Received: 12 August 2025 | Revised: 29 December 2025 | Accepted: 28 January 2026 | Published online: 12 February 2026

REVIEW

Artificial Intelligence and Applications
2026, Vol. 4(2) 149–164
DOI: 10.47852/bonviewAIA62027183
BON VIEW PUBLISHING

# Generative AI in Physics Education: A PRISMA-Guided Systematic Review of Empirical Studies

**Maria Moundridou[1,*], Nikolaos Moutis[1], Konstantinos Charalampopoulos[1] and Nikolaos Matzakos[1]**

*[1]Department of Education, School of Pedagogical and Technological Education, Greece*

**Abstract:** Generative artificial intelligence (GenAI) tools are redefining education, particularly in science disciplines such as physics. Even though these systems are widely used, there are still only a few studies that clearly summarize their use in teaching and their impact. In this PRISMA-guided systematic review, we analyzed 21 empirical studies (from 2023 to 2025) on the role of GenAI in physics education. We combined bibliometric mapping with content-based refinement and classified the studies by educational level, tools used, research methodology, and pedagogical role. Three main functions were identified: problem solver (57% of studies), student's tutor (38%), and teacher's assistant (33%). Problem-solving roles demonstrated variable accuracy, ranging from 100% to 0% across task types. Tutoring roles, though few, showed the strongest evidence of positive learning outcomes. Assistant roles mainly supported efficiency in grading and content generation, with less direct evidence of learning outcomes. The review contributes by providing this physics-specific PRISMA synthesis of GenAI in education and by identifying the roles GenAI tools assume in this context. This research is limited by the small, diverse set of included studies, which focus mainly on university contexts, and by its reliance on Scopus-indexed literature. Our findings suggest that GenAI tools can potentially improve conceptual understanding, formative assessment, and instructional efficiency. However, their effective integration requires human supervision, critical evaluation, and strategies to address issues of accuracy, equity, and multimodality.

**Keywords:** generative artificial intelligence, large language models, AI in education, physics education, systematic review

## 1. Introduction

Generative artificial intelligence (GenAI) tools, particularly large language models (LLMs) such as ChatGPT, are able to understand and respond to user prompts, simulating dialogue, and providing feedback. These features have attracted global attention for their potential to support and enhance student learning and teaching practices. Recent research explores the integration of GenAI in educational contexts for instructional design, content creation, automated feedback, and the facilitation of inquiry-based learning [1, 2].

Within science education—and physics in particular—GenAI systems may address several longstanding pedagogical challenges. Physics instruction demands that learners navigate abstract concepts, mathematical formulations, symbolic representations, and multistep problem-solving [3, 4]. Students must also shift between intuitive, everyday reasoning and formal scientific understanding [5, 6]. GenAI's capacity to offer timely feedback, translate between modalities, and personalize instruction suggests it could scaffold student thinking meaningfully—if integrated thoughtfully and ethically.

Recent studies have begun to explore how GenAI tools integrate into physics education across different learning environments and educational levels. These tools have been used in multiple pedagogical roles. As student tutors, they offer feedback, explanations, and guidance tailored to individual learners [7]. As teaching assistants, they can generate content, facilitate assessment, and support instructional design, helping reduce instructors' workloads [8]. As problem solvers, they simplify complex physics tasks and suggest solutions, though with varying reliability [9, 10]. These roles demonstrate GenAI's potential to support adaptive, personalized learning in physics. At the same time, they raise questions about the accuracy of responses, the transparency of its reasoning, and the maintenance of student agency [11, 12]. Understanding how these tools are implemented and their outcomes is essential for responsible and effective use.

Several systematic reviews have examined AI and GenAI in education, each with its own focus. Before GenAI became widespread, Xu and Ouyang [13] reviewed AI technologies in STEM education from 2011 to 2021, without physics-specific analysis. Similarly, Heeg and Avraamidou [14] focused on primary and secondary levels and studied AI in K-12 science education from 2010 to 2021. With the advent of GenAI, Batista et al. [15] conducted a broad systematic review of GenAI in higher education, without focusing on science disciplines. Aydin-Günbatar et al. [16] reviewed GenAI in all science education areas—chemistry, biology, and physics—but did not examine the unique challenges of each discipline. Although these reviews offer useful background, there remains a lack of a systematic examination of GenAI in physics education. Physics presents unique instructional and epistemological challenges: the coordination of multiple representations, the modeling of complex problems, and domain-specific misconceptions warrant dedicated attention. As GenAI adoption grows and research remains fragmented, a focused synthesis for physics education is both timely and necessary.

This research addresses that gap by reviewing empirical studies on GenAI in physics education using the PRISMA methodology. We analyze the use of these tools in physics learning environments, their effects on student outcomes and instructional practices, and identify emerging pedagogical trends. Based on 21 empirical studies, we summarize research on GenAI tools in physics education, highlight

[*]**Corresponding author:** Maria Moundridou, Department of Education, School of Pedagogical and Technological Education, Greece. Email: mariam@aspete.gr

opportunities and limitations, and provide recommendations for practice, policy, and future research.

## 2. Purpose of the Study and Research Questions

This study reviews empirical research from 2023 to 2025 on the use of GenAI tools in physics education. It examines how GenAI, particularly LLM-based tools, serves as a student tutor, teaching assistant, and problem solver. The study identifies common usage patterns, highlights effective practices, and provides guidance for future integration.

The review is structured around the following research questions:

**RQ1.** What are the characteristics of empirical studies that examine the integration of GenAI tools in physics education?

