Contemporary Educational Technology, 2026, **18**(3), ep667
ISSN: 1309-517X (Online)

OPEN ACCESS

Review Article

The integration of generative artificial intelligence in secondary education: A systematic review

Héctor Pérez-Montesdeoca [1*]
0000-0002-6728-8681

Daniel Rodríguez-Rodríguez [1]
0000-0002-9721-259X

Aitana Fernández-Sogorb [2]
0000-0003-2709-1099

[1] Universidad Europea de Canarias, Faculty of Social Sciences, Canary Islands, SPAIN
[2] University of Alicante, Alicante, SPAIN
[*] Corresponding author: hector.perez@universidadeuropea.es

Citation: Pérez-Montesdeoca, H., Rodríguez-Rodríguez, D., & Fernández-Sogorb, A. (2026). The integration of generative artificial intelligence in secondary education: A systematic review. *Contemporary Educational Technology, 18*(3), Article ep667. https://doi.org/10.30935/cedtech/18746

ARTICLE INFO

Received: 16 Jul 2025

Accepted: 20 May 2026

ABSTRACT

In recent years, the emergence of generative artificial intelligence (GenAI) has reshaped multiple domains of human knowledge, including education, giving rise to an emerging field of study that still lacks conceptual and empirical systematization—particularly at the secondary education level. Despite growing interest in exploring its pedagogical potential, existing studies remain fragmented, methodologically uneven, and often rooted in experimental or anecdotal contexts, which hinders the development of a robust evidence base regarding its actual impact on learning. In response to this situation, the present study conducts a systematic review of recent scientific literature with the aim of identifying the main uses of GenAI in secondary education and examining the improvements these uses bring to teaching and learning processes. The review follows the PRISMA protocol and includes a total of 33 studies selected based on explicit inclusion criteria, focusing on experiences involving generative tools. The findings reveal a diverse range of approaches to GenAI integration, with a predominance of applications in written production, STEM problem-solving, creative stimulation, and automated feedback—most of which are initiated by teachers and implemented in isolated or experimental settings. The review also identifies significant improvements in areas such as student motivation, autonomy, critical thinking, and digital competence. However, methodological limitations and gaps in pedagogical integration are also noted. These findings underscore the need to move towards more integrated and sustained pedagogical models and highlight the urgency of strengthening longitudinal and theoretically grounded research to gain deeper insights into the educational implications of this emerging technology.

Keywords: applications, benefits, educational technology, emerging technology, innovation

INTRODUCTION

Artificial Intelligence and Generative Artificial Intelligence in Education

Artificial intelligence (AI) has evolved significantly since its early conceptualization as the capacity of machines to simulate human intelligence (McCarthy et al., 2006). Advances in data availability, machine learning, and computational power have expanded its application across multiple domains, including education (Lucas et al., 2024). Today, AI is broadly understood as an interdisciplinary field focused on the development of systems capable of performing tasks that typically require human cognitive abilities, such as problem-solving, adaptation, and decision-making.

As AI increasingly overlaps with human capabilities, it is reshaping educational practices by introducing new opportunities and challenges (Lu & Fan, 2023; Vincent-Lancrin & van der Vlies, 2020). Within this context, artificial intelligence in education (AIEd) supports teaching and learning by automating administrative tasks, enabling personalized instruction, and providing adaptive learning experiences. These tools facilitate real-time identification of learning needs and offer tailored resources that can enhance student autonomy and performance (Martínez-Comesaña et al., 2023).

However, the integration of AIEd also raises important concerns. These include risks related to reduced teacher autonomy, ethical issues such as bias and transparency, and challenges linked to equity and the potential impact on students' social development (Akgun & Greenhow, 2022; Bhimdiwala et al., 2022). At the same time, a growing body of research highlights its potential to improve learning outcomes when implemented within pedagogically sound frameworks (Holmes et al., 2022).

Recent advances in machine learning, particularly deep learning, have led to the emergence of generative artificial intelligence (GenAI), marking a significant shift in the educational technology landscape. The public release of tools such as ChatGPT in 2022 accelerated both research and practical applications in educational contexts (Martínez-Comesaña et al., 2023).

Unlike earlier AI systems, GenAI is capable of generating original and context-sensitive content, including texts, images, and feedback. This enables new forms of interaction with knowledge, allowing students to engage more actively in learning processes through personalized support, immediate feedback, and adaptive resources (Kangasharju et al., 2022). As a result, traditional roles in education are being redefined, with increasing emphasis on student agency and teacher mediation.

GenAI offers several pedagogical benefits, including the generation of personalized materials, support for problem-solving and creativity, and the provision of real-time feedback that can enhance understanding and motivation (Ifenthaler & Schumacher, 2023; Tapalova et al., 2022). It also contributes to the efficiency of instructional design and assessment processes.

Nevertheless, these opportunities are accompanied by significant challenges. Issues related to reliability, bias, authorship, and overreliance on automated outputs highlight the need for a critical and balanced integration of GenAI within educational practice (Murphy, 2019). In this context, its educational value depends not on the technology itself, but on how it is pedagogically implemented.

Pedagogical Frameworks for Integrating GenAI in Education

Digital technologies do not inherently lead to meaningful educational transformation; their value depends on how they are integrated into teaching practices and learning activities (Radović, 2024). This is particularly relevant in the case of GenAI, as these tools not only generate content but also shape how students engage with knowledge. Accordingly, their analysis requires a conceptual grounding that accounts for both technological integration and pedagogical implications.

From a teaching perspective, the technological pedagogical content knowledge (TPACK) model (Mishra & Koehler, 2006) highlights that effective integration depends on the interplay between disciplinary, pedagogical, and technological knowledge. In parallel, the SAMR model (Puentedura, 2006) differentiates levels of technological use, from substitution to redefinition, allowing GenAI applications to be interpreted in terms of their transformative potential in learning tasks.

From a learning perspective, constructivist approaches emphasize the role of active knowledge construction, where GenAI can support processes such as exploration, questioning, and explanation (Doolittle et al., 2023). Similarly, self-regulated learning frameworks highlight its potential to enhance autonomy and metacognitive reflection through scaffolding mechanisms such as feedback and example generation (Banihashem et al., 2025). In addition, the ICAP framework (Chi & Wylie, 2014) provides a useful lens to analyze how different uses of GenAI may promote varying levels of cognitive engagement, from passive interaction to constructive and interactive learning processes.

Finally, these perspectives are complemented by frameworks related to digital competence and AI literacy, which stress the importance of developing not only technical skills but also critical awareness of the limitations, biases, and ethical implications of AI systems.

Taken together, these frameworks provide a structured basis for interpreting the educational use of GenAI beyond a purely functional perspective. On this basis, the present systematic review aims to analyze how GenAI is being integrated into secondary education and its implications for teaching and learning processes.

METHOD

This systematic review was conducted in accordance with the PRISMA guidelines (Page et al., 2021) and followed established procedures for transparent and replicable evidence synthesis (Gough et al., 2017).

Research Aims and Search Strategy

The review was guided by two research questions (RQs):

1. **RQ1.** How is GenAI currently used in secondary education? and
2. **RQ2.** What impact does its use have on students' learning processes?

Accordingly, the study aimed to identify both the main pedagogical uses of GenAI and its effects on teaching and learning.

Eligibility Criteria

This review focuses on studies that describe and examine the use of GenAI in secondary education. The year 2019 was selected as the starting point for the search to capture the period preceding the rapid expansion of research on GenAI following the public release of ChatGPT in 2022, which significantly accelerated interest in these technologies in educational contexts (Martínez-Comesaña et al., 2023). Accordingly, the review included empirical studies published between 2019 and 2024, reflecting the recent emergence of GenAI technologies, and considered qualitative, quantitative, and mixed-methods research designs. The studies included in this systematic review were required to meet the following inclusion criteria:

(1) studies published after 2019 in Web of Science and Scopus,

(2) studies published in English or Spanish,

(3) primary research articles,

(4) research papers published in peer-reviewed academic journals,

(5) studies focused on the secondary education level,

(6) studies analyzing GenAI, and

(7) studies with full-text availability.

