., Joosten-Ten Brinke, D., Vos, T. E. J., de Maillette Buy Wenniger, G., & Camp, G. (2025). Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. *Computers & Education, 234*, Article 105323. https://doi.org/10.1016/j.compedu.2025.105323

Deveci-Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies, 26*(5), 6241–6265. https://doi.org/10.1007/s10639-021-10627-8

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education, 20*(1), 63. https://doi.org/10.1186/s41239-023-00434-1

Duy, H. T., Ngoc, C. T., & Hai, N. T. (2024). Building and using chatbots in the process of self-studying physics to improve the quality of learners' knowledge. *International Journal of Education and*

Practice, 12(4), 1165–1185. https://doi.org/10.18488/61.v12i4.3859

Ekici, E. (2016). Why do i slog through the physics? Understanding high school students' difficulties in learning physics. *Journal of Education and Practice*, *7*(7). www.iiste.org

Erinosho, S. (2013). How do students perceive the difficulty of physics in secondary school? An exploratory study in Nigeria. *International Journal for Cross-Disciplinary Subjects in Education, 3*(Special 3), 1510–1515. https://doi.org/10.20533/ijcdse.2042.6364.2013.0212

Feldman-Maggor, Y., Blonder, R., & Alexandron, G. (2025). Perspectives of generative AI in chemistry education within the TPACK framework. *Journal of Science Education and Technology, 34*(1), 1–12. https://doi.org/10.1007/s10956-024-10147-3

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using chatGPT. *Journal of Chemical Education, 100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), Article 035021. https://doi.org/10.1088/1361-6552/acc299

Güldal, H., & Dinçer, E. O. (2024). Can rule-based educational chatbots be an acceptable alternative for students in higher education? *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12977-5

Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education, 100*(12), 4876–4883. https://doi.org/10.1021/acs.jchemed.3c00505

Heeg, D. M., & Avraamidou, L. (2023). The use of artificial intelligence in school science: A systematic literature review. *Educational Media International, 60*(2), 125–150. https://doi.org/10.1080/09523987.2023.2264990

Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., & Dimitrova, V. (2022). Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law. *Council of Europe*. https://book.coe.int/en/education-policy/11334-pdf-artificial-intelligence-and-education-a-critical-view-through-the-lens-of-human-rights-democracy-and-the-rule-of-law.html

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Holmes, W. (2023). *The unintended consequences of artificial intelligence and education*. Education International. https://discovery.ucl.ac.uk/id/eprint/10179267/1/Holmes%20-%202023%20-%20The%20Unintended%20Consequences%20of%20Artificial%20Intellig.pdf

Huang, H.-W., Teng, D.C.-E., & Tiangco, J. A. N. Z. (2024). The impact of AI chatbot-supported guided discovery learning on preservice teachers' learning performance and motivation. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-024-10179-9

Humphry, T., & Fuller, A. L. (2023). Potential ChatGPT use in undergraduate chemistry laboratories. *Journal of Chemical Education, 100*(4), 1434–1436. https://doi.org/10.1021/acs.jchemed.3c00006

Hwang, G.-J., & Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments, 31*(7), 4099–4112. https://doi.org/10.1080/10494820.2021.1952615

Johnson, L., & Ricker, M. C. (2017). *2017 Educator Quality of Work Life Survey*. American Federation of Teachers. https://www.aft.org/sites/default/files/media/2017/2017_eqwl_survey_web.pdf

Kılınç, S. (2023). Embracing the future of distance science education: Opportunities and challenges of ChatGPT integration. *Asian Journal of Distance Education*, *18*(1). https://doi.org/10.5281/zenodo.7857395

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kirschner, P. A., & Hendrick, C. (2024). *How learning happens: Seminal works in educational psychology and what they mean in practice* (2nd ed.). Routledge. https://doi.org/10.4324/9781003395713

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research, 19*(2), Article 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1), 56. https://doi.org/10.1186/s41239-023-00426-1

Latif, E., & Zhai, X. (2023). *Fine-tuning ChatGPT for automatic scoring*. http://arxiv.org/abs/2310.10072

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. *Cambridge University Press*.

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*, 1643–1660. https://doi.org/10.1109/TLT.2024.3401457

Lee, J., An, T., Chu, H.-E., Hong, H.-G., & Martin, S. N. (2023). Improving science conceptual understanding and attitudes in elementary science classes through the development and application of a rule-based AI chatbot. *Asia-Pacific Science Education*. https://doi.org/10.1163/23641177-bja10070

Li, T., Ji, Y., & Zhan, Z. (2024). Expert or machine? Comparing the effect of pairing student teacher with in-service teacher and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course. *Asia Pacific Journal of Education, 44*(1), 45–60. https://doi.org/10.1080/02188791.2024.2305163

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models* [Preprint, arXiv version 5]. arXiv. https://doi.org/10.48550/arXiv.2304.03271

Liang, Y., Zou, D., Xie, H., & Wang, F. L. (2023). Exploring the potential of using ChatGPT in physics education. *Smart Learning Environments, 10*(1), 52. https://doi.org/10.1186/s40561-023-00273-7

Lin, Y.-T., & Ye, J.-H. (2023). Development of an educational Chatbot system for enhancing students' biology learning performance. 網際網路技術學刊, *24*(2), 275–281. https://doi.org/10.53106/160792642023032402006

Linn, M. (2003). Technology and science education: Starting points, research programs, and trends. *International Journal of Science Education, 25*(6), 727–758. https://doi.org/10.1080/09500690305017

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policy-makers*. UNESCO. https://doi.org/10.54675/PCSP7350

Muijs, D., & Reynolds, D. (2017). *Effective teaching: Evidence and practice* (4th ed.). SAGE Publications Ltd.