**RQ2.** In what functional roles are GenAI tools employed, and how are these roles operationalized across empirical studies in physics education?

**RQ3.** How are GenAI tools employed in the role of a student tutor in physics education, and what effects do they have on students' learning outcomes and perceptions?

**RQ4.** How are GenAI tools employed in the role of a teaching assistant in physics education, and what effects do they have on assessment, instructional support, and perceptions of their use?

**RQ5.** How are GenAI tools employed in the role of a problem solver in physics education, and what patterns of performance, reasoning, and limitations emerge from their application across different types of physics tasks?

## 3. Methodology

To ensure a systematic, transparent, and replicable approach in this fast-paced field, the present review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [17], which stresses clear filtering, documentation, and stepwise decision-making. In addition to traditional bibliographic searches and screening, we used bibliometric mapping and content-based refinement. These tools helped narrow the scope and maintain correspondence to the research objectives. The next section describes the eligibility criteria for including or excluding records to ensure that only relevant studies with academic integrity were selected for in-depth analysis.

### 3.1. Search strategy and data sources

The main data source for this systematic review was the Scopus database. The search was conducted using the following query terms: *("Generative AI" OR GenAI OR LLM OR ChatGPT OR "Artificial Intelligence") AND physics AND education*, on May 14, 2025.

The review followed the PRISMA methodology, which includes four key stages: identification, screening, eligibility assessment, and final inclusion. Building on this framework, the subsequent stages are described in detail below.

At the identification stage, all potentially relevant articles were retrieved through Scopus using the specified keyword combination. As illustrated in Figure 1, the publication trend shows a significant increase after 2023, with the majority of documents originating from the United States. Notably, 46.3% of these are classified as conference papers, indicating the field's emerging, fast-growing nature. In terms of subject area, most documents were in Computer Science, Engineering, Social Sciences, Physics, and Mathematics.

Figure 2 shows a keyword co-occurrence network based on 523 Scopus records. Only publications from 2023 onward are included. This period marks the rise of GenAI tools such as ChatGPT in education-related research. The visualization was created with the *Bibliometrix*

**Figure 1**
**Distribution of GenAI-related publications in physics education (Scopus corpus)**

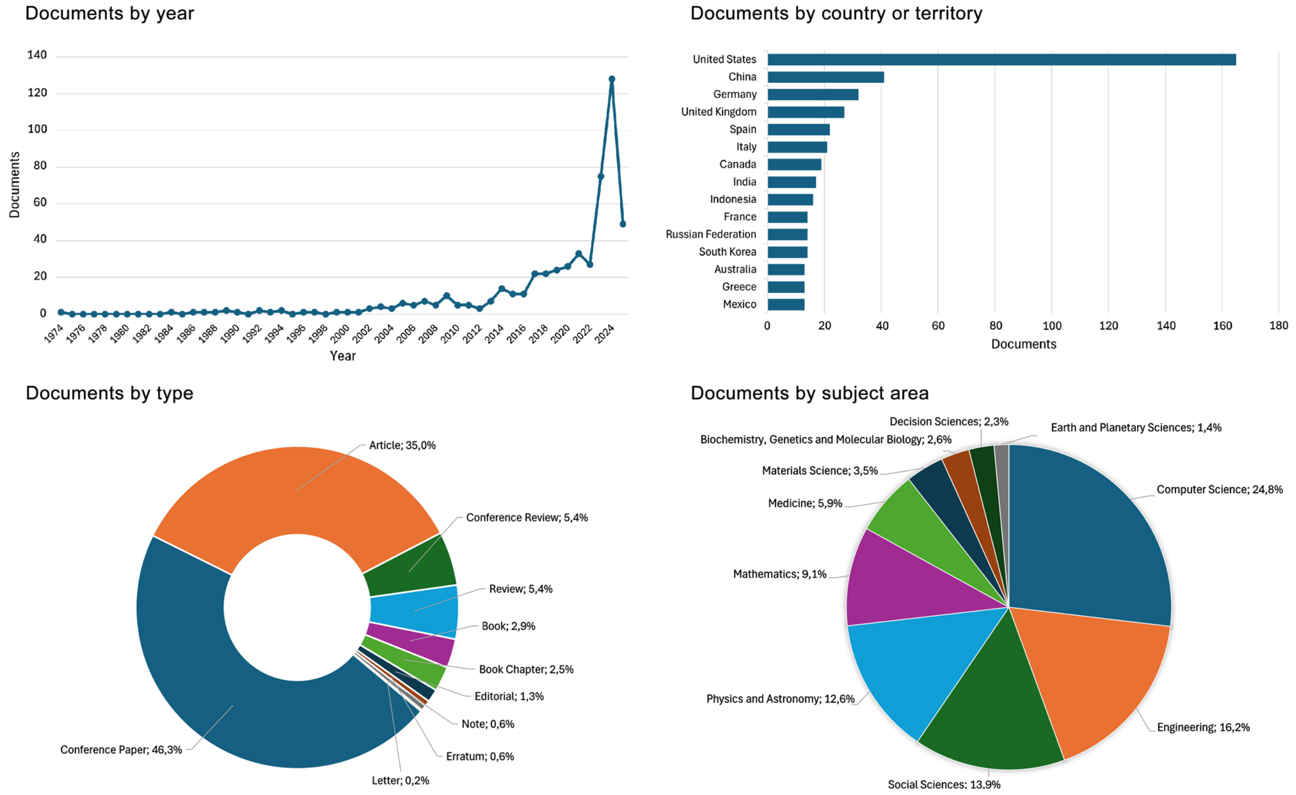
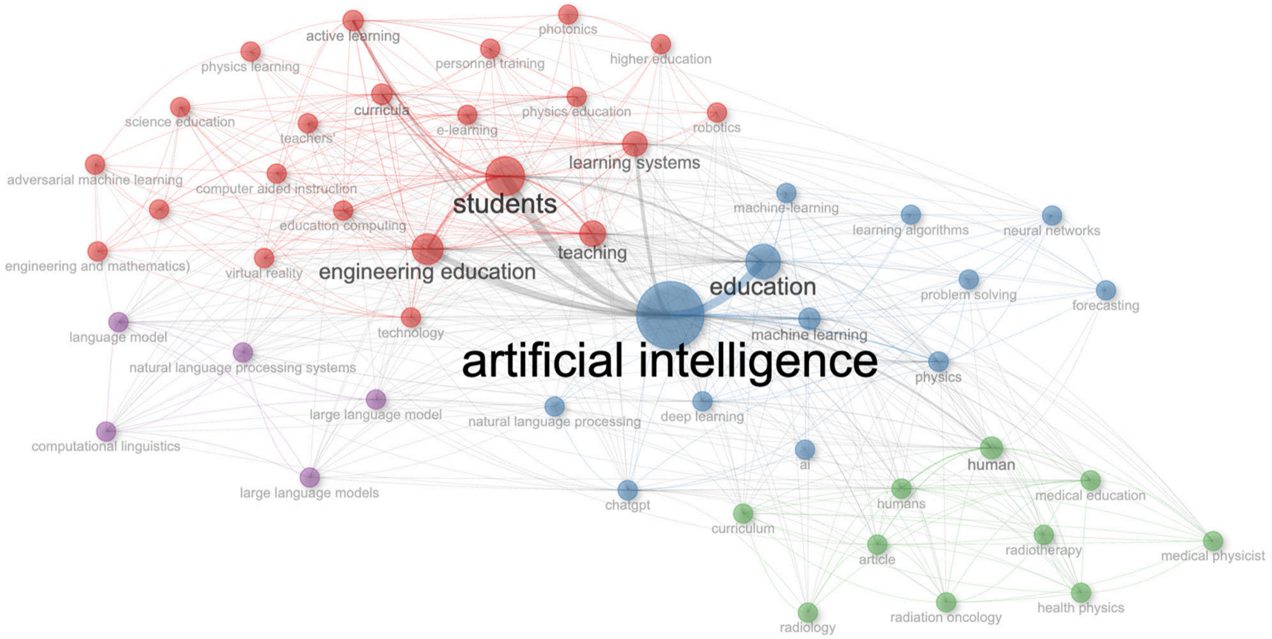