Consequently, the exclusion criteria applied in this study were as follows:

(1) studies published before 2019 in Web of Science or Scopus,

(2) studies published in languages other than English or Spanish,

(3) non-primary research articles, such as systematic reviews, literature reviews, or meta-analyses,

(4) documents not classified as peer-reviewed research articles, such as books, conference proceedings, doctoral theses, book chapters, manuals, etc.,

(5) studies focusing on educational levels other than secondary education,

(6) studies not addressing GenAI, and

(7) studies without full-text access.

Although the search was conducted in November 2024, some studies retrieved appear with a 2025 publication year due to the common practice of advance online publication, in which articles are assigned to subsequent volumes after their initial online release.

Search Strategy

On 11 November 2024, a systematic search was conducted across two electronic databases: Scopus and Web of Science. At the time of the search, keywords were grouped into two main categories:

(a) GenAI and

(b) educational level.

Variations and synonyms of the keywords were subsequently included to maximize retrieval and minimize semantic bias.

The keywords and their synonyms were incorporated into the following search strategy: ("Artificial Intelligence" OR AI OR "Inteligencia Artificial" OR IA OR "Generative Artificial Intelligence" OR "Generative AI" OR GenAI OR GAI OR "Inteligencia Artificial Generativa" OR "AI Generativa" OR "ChatGPT" OR GPT) AND ("High School" OR "Secondary Education" OR "Educación Secundaria" OR K-12 OR "Middle School" OR "Middle Education").

This search yielded a total of 3,381 studies. After applying the aforementioned inclusion and exclusion criteria and removing duplicates, 676 articles remained for further analysis.

Study Selection

The search results (n = 676) were assessed independently in two phases by two reviewers. First, titles and abstracts were screened to identify studies that met the inclusion criteria. In the second phase, the full texts of the remaining articles were reviewed to ensure alignment with the objectives of this systematic review.

In cases where the title and abstract did not provide sufficient information to determine eligibility, the article was assessed during the full-text screening stage. A high-sensitivity approach was adopted during selection; therefore, in cases of uncertainty, studies were included rather than excluded.

To ensure independence and inter-rater reliability, two authors independently reviewed the full texts. Disagreements regarding study eligibility were discussed until consensus was reached. When necessary, the reviewers revisited the predefined inclusion and exclusion criteria to guide the final decision. Inter-rater reliability was high (κ = .89).

After removing duplicates and applying the inclusion and exclusion criteria, a total of 676 studies were initially identified. Following the screening of titles and abstracts, 322 publications were excluded, along with 3 studies due to lack of access and 2 studies that had been retracted. This resulted in 349 articles eligible for full-text review.

Subsequently, 257 articles were excluded for not addressing GenAI, and a further 46 were excluded for not focusing on the use of GenAI tools or their contribution to learning improvement. Additionally, 13 articles were excluded for not being related to the secondary education level. The complete selection process is detailed in the flow diagram presented in **Figure 1**.

Quality Appraisal

To assess the methodological quality of the studies included in this review, a quality appraisal procedure was conducted using the mixed methods appraisal tool (MMAT) (Hong et al., 2018). This instrument was selected because it enables the evaluation of qualitative, quantitative, and mixed-methods research designs within a single framework, which was appropriate given the methodological diversity of the studies included in this review.

The MMAT was applied only to studies that reported original empirical data collection involving students, teachers, or classroom implementations. Studies focusing exclusively on technical evaluations of AI systems, benchmark analyses, or conceptual discussions did not involve human participants and were therefore not subjected to methodological quality appraisal. These studies were nevertheless retained in the review because they contributed to understanding the uses and capabilities of GenAI tools in secondary education.

Two reviewers independently assessed the methodological quality of the eligible studies according to the five MMAT criteria corresponding to their research design (qualitative, quantitative, or mixed methods). The appraisal considered aspects such as the clarity of the RQs, the appropriateness of the methodology, the adequacy of data collection procedures, and the coherence between data analysis and the conclusions drawn.

Discrepancies in the ratings were discussed until consensus was reached. The results of the quality assessment were used to interpret the findings of the review and to identify potential methodological limitations in the existing literature.

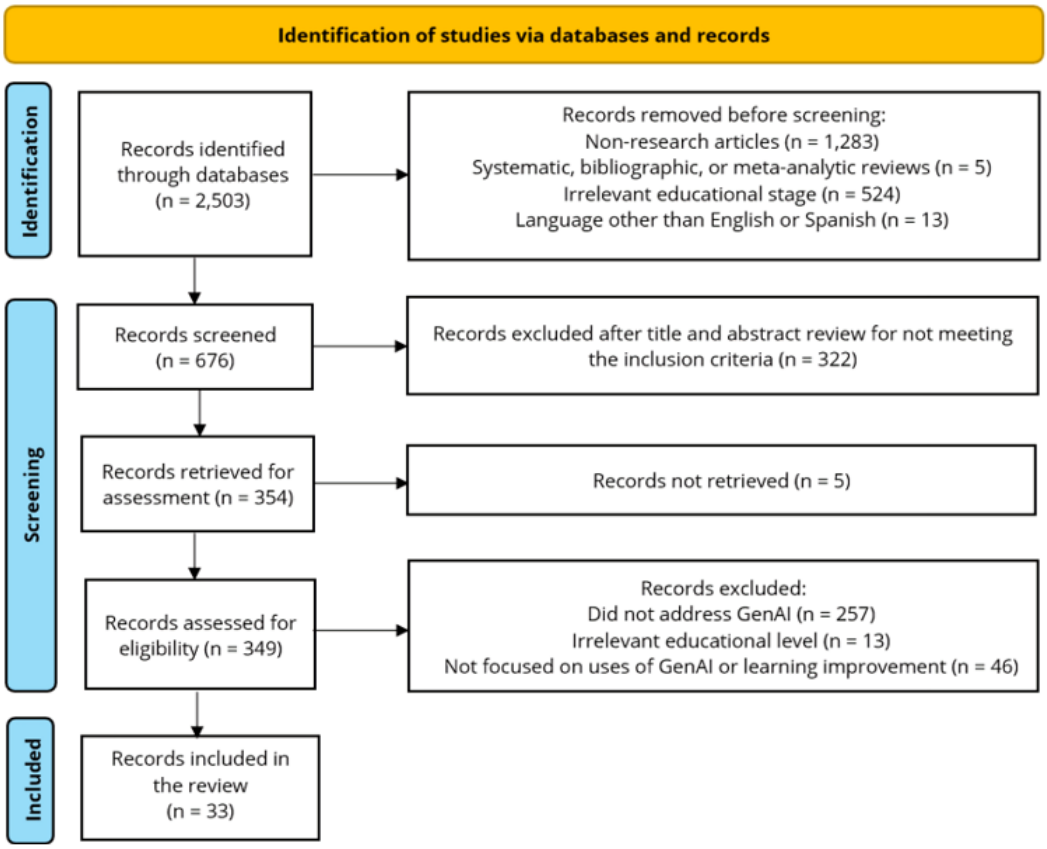

Figure 1. PRISMA flow diagram (Source: Authors' own elaboration)

RESULTS

Quality Appraisal of Included Studies

A methodological quality appraisal was conducted for empirical studies included in this review using the MMAT. Of the 33 included studies, 21 were empirical and 12 were technical or conceptual in nature. Of these, 21 reported original empirical data involving students, teachers, or classroom implementations and were therefore eligible for quality assessment. The remaining studies consisted of technical evaluations of AI systems, benchmark analyses, or conceptual discussions and were not subjected to MMAT appraisal, although they were retained in the review because they contributed to understanding the uses and capabilities of GenAI tools in secondary education.

Overall, the methodological quality of the empirical studies ranged from moderate to high. Most studies presented clearly defined research objectives and employed data collection methods appropriate to their research designs. In addition, the majority of studies demonstrated coherence between the RQs, the methodological approach, and the interpretation of results. However, several methodological limitations were also identified. Some studies relied on relatively small samples or exploratory designs, while others provided limited information regarding sampling strategies, instrument validation, or the integration of qualitative and quantitative data. These methodological differences should be considered when interpreting the reported findings and highlight the need for more robust and large-scale research on the educational use of GenAI in secondary education. The detailed results of the methodological quality appraisal are presented in **Table 1**.