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through <scp>ChatGPT</scp>: A pioneering pilot study. *British Journal of Educational Technology, 55*(4), 1328–1353. https://doi.org/10.1111/bjet.13454

Nguyen, H. (2023). Role design considerations of conversational agents to facilitate discussion and systems thinking. *Computers & Education, 192*, Article 104661. https://doi.org/10.1016/j.compedu.2022.104661

NIH. (2022). *National Institutes of Health, Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies, National Heart, Lung, and Blood Institute*. https://www.nhlbi.nih.gov/Health-pro/Guidelines/Indevelop/Cardiovascular-Risk-Reduction/Tools/Cohort.

Ojala, M. (2021). Safe spaces or a pedagogy of discomfort? Senior high-school teachers' meta-emotion philosophies and climate change education. *The Journal of Environmental Education, 52*(1), 40–52. https://doi.org/10.1080/00958964.2020.1845589

Okulu, H. Z., & Muslu, N. (2024). Designing a course for pre-service science teachers using ChatGPT: What ChatGPT brings to the table. *Interactive Learning Environments, 32*(10), 7450–7467. https://doi.org/10.1080/10494820.2024.2322462

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122–133. https://doi.org/10.1037/0022-0663.86.1.122

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, n160. https://doi.org/10.1136/bmj.n160

Park, H. K., & Martin, S. N. (2024). Exploring the role of ChatGPT in science education for Asia-Pacific and beyond: A systematic review. *Asia-Pacific Science Education, 10*(2), 233–264. https://doi.org/10.1163/23641177-bja10079

Peikos, G., & Stavrou, D. (2025). ChatGPT for science lesson planning: An exploratory study based on pedagogical content knowledge. *Education Sciences, 15*(3), 338. https://doi.org/10.3390/educsci15030338

Ramkorun, B. (2024). Graph plotting of 1-D motion in introductory physics education using scripts generated by ChatGPT 3.5. *Physics Education, 59*(2), Article 025020. https://doi.org/10.1088/1361-6552/ad2191

Revalde, G., Zholdakhmet, M., Abola, A., & Murzagaliyeva, A. (2025). Can ChatGPT pass a physics test? *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-025-09814-0

Ruff, E. F., Franz, J. L., & West, J. K. (2024). Using chatgpt for method development and green chemistry education in upper-level laboratory courses. *Journal of Chemical Education, 101*(8), 3224–3232. https://doi.org/10.1021/acs.jchemed.4c00193

Saçkes, M. (2015). Young children's ideas about earth and space science concepts. In *Research in Early Childhood Science Education* (pp. 35–65). Springer Netherlands. https://doi.org/10.1007/978-94-017-9505-0_3

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Mission AI. *Springer International Publishing*. https://doi.org/10.1007/978-3-031-21448-6

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4. https://doi.org/10.2307/1175860

Sperling, A., & Lincoln, J. (2024). Artificial intelligence and high school physics. *The Physics Teacher, 62*(4), 314–315. https://doi.org/10.1119/5.0202994

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education, 101*(6), 2475–2482. https://doi.org/10.1021/acs.jchemed.4c00212

Uğraş, H., & Uğraş, M. (2024). ChatGPT in early childhood STEM education: Can it be an innovative tool to overcome challenges? *Education and Information Technologies, 30*, 4277–4305. https://doi.org/10.1007/s10639-024-12960-0

Vasconcelos, M. A. R., & dos Santos, R. P. (2023). Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with: A case study. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7), Article em2296. https://doi.org/10.29333/ejmste/13313

Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research, 20*(1), Article 010152. https://doi.org/10.1103/PhysRevPhysEducRes.20.010152

Wang, J. T. H. (2023). Is the laboratory report dead? AI and ChatGPT. *Microbiology Australia, 44*(3), 144–148. https://doi.org/10.1071/MA23042

Wei, X., Wang, L., Lee, L.-K., & Liu, R. (2024). Multiple generative AI pedagogical agents in augmented reality environments: A study on implementing the 5E model in science education. *Journal of Educational Computing Research, 63*(2), 336–371. https://doi.org/10.1177/07356331241305519

Willingham, D. T. (2019). *How to teach critical thinking*. NSW Department of Education. https://education.nsw.gov.au/aboutus/educational-data/cese/publications/education-for-a-changing-world-pubs

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yin, J., Goh, T.-T., & Hu, Y. (2024a). Interactions with educational chatbots: The impact of induced emotions and students' learning motivation. *International Journal of Educational Technology in Higher Education, 21*(1), 47. https://doi.org/10.1186/s41239-024-00480-3

Yin, J., Zhu, Y., Goh, T.-T., Wu, W., & Hu, Y. (2024b). Using educational chatbots with metacognitive feedback to improve science learning. *Applied Sciences, 14*(20), 9345. https://doi.org/10.3390/app14209345

Zhang, R., Zou, D., & Cheng, G. (2024). A review of chatbot-assisted learning: Pedagogical approaches, implementations, factors leading to effectiveness, theories, and future directions. *Interactive Learning Environments, 32*(8), 4529–4557. https://doi.org/10.1080/10494820.2023.2202704