**Figure 2**
**Keyword co-occurrence network from the initial Scopus corpus (*n* = 523)**



R package [18]. It illustrates the corpus's semantic structure through keyword co-occurrences. Node sizes reflect term frequency. Link thickness shows co-occurrence strength between terms.

The term "artificial intelligence" sits at the center of the network, denoting its conceptual centrality across the dataset. Several major thematic clusters emerge from this core.

**Educational Cluster (Red):** This cluster connects terms such as "students," "engineering education," "physics education," and "teaching." It shows the dominant pedagogical discourse around GenAI tools in STEM education, focusing on instructional design and student-centered learning.

**Cognitive and Technological Cluster (Blue):** This cluster features terms like "ChatGPT," "natural language processing," "machine learning," "deep learning," "problem solving," and "physics." It shows how GenAI supports thinking, reasoning, and problem-solving in science education, especially physics.

**Applied Sciences Cluster (Green):** This cluster comprises terms like "medical education," "radiation oncology," and "health physics." It points to targeted GenAI use in clinical and health education.

**Linguistic Infrastructure Cluster (Purple):** This smaller cluster centers on terms such as "large language models," "natural language processing," and "computational linguistics." These terms represent the linguistic and algorithmic foundations of GenAI tools.

The network shows a strong link between AI and physics education. "Students," "teaching," and "active learning" are closely linked with "ChatGPT" and "physics." This structure guided the screening phase of the review and ensured that included studies were conceptually situated at the intersection of GenAI tools and physics learning.

Figure 3 shows a thematic map created from 523 Scopus records of studies published after 2023. The map uses Callon's centrality-density model to classify themes. The horizontal axis represents thematic relevance (centrality), while the vertical axis indicates internal development (density).

The cluster "artificial intelligence–education–learning systems" appears in the Basic Themes quadrant (bottom-right). These terms are conceptually central but show low internal development. This suggests that while they are widely used, their pedagogical implementations are still at an early or exploratory stage, especially in physics education.

The cluster "students–engineering education–active learning" sits between the basic and niche regions, suggesting relevance across STEM disciplines and moderate structural cohesion. This indicates growing pedagogical engagement with GenAI tools in active learning environments.

The "physics–humans–human" cluster lies in the middle-left region, indicating a specialized and possibly fragmented conceptual space. This may show that there are specific challenges in using GenAI for teaching physics, especially when it comes to modeling human-like reasoning.

Finally, "photonics–robotics–higher education" appears in the niche quadrant, suggesting these themes develop within isolated subdomains, possibly outside the core educational focus.

This mapping reinforces co-occurrence network findings by contextualizing the centrality of AI and education themes while indicating the emerging role of physics and student-centered learning in the evolving research landscape.

The keyword network and thematic map helped refine the review's scope. They guided the inclusion criteria and made sure the included studies matched the core pedagogical and disciplinary themes. The following section outlines the eligibility criteria and screening procedures used.