Description of Included Studies

The literature search for this systematic review yielded a total of 33 studies that met the established inclusion criteria, as detailed in **Appendix A**. In terms of geographical distribution, the majority of studies were conducted in the United States (n = 8), followed by China and Germany (n = 3), and Taiwan and Spain (n = 2). Additional studies were identified from other countries, including Canada, Finland, the Netherlands, Singapore, and Turkey.

Table 1. Methodological quality appraisal of empirical studies using the MMAT

Study	Design	C1	C2	C3	C4	C5	Quality
Ali et al. (2021)	Quantitative non-randomized	✓	✓	✓	✓	✓	High
Aprin et al. (2024)	Quantitative non-randomized	✓	✓	✓	✓	△	High
Chen and Chang (2024)	Quantitative randomized controlled	✓	✓	✓	✓	✓	High
Clark et al. (2024)	Qualitative	✓	✓	✓	✓	✓	High
Cohn et al. (2025)	Qualitative	✓	✓	✓	✓	△	High
de Souza et al. (2024)	Mixed methods	✓	✓	✓	△	△	Moderate
Duan and Zhao (2024)	Quantitative descriptive	✓	✓	✓	✓	✓	High
Ergene and Ergene (2025)	Mixed methods	✓	✓	✓	△	△	Moderate
Hsiao and Chang (2023)	Quantitative non-randomized	✓	✓	✓	✓	✓	High
Jiang et al. (2023)	Quantitative descriptive	✓	✓	✓	✓	△	High
Kangasharju et al. (2022)	Qualitative	✓	✓	✓	✓	✓	High
Khan et al. (2024)	Quantitative non-randomized	✓	✓	✓	✓	✓	High
Kong and Yang (2024)	Qualitative	✓	✓	✓	✓	△	High
Küchemann et al. (2023)	Quantitative descriptive	✓	✓	✓	△	△	Moderate
Lee et al. (2025)	Qualitative	✓	✓	✓	✓	△	High
Levine et al. (2025)	Qualitative	✓	✓	✓	✓	✓	High
Meyer et al. (2024)	Quantitative randomized controlled	✓	✓	✓	✓	✓	High
Shikun et al. (2024)	Mixed methods	✓	✓	✓	✓	△	High
Shin et al. (2021)	Mixed methods	✓	✓	✓	✓	✓	High
Tamimi et al. (2024)	Mixed methods	✓	✓	✓	△	△	Moderate
Tang et al. (2024)	Qualitative (action research)	✓	✓	✓	✓	△	High

Note. C1-C5 correspond to the methodological criteria of the MMAT according to the study design; ✓ is criterion met; △ is partially met; & Only studies reporting empirical data involving human participants were subjected to methodological quality appraisal

With regard to temporal distribution, there has been a noticeable increase in the number of publications in recent years. Specifically, 2 studies were published in 2021, 2 in 2022, 6 in 2023, and the majority (n = 23) in 2024. This trend reflects the growing academic interest in GenAI within the context of secondary education.

In terms of thematic focus, the studies included in this review can be grouped into two main lines of research. A total of 26 studies concentrated on analyzing the uses and applications of GenAI in secondary education contexts, addressing topics such as content generation, language learning, question answering, among others. On the other hand, 7 studies evaluated the impact of these technologies on the improvement of teaching and learning processes, including teacher training, students' ability to detect deepfakes, understood as AI-generated synthetic media capable of manipulating images, audio, or video to create highly realistic but fabricated content, and the benefits derived from GenAI-generated content. Additionally, one study combined both lines of inquiry. The sample sizes analyzed in the studies included in the review showed considerable variability. Four studies involved samples ranging from 1 to 20 students, one of which also included the participation of four teachers. Similarly, three studies worked with samples between 21 and 40 students, while four studies included between 41 and 70 participants.

In addition, two investigations involved samples ranging from 90 to 160 students, one of which also included 80 participating teachers. Likewise, two studies had samples ranging from 200 to 460 students, and one study reported a significantly larger sample composed of 1200 students. Moreover, three studies were identified whose samples consisted exclusively of teachers. Of these, two studies included between 4 and 35 participants, while another reported a significantly larger sample involving 3,300 teachers. One further study involved 26 prospective secondary education teachers. It is also worth noting that a total of 12 studies did not include human participants, as their focus was on analyzing the effectiveness of various GenAI tools within the context of secondary education.

RQ1. How is GenAI Being Used in Secondary Education?

The analysis of the 33 studies included in this systematic review showed that 26 focused on the use of GenAI in secondary education, grouped into distinct categories reflecting its educational applicability. These included the creation of narratives and materials (2 studies), question and activity generation (3), writing support (1), and image generation (2). A larger subset examined GenAI in problem-solving tasks across subjects (7), foreign language learning (2), assessment processes (6), and teacher training (3). The following section details the findings in each category, outlining the main trends and applications identified.

Story and narrative creation

Bulut and Yildirim-Erbasli (2022) noted that GenAI facilitates the creation of narratives by generating fluent, coherent, and grammatically correct stories that are comparable to those produced by humans. This capability enables teachers to develop narratives quickly and efficiently, without the need to invest time in sourcing additional materials. However, the authors caution that the generated texts require review and editing to correct potential semantic errors, repetitions, or abrupt thematic shifts.

Similarly, Carrasco (2023) highlighted that the use of GenAI tools in teaching Modern History allows for the automated generation of theoretical content, educational resources, and practical activities. This enriches the teaching and learning process by providing teachers with greater flexibility in content creation, facilitating the adaptation of materials to students' needs, and optimizing lesson preparation time.

Question and activity creation

Tang et al. (2024) identified five categories of student questions when interacting with ChatGPT: elaborate (seeking specificity), expand (introducing new perspectives), contextualize (understanding background), verify (assessing source reliability), and clarify (metacognitive reflection). Similarly, Küchemann et al. (2023) compared physics questions created by ChatGPT and prospective teachers. Both achieved high conceptual accuracy, but ChatGPT's lacked contextualization unless explicitly prompted—highlighting user limitations rather than flaws in the tool itself. The authors concluded that GenAI can effectively support assessment design if paired with teacher review. Likewise, Araújo and Saúde (2024) found that while ChatGPT accurately outlined procedures for lab tasks, inconsistencies in scientific and pedagogical content sometimes hindered learning objectives, reaffirming the importance of teacher supervision in guiding its use.

Writing support

The study by Levine et al. (2025) explored how students engaged with ChatGPT during writing tasks, identifying its supportive role in planning, revision, and translation. Planning was the most common use, involving goal setting, idea generation, and content organization. Importantly, students did not copy ChatGPT's responses but used them as a foundation for their own work. During revision, they relied on the tool for grammar, syntax, and style improvements, though often accepted or rejected wholesale suggestions without detailed evaluation. For translation, students used ChatGPT's examples as models to construct original sentences, rather than replicating its outputs.

Image generation

The research conducted by Díaz-Sánchez and Chapinal-Heras (2024) highlighted that GenAI-based image generation tools enable students to visualize and reinterpret literary sources through visual representations, thereby facilitating communication and the reprocessing of textual information. Additionally, displaying the generated images in the classroom fosters discussion around their fidelity to the original descriptions and the sources used.

Similarly, the study by de Souza et al. (2024) demonstrated that these tools allow teachers to assess students' understanding of complex concepts. Students who achieved higher scores in their evaluations tended to generate more accurate images focused on core subject principles, whereas those with lower performance had greater difficulty visually representing concepts, reflecting gaps in their understanding. Although some of the outcomes did not align with students' initial expectations, the researchers observed that the creative process encouraged deeper reflection on the subject matter.