## 3.2. Eligibility criteria and screening procedure

Figure 4 shows the complete PRISMA flow diagram, which outlines the steps of identification, screening, and inclusion in the review process. Starting with 523 records from Scopus, we filtered them to exclude publications from before 2023, those not in English, and those listed as literature reviews or conference proceedings.

After automated filtering, 105 records were screened by title and abstract. Studies not related to physics education, including those on mathematics, medicine, or environmental science, were excluded at this stage (*n* = 47).

**Figure 3**
**Thematic map of keyword clusters (2023–2025), highlighting the conceptual positioning of GenAI-related themes in physics education based on their centrality and density**

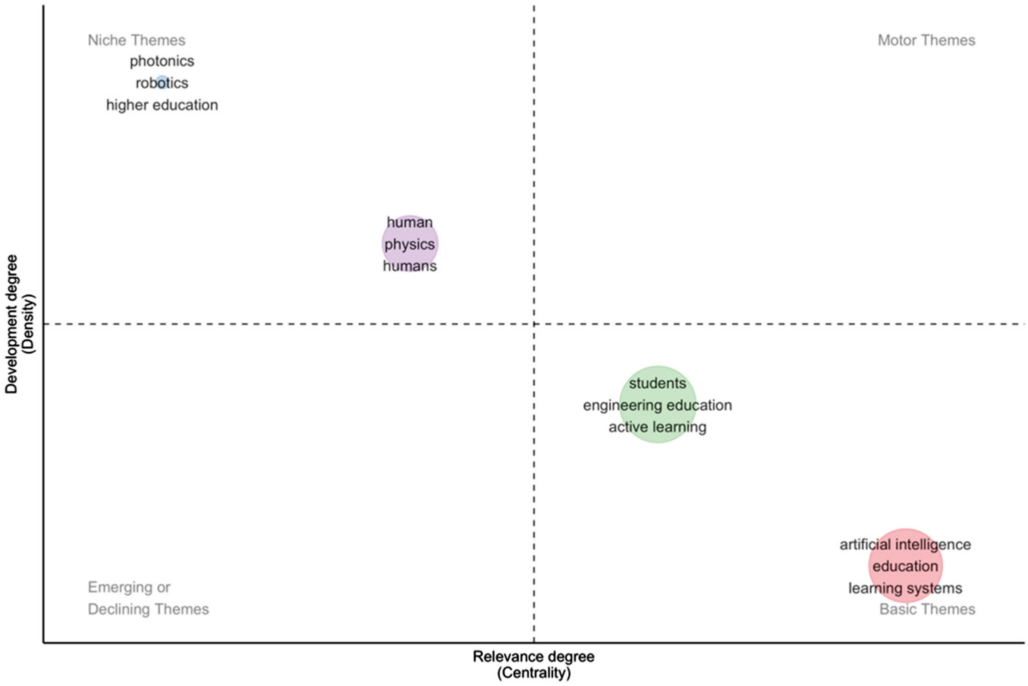
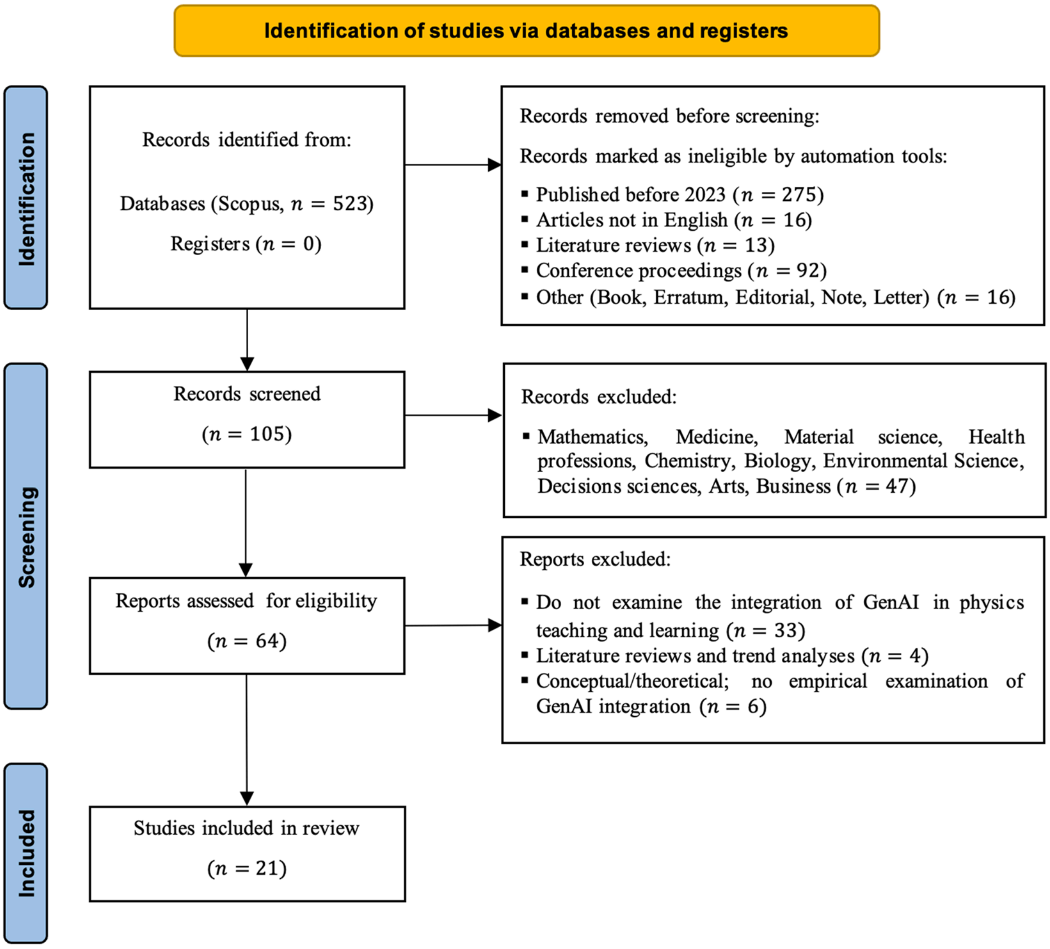


**Figure 4**
**Prisma flow diagram**


Identification of studies via databases and registers
Identification
Records identified from:
Databases (Scopus, $n$ = 523)
Registers ($n$ = 0)
Records removed before screening:
Records marked as ineligible by automation tools:
▪ Published before 2023 ($n$ = 275)
▪ Articles not in English ($n$ = 16)
▪ Literature reviews ($n$ = 13)
▪ Conference proceedings ($n$ = 92)
▪ Other (Book, Erratum, Editorial, Note, Letter) ($n$ = 16)
Screening
Records screened ($n$ = 105)
Records excluded:
▪ Mathematics, Medicine, Material science, Health professions, Chemistry, Biology, Environmental Science, Decisions sciences, Arts, Business ($n$ = 47)
Reports assessed for eligibility ($n$ = 64)
Reports excluded:
▪ Do not examine the integration of GenAI in physics teaching and learning ($n$ = 33)
▪ Literature reviews and trend analyses ($n$ = 4)
▪ Conceptual/theoretical; no empirical examination of GenAI integration ($n$ = 6)
Included
Studies included in review ($n$ = 21)

The remaining 64 records were then assessed for eligibility through full-text review. Studies were excluded if they did not examine the integration of GenAI in physics teaching and learning (*n* = 33), were secondary analyses or reviews (*n* = 4), or were conceptual/theoretical papers lacking empirical data (*n* = 6).

To be included, studies had to meet the following criteria:

1) Be peer-reviewed and published in English from 2023 onward
2) Focus explicitly on physics education
3) Present empirical findings on the use of GenAI tools (e.g., GPT-based systems, LLMs) in physics teaching or learning contexts

21 empirical studies met all eligibility criteria and were included in the analysis and synthesis.

## 3.3. Study selection and thematic coding

A structured extraction matrix was developed to capture study characteristics, research design, GenAI tools employed, educational levels, and measured outcomes. Thematic coding was then applied to categorize studies according to the pedagogical roles fulfilled by GenAI tools. Three recurring roles were identified: student tutor, teaching assistant, and problem solver, with some studies incorporating more than one role simultaneously. This categorization is reflected in Tables 1–5 and provides the framework for the comparative synthesis presented in Sections 4 and 5.

## 3.4. Conceptual orientation through bibliometric mapping

To support the conceptual alignment of selected studies, bibliometric visualizations were conducted in two phases using the Biblioshiny interface [18]. In the first phase, we analyzed the full set of 523 documents retrieved from Scopus; this analysis generated Figures 1–3, presented in Section 3.1. In the second phase, bibliometric techniques were applied to the final set of 21 included studies, resulting in Figures 5 and 6. In particular, Figure 5 presents a word cloud illustrating the most frequently occurring keywords across the 21 selected articles and reveals an emphasis on ChatGPT, large language models, problem solving, students, teaching, and curricula.

In addition to keyword frequency, the thematic cluster map in Figure 6 illustrates conceptual groupings and co-occurrence patterns across the 21 studies. It shows the internal thematic density and the centrality of research topics within the selected empirical corpus. Nodes indicate high-frequency keywords. Edges represent the strength of co-occurrence relationships, based on full counting. The layout and color coding use modularity-based clustering to highlight thematic proximities and conceptual groupings in the literature.

At the network's center, the terms "chatgpt" and "problem solving" emerge as conceptual hubs. They are tightly connected to both AI-specific terms (such as "genai", "gpt-4", and "chatbots") and educational constructs (such as "educational process", "context-rich problems", and "higher-order thinking skills"). This points to a growing interest in using GenAI tools to support problem-based learning and reasoning in science education.

The blue cluster highlights that "physics education" and students are part of a well-developed area of educational research. These concepts are closely linked to themes such as "curricula", "teaching", and "assessment", indicating that this field has a solid foundation. Within this context, GenAI tools are now starting to be integrated, extending existing approaches rather than replacing them.

The red cluster covers topics like "artificial intelligence", "ethical aspects", and "educational perceptions". It suggests there is a separate conversation about how AI affects society and classroom teaching.

Smaller clusters, like the green (e.g., "ai benchmark", "higher education") and the purple (e.g., "correlation", "ai in education"), show new or less central topics that could become main themes as the field develops.