Question, exercise, and exam solving

A range of studies has assessed the performance of GenAI tools across scientific disciplines, highlighting strong outcomes in basic tasks but notable limitations with increasing complexity. Dao and Le (2023) found BingChat and Bard highly accurate (92.11%) in biology knowledge questions, yet all tools struggled with higher-order items (below 20% accuracy), with ChatGPT showing greater consistency but lower overall scores. Similar patterns were observed in other areas: de Winter (2024) reported improved performance in English reading comprehension with GPT-4 models, and Yeadon and Hardy (2024) noted a decline in ChatGPT-3.5's

accuracy from 83.4% to 63.8% as physics content became more advanced. LLMs also struggled with large numerical calculations and automated grading, showing only 50.8% agreement with human evaluators. Polverini and Gregorcic (2024) showed ChatGPT's interpretation of kinematic graphs mirrored student performance but was hindered by visual misinterpretation. Zhai et al. (2025) found both ChatGPT and GPT-4 outperformed most 8th graders and 12th graders in science NAEP tests, with stable performance regardless of cognitive demand. In mathematics, Ergene and Ergene (2025) noted that GPT-4 and GPT-4o achieved high success rates (82%-87%) across problem types, while Gemini performed poorly (32%). Parra et al. (2024) identified critical limitations in geometry, with ChatGPT-3.5 and Bard failing to accurately convert descriptions into visual representations or apply conceptual reasoning.

Foreign language learning

Shin et al. (2021) analyzed the use of GenAI as a conversational partner in second language learning. The interaction was generally smooth; however, students played a passive role, generating less than half as many utterances as the GenAI system. Despite high levels of satisfaction, the researchers noted that overly lengthy responses and irrelevant comments hindered both comprehension and the fluency of dialogue.

In a related study, Shikun et al. (2024) investigated the impact of GenAI on the speaking skills of students learning English as a foreign language. The results revealed improvements in pronunciation, intonation, and stress, particularly among intermediate-level learners. However, fluency in reading aloud showed progress only in this group. Additionally, GenAI supported the development of writing skills by providing grammatical corrections and encouraging debate through immediate feedback.

Assessment

Several studies highlight the growing role of GenAI in assessment. Ciampa et al. (2025) outlined five levels of teacher-led integration, ranging from automated grading of closed questions to student-led critique of GenAI-generated content fostering critical thinking. Meyer et al. (2024) found that ChatGPT-3.5-turbo feedback significantly improved students' writing revisions and motivation. Similarly, Cohn et al. (2025) noted that GenAI facilitated real-time progress tracking and informed lesson planning. Jiang et al. (2023) reported high alignment between LLMs and human feedback in writing correction, with GPT-4 showing the greatest accuracy. Latif and Zhai (2024) demonstrated that fine-tuned ChatGPT-3.5-turbo outperformed BERT in evaluating scientific responses, especially in multi-class tasks. Khan et al. (2024) further evidenced the positive impact of GenAI in L2 vocabulary assessment, where immediate, tailored feedback nearly doubled student scores and reduced the need for teacher intervention.

Teacher training

The integration of GenAI into teacher training has yielded notable gains in autonomy and professional growth. Duan and Zhao (2024) observed that teachers using GenAI tools demonstrated greater autonomy and significant improvements in professional competencies during online teaching, also benefiting from reduced digital burnout. Kong and Yang (2024), applying the TPACK framework, reported enhancements in teachers' content, pedagogical, and technological knowledge, which translated into improved student engagement and satisfaction. Similarly, Clark et al. (2024) noted that ChatGPT supported instructional planning and assessment design, though concerns were raised about the originality and subject-specific quality of the generated content.

Taken together, these studies indicate that GenAI holds significant potential to expand and enrich assessment practices in secondary education, particularly in areas such as formative feedback, automated scoring, and real-time monitoring of student progress. In several cases, GenAI-generated feedback produced outcomes comparable to—or exceeding—those generated by human evaluators. However, the evidence also points to important limitations: inconsistencies in feedback quality, reduced accuracy in evaluating complex discursive elements, and a tendency to produce uniform responses that may not adapt to individual student profiles. These findings suggest that the most productive role for GenAI in assessment is as a complement to teacher judgment rather than a replacement for it.

RQ2. What Improvements Does the Use of GenAI Tools Bring to the Learning Process of Secondary School Students?

In line with the findings of this systematic review, seven studies examined the improvements that GenAI generates in the learning process of secondary education students. Firstly, two studies addressed deepfake detection, assessing its impact on critical thinking and the ability to identify manipulated content. Three investigations explored broader improvements resulting from the use of GenAI, such as reduced cognitive load, understood as the amount of mental effort required to process information during learning tasks, and enhanced academic performance. Additionally, two studies focused on the influence of GenAI on grammatical improvement and writing development in foreign language learning. The specific findings within each of these categories are presented below.

Detection of deepfakes

The study conducted by Ali et al. (2021) found no significant improvement in students' ability to identify deepfakes after participating in workshops on the generation and dissemination of such content. Pre-test (mean [M] = 54.68) and post-test (M = 53.87) scores remained virtually unchanged. However, the research highlighted the value of GenAI in fostering critical thinking by providing technical knowledge about content manipulation on social media platforms. Following the intervention, students identified key criteria for assessing the credibility of information, such as the possibility of verifying sources (53.33%), the reliability of the sender (46.66%), and the knowledge of the author (26.66%), suggesting increased skepticism and a more reflective approach to consuming digital content.

Complementing these findings, Aprin et al. (2024) showed that the use of GenAI-based tools improved students' accuracy in judging altered images compared to those who completed the task without AI support. However, the improvement was less pronounced in cases where the image lacked sufficient online information, making the verification process more difficult.

Holistic improvement

Chen and Chang (2024) found that integrating ChatGPT with digital game-based learning (DGBL) significantly improved students' academic performance (M = 49.06 vs. M = 44.46) and reduced cognitive load (M = 3.24/3.32 vs. 3.74). While motivation levels were similar across groups, perceived competence was higher among those using ChatGPT (M = 2.63/2.54 vs. 2.21), indicating increased confidence. Additionally, these students displayed more reflective and strategic problem-solving behaviors, contrasting with the less analytical approach observed in the DGBL-only group.

On the other hand, the study conducted by Tamimi et al. (2024) found that 37.4% of students reported improvements in academic performance following the use of GenAI-powered tools. The most commonly cited uses of these tools included receiving immediate feedback on completed tasks (24.3%), enhancing comprehension of learning materials and accessing academic support (20%), as well as optimizing study time (16.5%). However, 44.3% of students indicated that they did not experience any significant changes in their personal productivity. Similarly, Lee et al. (2025) identified three interaction patterns with ChatGPT: full delegation of tasks, complementary use for comparison and improvement, and complete non-use. The first approach limited learning opportunities, while the latter two fostered deeper engagement, critical thinking, and ownership of the research process.

Linguistic improvement

The study by Hsiao and Chang (2024) demonstrated that the integration of GenAI-supported tools into real-world contexts enabled students to better understand their practical application and benefit from immediate feedback. This not only facilitated the acquisition of linguistic knowledge but also fostered greater interaction and engagement throughout the learning process. From another perspective, the research conducted by Kangasharju et al. (2022) showed that these tools can support students in composing poetry by providing draft versions, structural assistance, and guidance on the use of poetic elements. Moreover, GenAI was found to enable learners to experiment with new literary forms in a more motivating and less intimidating manner. A significant finding was that this support benefited all students equally, regardless of their academic

level, as no significant differences were observed in the improvements of written texts according to students' previous grades.

Taken together, the improvements reported across these categories are neither automatic nor uniform. They emerge most consistently in contexts where GenAI is embedded within structured pedagogical frameworks, supported by active teacher mediation, and used in ways that promote student agency rather than passive dependence on the tool. Conversely, when GenAI is used without clear learning objectives or critical guidance, its impact on learning tends to be limited or, in some cases, counterproductive. These findings underscore the importance of moving beyond questions of whether GenAI improves learning outcomes, towards a more nuanced understanding of the conditions under which such improvements occur.

Overall, these results indicate that GenAI is predominantly positioned as a pedagogical support tool that enhances efficiency, adaptability, and engagement, while simultaneously introducing new instructional, ethical, and organizational challenges that require careful mediation.

DISCUSSION

The emergence of GenAI in the educational field has marked an unprecedented turning point—not only due to its technological innovation capacity, but also because of the depth with which it impacts the dynamics of knowledge production, transmission, and assessment. Unlike previous technological waves that focused mainly on automating repetitive tasks, GenAI introduces creative, adaptive, and discursive functionalities. These capabilities have the potential to substantially reshape traditional pedagogical practices. This transformation affects both the structure of learning activities and the traditional roles attributed to teachers and students. At the same time, it generates new pedagogical, ethical, and organizational tensions and opportunities that are still under investigation.