**Figure 5**
**Thematic keyword cloud from the PRISMA-selected studies (*n* = 21)**

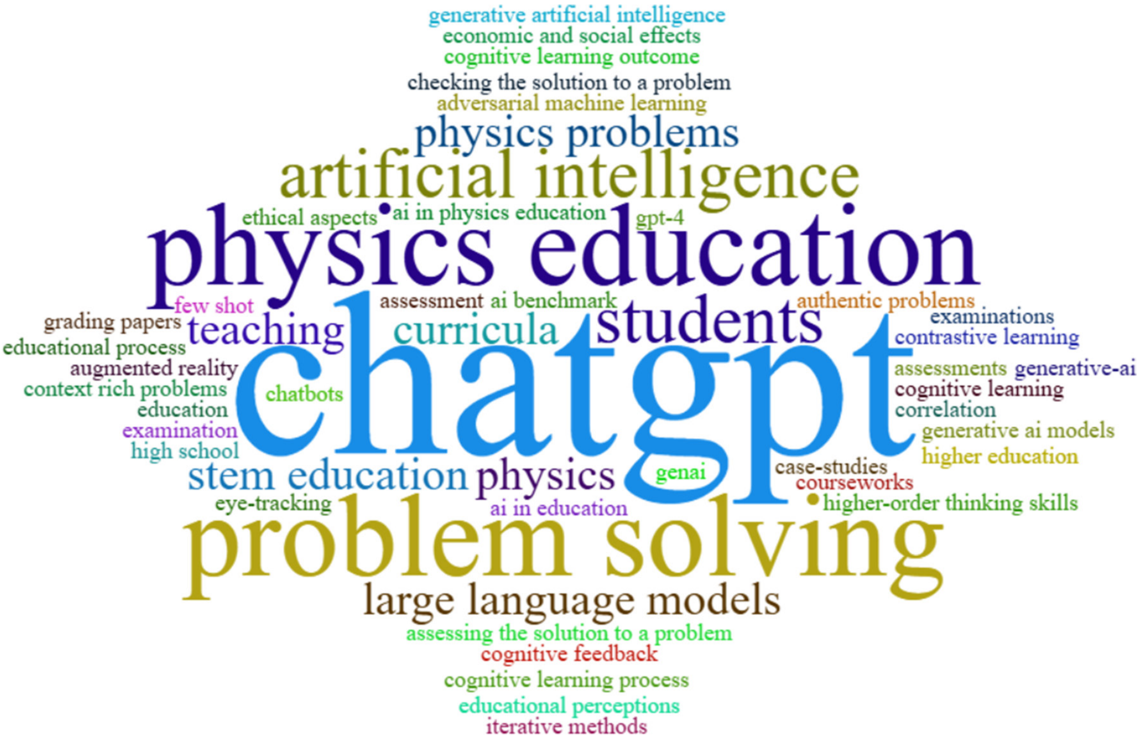
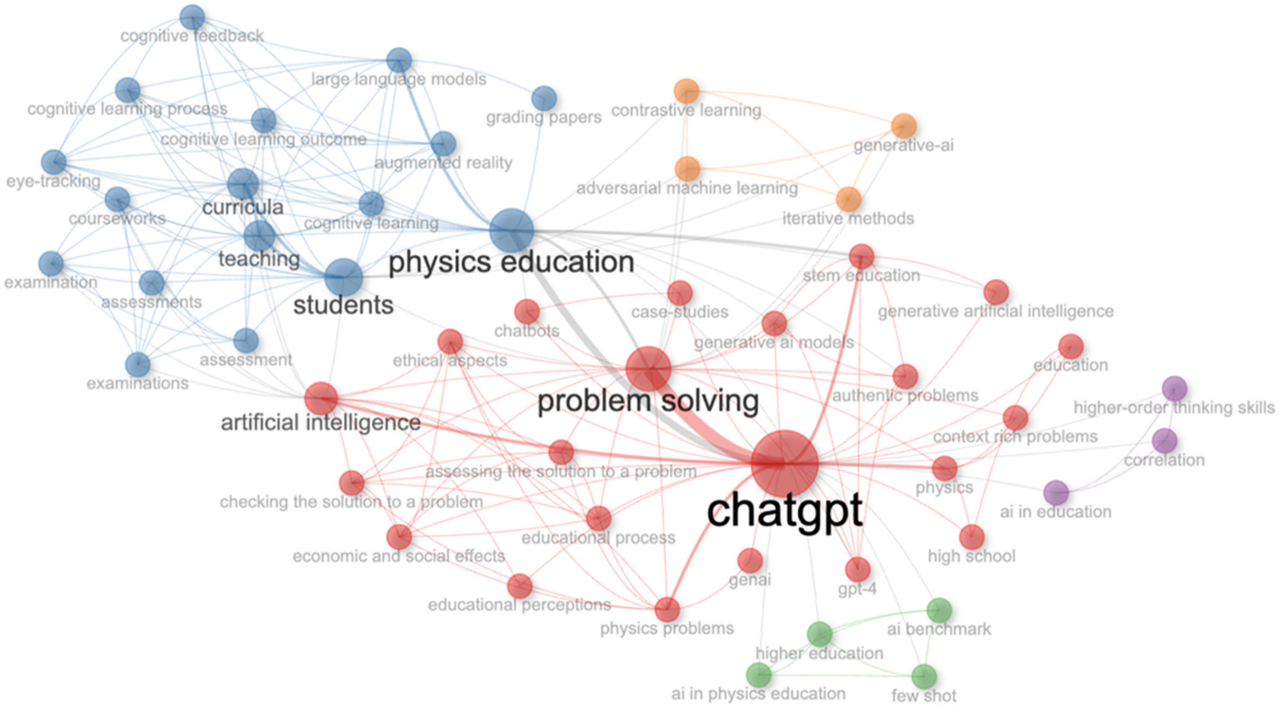


**Figure 6**
**Thematic cluster map based on the included studies (*n* = 21)**



The mapping shows that ChatGPT and problem-solving methods are central in recent research. It also highlights an emerging overlap between physics education, assessment methods, and GenAI tools.

The bibliometric analysis offers a conceptual framework for interpreting the main themes and structure of the selected corpus. This approach helps organize the synthesis presented in Section 4 and guides the analysis of pedagogical roles, impacts, and limitations discussed in Section 5.