In this rapidly changing scenario, it is imperative to understand how the integration of GenAI is being articulated within compulsory education, particularly at the secondary level. To this end, the main objective of the present systematic review was to analyze the current state of scientific knowledge regarding this integration, with a specific focus on both the concrete uses being implemented in school settings and the empirical evidence available concerning its effects on learning processes. The study was structured around two interrelated lines of inquiry. Firstly, it explored the specific ways in which teachers, educational institutions, or other actors involved in pedagogical practice are incorporating GenAI—whether in teaching activities, assessment, material development, or personalized student support. Secondly, it examined the impact of these applications on various dimensions of the educational process, such as academic achievement, the development of key competences, student motivation and engagement, and the promotion of critical thinking.

The analysis of the selected studies makes it possible to answer the RQs from a comprehensive perspective. The corpus includes studies with diverse methodological approaches, disciplinary backgrounds, and geographical contexts. This plurality has provided a representative overview of the ways in which GenAI is beginning to take root in Secondary Education, thus enabling a preliminary mapping of its level of classroom penetration and its transformative potential in both pedagogical and educational terms.

These findings should also be interpreted in light of the methodological quality of the available evidence. The quality appraisal conducted using the MMAT indicated that most empirical studies presented moderate to high methodological quality, although several relied on relatively small samples or exploratory designs. Consequently, while the reported results provide valuable insights into emerging pedagogical uses of GenAI, they should be interpreted cautiously, particularly when considering the generalizability of the reported learning improvements.

These patterns can also be interpreted through cognitive engagement frameworks such as ICAP (Chi & Wylie, 2014), which distinguish between passive, active, constructive, and interactive learning processes, suggesting that the pedagogical value of GenAI depends on the extent to which it promotes deeper levels of student engagement. Building on these results, the following discussion is structured around three key dimensions: uses and limitations of GenAI in secondary education, observed improvements in student learning, and implications for practice.

Uses and Limitations of GenAI in Secondary Education

Pedagogical uses of GenAI

The reviewed literature reveals diverse approaches to integrating GenAI in secondary classrooms, shaped by varying pedagogical frameworks, technological maturity, and levels of teacher involvement. Rather than a unified practice, its use spans from established applications to exploratory efforts with developmental potential.

A prominent use case is GenAI as a writing support tool, particularly in language and humanities. It is employed for composing texts, reformulating content, constructing arguments, and synthesizing ideas—often within revision-focused workflows that prompt students to make linguistic choices. In advanced instructional designs, teacher guidance enhances this process, fostering critical and reflective tool use.

Nonetheless, several limitations emerge. GenAI-generated texts may contain semantic errors, conceptual inaccuracies, contextual biases, and methodological gaps—especially in the absence of teacher oversight. Moreover, some tasks lack adequate educational contextualization, reducing their alignment with curricular goals and classroom relevance.

In the field of foreign language teaching and learning, GenAI is being used as a conversational partner, writing assistant, and linguistic corrector. The reviewed studies report significant improvements among students in pronunciation, fluency, writing, and motivation. However, they also reveal limitations related to the superficial nature of interactions and the one-sidedness of the communicative exchange.

In many cases, the responses generated are overly long, insufficiently adaptive, and offer limited opportunities for real-time correction, thereby diminishing the dialogic and formative nature of language practice. Furthermore, researchers have noted a tendency among students to accept GenAI's corrections or suggestions without question, which may negatively impact the development of linguistic judgment and critical autonomy.

In another widely represented area, GenAI is also used as an assistant in STEM tasks. Its use encompasses problem-solving, hypothesis development, and data interpretation. These experiences highlight the AI's ability to support logical reasoning, computational modeling, and result verification. While these functions can reduce cognitive load and promote autonomous learning, they also carry risks related to the reliability of the responses.

Generative models often produce answers that appear plausible but are incorrect when facing medium or high-complexity tasks, which can lead to misunderstandings or reinforce misconceptions. In subjects like mathematics or physics, for instance, a decline in performance has been observed as problem difficulty increases, along with errors in basic operations or in interpreting graphs.

To a lesser extent, though with innovative proposals, the use of GenAI has been identified in activities aimed at fostering creativity. These experiences include the development of interactive stories, graphic design, character creation, and scriptwriting. In these contexts, the tool becomes an ally of imagination, facilitating expressive and narrative processes that stimulate divergent thinking.

The pedagogical value of these proposals lies in their ability to connect students with alternative forms of symbolic representation, as well as in their integration with active methodologies such as project-based learning. However, their limited presence among the reviewed studies suggests that this potential is still far from being fully leveraged.

Another emerging field is that of GenAI-assisted assessment, both in its formative and summative dimensions. The tools analyzed have been used to grade written texts, provide personalized feedback, and conduct progress analysis based on student interactions. These functions offer advantages in terms of personalization, speed, and reducing teacher workload, and in some cases show accuracy levels comparable to those of human evaluators.

Challenges and institutional considerations for the integration of GenAI

Beyond its pedagogical potential, the integration of GenAI also raises a number of challenges related to its effective and responsible use in educational contexts. Despite its advantages, GenAI presents limitations

in evaluating complex discursive elements such as argumentation, metaphor use, and textual coherence. Inconsistencies in feedback have also been observed, particularly when the tool fails to adapt its suggestions to students' individual profiles or the cognitive demands of tasks. These limitations underscore the importance of using GenAI as a complement—rather than a substitute—for teacher judgment in assessment.

In addition, an emerging application of GenAI involves supporting teacher training and instructional planning. Generative models can assist in designing activities, lesson sequences, rubrics, and subject-specific explanations, functioning as cognitive aids for educators. While this may encourage pedagogical innovation, studies caution against overreliance on automated outputs. Uncritical adoption of GenAI-generated content may weaken instructional intent, lead to misaligned teaching strategies, and highlight the need for institutional guidance and professional development to support its appropriate use in classroom contexts.

Observed Improvements in Student Learning

Multidimensional learning gains associated with GenAI

The findings of this review indicate that GenAI can support student learning across multiple and interconnected dimensions, including cognitive, communicative, metacognitive, and motivational domains. Rather than producing isolated effects, these dimensions appear to reinforce one another, suggesting that GenAI may contribute to a more comprehensive transformation of learning processes.

At the cognitive level, the literature reports improvements in the understanding of complex concepts, problem-solving, disciplinary content learning, and analytical capacity across subject areas. Generative models facilitate access to content through multiple formats—such as explanations, examples, and visual representations—while also helping to reduce intrinsic cognitive load. This allows students to process new information more efficiently without overloading their mental resources.

In parallel, GenAI contributes to communicative and expressive development, particularly in written and oral skills in both first and second languages. These tools support key stages of the writing process—planning, drafting, and revision—as well as oral fluency, pronunciation, and vocabulary expansion. When used as a scaffold rather than a substitute, GenAI can foster more sophisticated forms of expression and support the development of communicative competence.

Metacognitive and critical thinking gains are also highlighted in the reviewed studies. In particular, when students are encouraged to evaluate, refine, or challenge AI-generated responses, GenAI can function as a catalyst for reflection, argumentation, and analytical reasoning. In this sense, generative models act as imperfect interlocutors that stimulate evaluation and judgment, supporting the development of information literacy and independent thinking.

At the motivational level, GenAI appears to enhance student engagement due to its immediacy, interactivity, and novelty. Increased participation and more positive attitudes toward learning tasks are commonly reported. However, these effects are not always sustained over time, as initial enthusiasm may diminish if activities are not supported by clear pedagogical objectives and meaningful task design.

However, these findings should be interpreted with caution, as several of the reviewed studies rely on small sample sizes, short intervention periods, or exploratory designs, which may limit the generalizability of the reported effects.

Conditions shaping the effectiveness of GenAI for learning

However, the improvements associated with GenAI are neither uniform nor guaranteed. Their effectiveness depends on a range of pedagogical, contextual, and individual factors, including task design, teacher guidance, institutional conditions, and students' levels of digital readiness. Evidence consistently suggests that the impact of GenAI is not determined by the technology itself, but by how it is integrated into instructional practices.