## 4. Results

This section presents the findings of the systematic review in two stages. First, a descriptive analysis of the 21 included studies is provided, organized through a series of summary tables (Tables 1–5). These tables highlight study characteristics, the pedagogical roles of GenAI tools, and reported outcomes. Second, a cross-study synthesis is conducted to identify broader patterns and trends across the corpus, including aggregated role distributions, research designs, performance metrics, and participant perceptions.

### 4.1. Descriptive characteristics

Following the PRISMA screening process, 21 empirical studies were selected for inclusion in this review. Their features and outcomes are synthesized in Tables 1–5, each focusing on a distinct analytical dimension. Collectively, these tables capture the diversity of study designs, participant groups, educational levels, GenAI tools employed, and the pedagogical roles investigated. This descriptive mapping provides the foundation for the cross-study synthesis presented in Section 4.2.

Table 1 presents the main features of the studies included in this review. It lists the users, the course or context and educational level, the GenAI tools and their roles, and the research focus, design, and methods. This overview establishes a foundation for interpreting the thematic analyses in the following tables.

Table 1 gives an overview of the main features of the empirical studies reviewed. Most of these studies involved university participants, such as undergraduates, pre-service teachers, and faculty researchers. Fewer studies focused on high school or upper-secondary students. ChatGPT, especially versions 3.5 and 4, was the most widely used tool, though some studies also used newer models like GPT-4o, Gemini, and Claude. GenAI tools were mainly used as problem solvers, student tutors, or teacher's assistants, and sometimes filled more than one of these roles. The research mainly examined student learning outcomes, user perceptions, and GenAI performance on physics-related tasks. Most studies employed non-experimental methods, like performance tests or surveys, while a smaller number used experimental or quasi-experimental designs. In general, the table shows that research in this area is still developing, with most studies being exploratory.

Table 2 summarizes studies on GenAI tools as teaching assistants for content creation, assessment design, and grading. These studies address technical feasibility and pedagogical effectiveness.

The findings indicate that GenAI tools such as GPT-4 can grade accurately under controlled conditions. Still, there are issues with reading diagrams, keeping grading consistent, and getting reliable results from different models. Addressing these issues requires structured workflows, OCR integration, and human oversight.

Table 3 looks at the capacity of GenAI tools to autonomously solve physics problems. The included studies assess their performance on traditional exercises, multistep reasoning problems, and context-rich tasks.

Although GenAI tools can effectively address routine, well-defined problems, they face challenges with open-ended or complex problems that require modeling, assumptions, and context-based judgment. This shows the current boundaries of LLMs in higher-order scientific reasoning and suggests that GenAI should be used carefully in student problem-solving.

Table 4 presents studies in which GenAI acts as a student tutor to provide individualized feedback, scaffolding, and support for conceptual understanding during experimental or instructional tasks.

**Table 1**
**Descriptive characteristics and research designs of the empirical studies on GenAI in physics education**

| Authors | Users\| Educational level\| Course | Tools used | Role of GenAI tools | Research focus | Design\| Methods |
|---|---|---|---|---|---|
| Alarbi et al. [19] | Students\| High school\| Physics | ChatGPT | - Student's tutor | - Learning outcomes | Quasi-experimental\| Quantitative |
| Coban et al. [7] | Students\| University\| Quantum Cryptography | ChatGPT 4 | - Student's tutor | - Learning outcomes | Experimental\| Mixed-methods |
| Dahlkemper et al. [20] | Researchers\| University\| N/A | ChatGPT 3.5 | - Student's tutor | - Perceptions<br>- GenAI's performance | Non-experimental\| Quantitative |
| Ding et al. [9] | Students\| College\| Introductory physics | ChatGPT | - Student's tutor | - Perceptions<br>- GenAI's performance | Non-experimental\| Mixed-methods |
| Fadillah et al. [21] | Students\| Upper Secondary\| N/A | ChatGPT | - Student's tutor | - Perceptions | Non-experimental\| Quantitative |
| Fontao et al. [22] | Prospective Teacher Students\| Master's level \| N/A | ChatGPT | - Student's tutor<br>- Teacher's assistant | - Perceptions | Non-experimental\| Quantitative |
| Horchani [10] | Professors-Researchers\| University\| N/A | ChatGPT | - Problem-solver | - GenAI's performance | Experimental\| Performance-based (Quantitative) |
| Jang and Choi [23] | Teachers\| High school and University\| N/A | ChatGPT | - Problem-solver<br>- Student's tutor<br>- Teacher's assistant | - Perceptions<br>- GenAI's performance | Non-experimental\| Qualitative |
| Kieser et al. [24] | Professors-Researchers\| University\| N/A | ChatGPT | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Kortemeyer [25] | Professors-Researchers\| University\| N/A | ChatGPT | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Kortemeyer et al. [26] | Researchers\| University\| Thermodynamics | ChatGPT, MathPix | - Teacher's assistant | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Küchemann et al. [8] | Prospective Teacher Students\| University\| N/A | ChatGPT 3.5, Textbook | - Teacher's assistant | - Perceptions<br>- GenAI's performance | Experimental\| Mixed methods |
| López-Simó and Rezende [12] | Researchers\| Upper-secondary\| N/A | ChatGPT 3.5 | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Mok et al. [27] | Researchers\| Undergraduate\| N/A | GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro | - Teacher's assistant | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Pimbblet and Morrell [28] | Researchers\| Undergraduate\| N/A | ChatGPT-4 | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Quantitative) |
| Polverini and Gregorcic [29] | Researchers\| High school\| N/A | ChatGPT-4 | - Problem-solver | - GenAI's performance | Non-experimental\| Mixed-methods |
| Revalde et al. [30] | Researchers\| College/ University\| N/A | ChatGPT-3.5 | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Mixed Methods) |
| Shamshin [31] | Students\| University\| N/A | OPS (ChatGPT, Gemini, Copilot, Claude) | - Problem-solver<br>- Student's tutor<br>- Teacher's assistant | - Learning outcomes<br>- Perceptions | Experimental\| Mixed-methods |
| Sirnoorkar et al. [32] | Researchers /Students\| University\| N/A | ChatGPT 3.5-4o | - Problem-solver | - GenAI's performance | Non-experimental\| Mixed-methods |
| Wang et al. [11] | Researchers\| College\| N/A | ChatGPT | - Problem-solver | - GenAI's performance | Non-experimental\| Performance-based (Mixed Methods) |