When GenAI is embedded within coherent pedagogical frameworks—aligned with curricular goals and supported by active teacher mediation—it can significantly enhance learning outcomes. In contrast, when used as a superficial add-on or disconnected from assessment and instructional design, its benefits tend to

be limited. These findings reinforce the importance of intentional, context-sensitive integration strategies to ensure that GenAI contributes meaningfully to student learning.

Practical Implications for the Responsible Implementation of GenAI in Secondary Education

Based on the findings of this review, several practical implications can be identified to support the responsible and pedagogically meaningful integration of GenAI in secondary education.

First, at the classroom level, the results emphasize that GenAI should be used as support for learning rather than a substitute for the educational process. The evidence suggests that its benefits are most evident when it is embedded within carefully designed activities, guided by explicit pedagogical objectives and active teacher mediation. In this context, teachers can use GenAI to support tasks such as generating alternative explanations, designing personalized exercises, or revising student work, while fostering students' critical engagement with AI-generated responses.

Second, from an institutional and policy perspective, the findings highlight the need for frameworks that guide the educational use of GenAI. This includes promoting targeted teacher training programs, establishing clear ethical guidelines for its use in learning and assessment contexts, and ensuring alignment with curricular goals and principles of educational equity. Effective integration depends not only on technological availability but also on institutional conditions that support reflective and pedagogically grounded use.

Finally, for developers of AI-based educational technologies, the results underline the importance of designing systems that are responsive to real classroom needs. GenAI tools should prioritize adaptability to different learning levels, transparency in content generation, and the provision of formative feedback. Equally important is the incorporation of mechanisms that help users identify potential errors, biases, or limitations in AI-generated outputs, thereby promoting more critical and responsible use in educational contexts.

LIMITATIONS AND FUTURE DIRECTIONS

This study presents several limitations that define the scope of its findings and highlight areas for improvement. These do not invalidate the results but do affect their generalizability. Firstly, the emerging nature of GenAI in secondary education—characterized by limited research and conceptual heterogeneity—requires the findings to be interpreted as part of an initial phase of development. A key methodological challenge was the inconsistent terminology in the literature, which often failed to distinguish between GenAI and traditional AI systems. This ambiguity necessitated a detailed review of each study's relevance and reflects a broader lack of consensus in the field.

Secondly, the dominance of short-term qualitative studies limits the ability to establish generalizable or causal conclusions. While such studies provide contextual insights, they fall short in assessing long-term impacts on student learning.

Another limitation relates to the methodological quality and heterogeneity of the empirical evidence available. Although the MMAT appraisal indicated that most studies met several quality criteria, a number of investigations relied on exploratory designs, small samples, or limited reporting of methodological procedures. These characteristics may affect the robustness and generalizability of some findings, reinforcing the need for more rigorous and large-scale research on the educational use of GenAI in secondary education.

Together, these limitations point to the need for clearer definitions, stronger methodological designs, and deeper analytical approaches. Recognizing these constraints is essential for an honest and critical interpretation of the emerging evidence.

Building on the findings of this review, several avenues for future research emerge that may contribute to a deeper understanding of the role of GenAI in secondary education. First, longitudinal studies are needed to examine how the impact of these tools on student learning evolves over time, as well as their potential effects on educational equity across students with varying levels of technological access, institutional support, and digital literacy.

Second, the results highlight the central role of teacher mediation and pedagogical design in the educational use of GenAI, suggesting the need to investigate effective models of teacher professional

development that prepare educators to integrate these technologies in a critical, reflective, and curriculum-aligned manner.

Finally, future research should further develop ethical and pedagogical frameworks for the use of GenAI in educational contexts, addressing issues such as the reliability of generated outputs, algorithmic transparency, authorship of AI-assisted knowledge production, and the role of these tools in fostering students' critical thinking.

CONCLUSIONS

This research has provided an up-to-date and critical overview of the use of GenAI in the field of Secondary Education, based on a systematic review of recent scientific literature. The analysis carried out makes it possible to identify a series of emerging trends regarding how this technology is being implemented in classrooms, as well as an emergent typology of pedagogical uses that, although diverse, reveals common patterns in objectives and methodological approaches.

Far from being a static technology, GenAI appears in the reviewed studies as a versatile tool whose effectiveness depends largely on the instructional frameworks in which it is embedded. As a writing assistant, conversational tutor, or catalyst for creative processes, it can offer significant educational value — but only when implemented with clear learning objectives, active teacher mediation, and explicit ethical and pedagogical criteria.

The findings show that, under certain conditions, the use of GenAI can lead to significant improvements in writing production, motivation, critical thinking, student autonomy, and the development of digital and metacognitive competencies. However, these improvements are neither automatic nor universal; they arise in contexts where technology is harnessed in support of intentional, reflective, and context-sensitive teaching.

Conversely, the results also highlight potential risks associated with superficial or decontextualized uses, the uncritical reproduction of generated content, or functional dependence on the tool—particularly when students are not provided with the necessary guidance and support.

Author contributions: HP-M: investigation, writing–original draft, and writing–review & editing; **DR-R:** writing–review & editing; & **AF-S:** writing–original draft and writing–review & editing. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: The authors stated that no ethical approval was required for this study.

AI statement: During the preparation of this work, the author(s) used Claude (Claude Opus 4.7, Anthropic) for the sole purpose of reformatting the in-text citations and the reference list in accordance with APA (7th edition) style requirements.

Declaration of interest: The authors declared no competing interests.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics, 2*, 431-432. https://doi.org/10.1007/s43681-021-00096-7

Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence, 2*, Article 100040. https://doi.org/10.1016/j.caeai.2021.100040

Aprin, F., Peters, P., & Hoppe, H. U. (2024). The effectiveness of a virtual learning companion for supporting the critical judgment of social media content. *Education and Information Technologies, 29*(10), 12797-12830. https://doi.org/10.1007/s10639-023-12275-6

Araújo, J. L., & Saúde, I. (2024). Can ChatGPT enhance chemistry laboratory teaching? Using prompt engineering to enable AI in generating laboratory activities. *Journal of Chemical Education, 101*(5), 1858-1864. https://doi.org/10.1021/acs.jchemed.3c00745

Banihashem, S. K., Bond, M., Bergdahl, N., Khosravi, H., & Noroozi, O. (2025). A systematic mapping review at the intersection of artificial intelligence and self-regulated learning. *International Journal of Educational Technology in Higher Education, 22*, Article 50. https://doi.org/10.1186/s41239-025-00548-8

Bhimdiwala, A., Neri, R. C., & Gomez, L. M. (2022). Advancing the design and implementation of artificial intelligence in education through continuous improvement. *International Journal of Artificial Intelligence in Education, 32*(3), 756-782. https://doi.org/10.1007/s40593-021-00278-8

Bulut, O., & Yildirim-Erbasli, S. N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education, 9*, 72-87. https://doi.org/10.21449/ijate.1124382

Carrasco Rodríguez, A. (2023). Reinventing the teaching of early modern history in secondary school: The use of ChatGPT to enhance learning and educational innovation. *Studia Historica: Historia Moderna, 45*(1), 101-145. https://doi.org/10.14201/shhmo2023451101146

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

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Ciampa, K., Wolfe, Z., & Hensley, M. (2025). From entry to transformation: Exploring AI integration in teachers' K-12 assessment practices. *Technology, Pedagogy and Education, 34*(2), 141-160. https://doi.org/10.1080/1475939X.2024.2413378

Clark, T. M., Fhaner, M., Stoltzfus, M., & Queen, M. S. (2024). Using ChatGPT to support lesson planning for the historical experiments of Thomson, Millikan, and Rutherford. *Journal of Chemical Education, 101*(5), 1992-1999. https://doi.org/10.1021/acs.jchemed.4c00200

Cohn, C., Snyder, C., Fonteles, J. H., TS, A., Montenegro, J., & Biswas, G. (2025). A multimodal approach to support teacher, researcher and AI collaboration in STEM+ C learning environments. *British Journal of Educational Technology, 56*(2), 595-620. https://doi.org/10.1111/bjet.13518

Dao, X. Q., & Le, N. B. (2023). LLMs performance on Vietnamese high school biology examination. *International Journal of Modern Education and Computer Science, 15*(6), 14-30. https://doi.org/10.5815/ijmecs.2023.06.02

de Souza, M. G., Won, M., Treagust, D., & Serrano, A. (2024). Visualising relativity: Assessing high school students' understanding of complex physics concepts through AI-generated images. *Physics Education, 59*(2), Article 025018. https://doi.org/10.1088/1361-6552/ad1e71

de Winter, J. C. (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

Díaz-Sánchez, C., & Chapinal-Heras, D. (2024). Use of open access AI in teaching classical antiquity. A methodological proposal. *Journal of Classics Teaching, 25(*49), 17-21. https://doi.org/10.1017/S2058631023000739

Doolittle, P., Wojdak, K., & Walters, A. (2023). Defining active learning: A restricted systemic review. *Teaching and Learning Inquiry, 11*, 1-24. https://doi.org/10.20343/teachlearninqu.11.25

Duan, H., & Zhao, W. (2024). The effects of educational artificial intelligence-powered applications on teachers' perceived autonomy, professional development for online teaching, and digital burnout. *The International Review of Research in Open and Distributed Learning, 25*(3), 57-76. https://doi.org/10.19173/irrodl.v25i3.7659

Ergene, O., & Ergene, B. C. (2025). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. *Education and Information Technologies, 30*(1), 509-545. https://doi.org/10.1007/s10639-024-13121-z

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). Sage Publications Ltd. https://doi.org/10.4135/9781036234942

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 Council of Europe Publishing. https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285-291. https://doi.org/10.3233/EFI-180221

Hsiao, J. C., & Chang, J. S. (2024). Enhancing EFL reading and writing through AI-powered tools: Design, implementation, and evaluation of an online course. *Interactive Learning Environments, 32*(9), 4934-4949. https://doi.org/10.1080/10494820.2023.2207187

Ifenthaler, D., & Schumacher, C. (2023). Reciprocal issues of artificial and human intelligence in education. *Journal of Research on Technology in Education, 55*(1), 1-6. https://doi.org/10.1080/15391523.2022.2154511

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages, 8*(4), Article 247. https://doi.org/10.3390/languages8040247

Kangasharju, A., Ilomaki, L., Lakkala, M., & Toom, A. (2022). Lower secondary students' poetry writing with the AI-based poetry machine. *Computers and Education: Artificial Intelligence, 3*, Article 100048. https://doi.org/10.1016/j.caeai.2022.100048

Khan, M. A., Kurbonova, O., Abdullaev, D., Radie, A. H., & Basim, N. (2024). Is AI-assisted assessment liable to evaluate young learners? Parents support, teacher support, immunity, and resilience are in focus in testing vocabulary learning. *Language Testing in Asia, 14*, Article 48. https://doi.org/10.1186/s40468-024-00324-x

Kong, S., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K-12 settings. *IEEE Transactions on Learning Technologies, 17*, 1562-1573. https://doi.org/10.1109/TLT.2024.3392830

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*, Article 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, M., Yi, T. R. J., Der-Thanq, C., Song, H. J., & David, H. W. L. (2025). Exploring interactions between learners and ChatGPT from a learner agency perspective: A multiple case study on historical inquiry. *Education and Information Technologies, 30*, 7167-7189. https://doi.org/10.1007/s10639-024-13114-y

Levine, S., Beck, S. W., Mah, C., Phalen, L., & Pittman, J. (2025). How do students use ChatGPT as a writing support? *Journal of Adolescent & Adult Literacy, 68*(5), 445-457. https://doi.org/10.1002/jaal.1373

Lu, W. Y., & Fan, S. C. (2023). Developing a weather prediction project-based machine learning course in facilitating AI learning among high school students. *Computers and Education: Artificial Intelligence, 5*, Article 100154. https://doi.org/10.1016/j.caeai.2023.100154

Lucas, M., Zhang, Y., Bem-Haja, P., & Vicente, P. N. (2024). The interplay between teachers' trust in artificial intelligence and digital competence. *Education and Information Technologies, 29*, 22991-23010. https://doi.org/10.1007/s10639-024-12772-2

Martínez-Comesaña, M., Rigueira-Díaz, X., Larrañaga-Janeiro, A., Martínez-Torres, J., Ocarranza-Prado, I., & Kreibel, D. (2023). Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review. *Revista de Psicodidáctica, 28*, 93-103. https://doi.org/10.1016/j.psicod.2023.06.001

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine, 27*(4), 12-12. https://www.semanticscholar.org/paper/A-Proposal-for-the-Dartmouth-Summer-Research-on-31,-McCarthy-Minsky/38e61d9a65aa483ad0fb4a219fe54d4e6a2f6c36

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record: The Voice of Scholarship in Education, 108*(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Murphy, R. F. (2019, January 23). *Artificial intelligence applications to support K-12 teachers and teaching: A review of promising applications, challenges, and risks*. RAND Corporation. https://doi.org/10.7249/PE315

Page, M. J., McKenzie, J. E., 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., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Parra, V., Sureda, P., Corica, A., Schiaffino, S., & Godoy, D. (2024). Can generative AI solve geometry problems? Strengths and weaknesses of LLMs for geometric reasoning in Spanish. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(5), 65-74. https://doi.org/10.9781/ijimai.2024.02.009

Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the test of understanding graphs in kinematics. *Physical Review Physics Education Research, 20*, Article 010109. https://doi.org/10.1103/PhysRevPhysEducRes.20.010109

Puentedura, R. B. (2006). *Transformation, technology, and education*. Hippasus. http://hippasus.com/resources/tte/

Radović, S. (2024). Is it only about technology? The interplay between educational technology for mathematics homework, teaching practice, and students' activities. *Journal of Computers in Education, 11*, 743-762 (2024). https://doi.org/10.1007/s40692-023-00277-9

Shikun, S., Grigoryan, G., Huichun, N., & Harutyunyan, H. (2024). AI chatbots: Developing English language proficiency in EFL classroom. *Arab World English Journal*, *15*(1), 292-305. https://doi.org/10.24093/awej/ChatGPT.20

Shin, D., Kim, H., Lee, J. H., & Yang, H. (2021). Exploring the use of an artificial intelligence chatbot as second language conversation partners. *Korean Journal of English Language and Linguistics, 21*, 375-391. https://doi.org/10.15738/kjell.21..202104.375

Tamimi, J., Addichane, E., & Alaoui, S. M. (2024). Evaluating the effects of artificial intelligence homework assistance tools on high school students' academic performance and personal development. *Arab World English Journal, 10*, 36-42. https://doi.org/10.24093/awej/call10.3

Tang, K. S., Cooper, G., Rappa, N., Cooper, M., Sims, C., & Nonis, K. (2024). A dialogic approach to transform teaching, learning & assessment with generative AI in secondary education: A proof of concept. *Pedagogies: An International Journal, 19*(3), 493-503. https://doi.org/10.1080/1554480X.2024.2379774

Tapalova, O., Zhiyenbayeva, N., & Gura, D. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *Electronic Journal of e-Learning, 20*(5), 639-653. https://doi.org/10.34190/ejel.20.5.2597

Vincent-Lancrin, S., & Van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD Education Working Paper No. 218,* 1-18. https://www.oecd.org/content/dam/oecd/en/publications/reports/2020/04/trustworthy-artificial-intelligence-ai-in-education_f1a7c415/a6c90fa9-en.pdf

Yeadon, W., & Hardy, T. (2024). The impact of AI in physics education: A comprehensive review from GCSE to university levels. *Physics Education, 59*(2), Article 025010. https://doi.org/10.1088/1361-6552/ad1fa2

Zhai, X., Nyaaba, M., & Ma, W. (2025). Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science? *Science & Education, 34*, 649-670. https://doi.org/10.1007/s11191-024-00496-1

APPENDIX A

Table A1. Results of the systematic review

Reference	Category	Subcategory	Participants	Findings
Ali et al. (2021)	Learning improvement	Deepfakes	38 students	Students developed awareness of the harmful potential of deepfakes and their role in the spread of misinformation.
Aprin et al. (2024)	Learning improvement	Deepfakes	22 students	An increase in the accuracy of judgements regarding the veracity of images in 57% of cases after reviewing recommendations and external sources.
Araújo and Saúde (2024)	Uses of GenAI	Question and activity creation	-	ChatGPT demonstrated capabilities in interpreting the symbolic language of chemistry and conceptualizing lab problems and activities. Limitations included confusion regarding chemical reactions and a lack of detail in safety protocols.
Bulut and Yildirim-Erbasli (2022)	Uses of GenAI	Story and narrative creation	-	The models generated texts and items reasonably well, but human evaluation and further adjustments are needed before they can be used with students.
Carrasco (2023)	Uses of GenAI	Story and narrative creation	-	ChatGPT was able to produce summaries, supplementary explanations, quizzes, practical exercises, and activity proposals, although some inaccuracies were observed.
Chen and Chang (2024)	Learning improvement	Holistic improvement	202 students	Students in the group that used examples provided by ChatGPT as a complement to DGBL achieved the highest post-test scores. They also reported a greater sense of competence and demonstrated more reflective and organized learning strategies. The DGBL group experienced higher intrinsic cognitive load, suggesting that the tasks were more difficult without AI assistance.
Ciampa et al. (2025)	Uses of GenAI	Assessment	-	Levels of integration included personalized learning, real-time feedback, and collaborative assessment. Challenges involved the time required to develop AI-related competencies, ethical concerns, and the need for ongoing teacher training. Benefits included the automation of administrative tasks, personalized assessments, and improved feedback for students.
Clark et al. (2024)	Uses of GenAI	Teacher training	4 teachers	The chatbot assisted instructors in structuring lesson plans, enhancing their content knowledge, and generating assessment questions. It served as a "thinking agent," encouraging reflection and the exploration of ideas.
Cohn et al. (2025)	Uses of GenAI	Assessment	4 students and 1 teacher	The multimodal timeline helped the teacher identify two key inflection points: Difficulty threshold when students encounter a challenge; Intervention point–the optimal moment to provide feedback. The teacher was able to observe productive pauses and signs of frustration, using this data to determine when to intervene.
Dao and Le (2023)	Uses of GenAI	Question, exercise, and exam solving	-	Bard achieved the highest average accuracy (69.5%) but showed variability in its responses. BingChat reached an average accuracy of 69.0%, with consistent performance. ChatGPT obtained the lowest accuracy (58%) but demonstrated high consistency in its responses.
de Souza et al. (2024)	Uses of GenAI	Image generation	10 students	Students with a stronger conceptual understanding created clearer prompts, resulting in images that more accurately reflected concepts related to relativity. Students with lower understanding struggled to articulate their ideas, leading to less representative images.
de Winter (2024)	Uses of GenAI	Question, exercise, and exam solving	-	GPT-4 achieved better results, with an average score of 8.3, outperforming GPT-3.5 in the assessments. ChatGPT-3.5 obtained an average score of 7.3, comparable to the average score of students in the Netherlands (6.99).
Díaz-Sánchez and Chapinal-Heras (2024)	Uses of GenAI	Image generation	-	Improved comprehension and development of digital competencies.
Duan and Zhao (2024)	Uses of GenAI	Teacher training	3,330 teachers	Significant increase in autonomy, improvement in professional development for online teaching, and reduction in digital burnout.
Ergene and Ergene (2025)	Uses of GenAI	Question, exercise, and exam solving	160 students & 80 teachers	GPT-4o showed the highest success rate with 87% of problems solved correctly, followed by GPT-4 (82%), MathGPT (50%), GPT-3.5 (36%), and Gemini (32%).
Hsiao and Chang (2024)	Learning improvement	Linguistic improvement	43 students	By incorporating AI-powered tools into real-world scenarios, students were able to see the practical application and benefits of these tools in their lives. By presenting their experiences with these tools to the class, they engaged in interpersonal interactions that resulted in the highest level of engagement.
Jiang et al. (2023)	Uses of GenAI	Assessment	41 students	GPT-4 excelled in accuracy (88%) but showed lower recall compared to the other models. iFLYTEK and GPT-3.5 demonstrated similar performance, with higher recall scores. The results highlighted discrepancies between the models and human evaluators, indicating areas for improvement in detecting grammatical structures and redundancies.

Table A1 (Continued).

Reference	Category	Subcategory	Participants	Findings
Kangasharju et al. (2022)	Learning improvement	Linguistic improvement	61 students	The structure of the poems evolved from drafts with "nonsensical" forms to descriptive-meditative or narrative styles. This tool enabled students to experiment with poetic features, fostering creativity and reducing the perception that writing poetry is difficult.
Khan et al. (2024)	Uses of GenAI	Assessment	60 students	The experimental group that used AI-assisted assessment outperformed the control group in vocabulary knowledge on the post-test. AI-assisted assessment significantly enhanced academic immunity* and resilience compared to the control group.
Kong and Yang (2024)	Uses of GenAI	Teacher training	31 teachers	Teachers reported an increased ability to design effective instructional materials, enhancing students' attention, relevance, confidence, and satisfaction.
Küchemann et al. (2023)	Uses of GenAI	Question and ctivity creation	26 participants	No significant differences were found in the quality of the tasks generated.
Latif and Zhai (2024)	Uses of GenAI	Assessment	1,200 students	GPT-3.5 showed an average increase of 9.1% in automated grading accuracy compared to BERT. For multi-label tasks, GPT-3.5 demonstrated significantly higher accuracy across all labels. In multi-class tasks, GPT-3.5 also outperformed BERT with an average accuracy increase of 10.6%.
Lee et al. (2025)	Learning improvement	Holistic improvement	3 students	Different patterns of ChatGPT use were identified: as a primary source of information, as a means of obtaining feedback and comparing information, and as a tool that was rejected by some students.
Levine et al. (2025)	Uses of GenAI	Writing support	12 students	Students interacted with ChatGPT primarily to plan, translate, and revise their texts. Most used the responses as starting points for their own arguments rather than copying and pasting directly
Meyer et al. (2024)	Uses of GenAI	Assessment	459 students	Students who received feedback generated by LLMs significantly improved their text revisions. There was an increase in motivation for future tasks and a rise in positive emotions such as enjoyment and pride. The feedback was perceived as helpful, although areas for improvement were identified.
Parra et al. (2024)	Uses of GenAI	Question, exercise, and exam solving	-	Only one of the responses generated by the chatbots matched the correct result, but it contained conceptual errors. Chatbots are not reliable for solving secondary-level geometry problems and may introduce or reinforce misconceptions.
Polverini and Gregorcic (2024)	Uses of GenAI	Question, exercise, and exam solving	-	ChatGPT achieved an average score of 41.7% correct responses, similar to that of secondary school students (47%). However, the distribution of its responses was much narrower (less variability) than that of the students. ChatGPT demonstrated adequate reasoning skills in 69.7% of responses. Its visual interpretation of graphs was correct in only 30.9% of cases.
Shikun et al. (2024)	Uses of GenAI	Foreign language learning	90 students	Significant improvement in pronunciation, intonation, and the ability to respond to questions in the target language. Notable progress in reading fluency. Increased motivation and confidence in language practice.
Shin et al. (2021)	Uses of GenAI	Foreign language learning	26 students	The chatbot serves as an effective conversational companion for learning another language.
Tamimi et al. (2024)	Learning improvement	Holistic improvement	115 students	Learning impact: 37.4% reported improved academic performance, 44.3% reported no significant changes, and 18.3% reported a decline in performance due to overreliance on the tools.
Tang et al. (2024)	Uses of GenAI	Question and activity creation	16 students and 4 teachers	Pedagogical outcomes: Students demonstrated critical skills in evaluating the relevance and accuracy of AI-generated responses. Active engagement was observed, reflecting critical thinking and creativity. A dialogic role with GenAI emerged, with students using it to support the co-construction of knowledge.
Yeadon and Hardy (2024)	Uses of GenAI	Question, exercise, and exam solving	-	ChatGPT showed variable performance depending on the educational level of the questions, with 83.4% and 63.8% accuracy rates. It also exhibited the following limitations: Mathematical calculations: A success rate of 45.2% in basic arithmetic operations. Conceptual errors: The model frequently produced plausible but incorrect answers.
Zhai et al. (2025)	Uses of GenAI	Question, exercise, and exam solving	-	ChatGPT and GPT-4 outperformed the majority of students in science ssessments. Their performance was not significantly affected by increasing levels of difficulty.

Note. *Academic immunity refers to students' capacity to maintain academic engagement and performance despite learning difficulties or academic challenges