**Table 1**
**(*Continued*)**

| Authors | Users\| Educational level\| Course | Tools used | Role of GenAI tools | Research focus | Design\| Methods |
|---|---|---|---|---|---|
| Yeadon and Hardy [33] | Researchers\| Upper-secondary, University\| N/A | ChatGPT 3.5 | - Problem-solver<br>- Teacher's assistant | - GenAI's performance | Non-experimental\| Performance-based (Mixed Methods) |

**Table 2**
**GenAI as teacher's assistant**

| Authors | Study's Research Questions/Objectives | Results |
|---|---|---|
| Kortemeyer et al. [26] | - AI support for grading handwritten physics exams | - LLMs can partially support grading if OCR converts handwriting accurately<br>- MathPix: effective for standard text; issues with diagrams, unusual layouts, poor handwriting<br>- Part-based workflow (GPT-4 + MathPix) outperformed human grading<br>- Rubric-based failed (memory limits); problem-based caused aggregation errors<br>- Precision ~95%, recall ~47% (often missed correct answers) |
| Küchemann et al. [8] | - Quality/type of tasks: ChatGPT vs textbook (prospective teachers)<br>- Task improvements made<br>- Teacher ratings: usability, usefulness, output quality | - ChatGPT tasks: correct difficulty, lower clarity/context than textbook; both groups weak on specificity<br>- Task types: ChatGPT → "Understand"; textbook → "Apply/Evaluate"<br>- 75% ChatGPT tasks used as-is (24% edited); textbook group mostly wrote original tasks<br>- ChatGPT edits improved clarity; some textbook edits introduced errors<br>- Ratings: high usability, neutral usefulness, slightly below neutral output quality |
| Mok et al. [27] | - LLM grading accuracy vs human markers<br>- Mark scheme effects<br>- Grading consistency across models | - Current state: unreliable without additional support; struggles with mathematical equivalence, hallucinations, error penalization<br>- With mark scheme: GPT-4 improved significantly, closer to human standards<br>- Consistency: GPT-4 most consistent, Gemini least |

**Table 3**
**GenAI as problem-solver**

| Authors | Study's Research Questions/Objectives | Results |
|---|---|---|
| Horchani [10] | - ChatGPT-4 performance: traditional vs context-rich problems (CRPs)<br>- Types/locations of reasoning failures | - Success rates: 77% traditional, 41% CRPs<br>- Main failures: modeling & solution planning stages<br>- Struggled with reasonable assumptions and valid physical models when data missing or real-world interpretation needed |
| Kieser et al. [24] | - ChatGPT conceptual understanding (force concept)<br>- Ability to simulate student cohort understanding<br>- Simulation of specific student preconceptions | - Force Concept Inventory: 83% accuracy (vs 14% real students)<br>- Cohort simulation: no significant response pattern variation<br>- Preconception simulation (impetus, centrifugal force): realistic variability, resembling actual student distributions |
| Kortemeyer [25] | - ChatGPT performance in standard university physics course<br>- Performance across assessment types<br>- Implications for instruction/assessment | - Overall: 53% (above pass, below graduation level)<br>- By type: 55% homework, 47% exams, 90% programming, 93% clickers<br>- Performance similar to struggling-but-passing student<br>- Implications: AI-enabled grade inflation concerns, need for AI-aware assessment, critical thinking/symbolic reasoning emphasis |

**Table 3**
**(*Continued*)**

| Authors | Study's Research Questions/Objectives | Results |
|---|---|---|
| López-Simó and Rezende [12] | - Question type effects on ChatGPT correctness/ variability | - Definition: 10/10 correct, low variability<br>- Simple calculation: 7/10 correct, minor errors<br>- Multistep calculation: 0/10 correct, major errors (trig, concepts, friction)<br>- Conceptual reasoning: 0/10 correct, ignored Galilean relativity<br>- Fermi problem: 3/10 close to magnitude, high variability, flawed reasoning |
| Pimbblet and Morrell [28] | - GPT-4 completion of UK undergraduate physics degree<br>- Performance variation across assessment formats<br>- Assessment design changes for academic integrity | - Overall: 65% (2.1 honors), unofficial pass (failed in-person lab & viva)<br>- Strong: programming, simple calculations, structured responses<br>- Weak: multi-step reasoning, complex problem-solving, diagram interpretation<br>- Recommendations: in-person assessments (exams, vivas, fieldwork, lab skills, oral presentations) |
| Polverini and Gregorcic [29] | - ChatGPT performance on kinematics graph interpretation<br>- Strengths/weaknesses analysis | - TUG-K test: performance comparable to HS students (avg. 41.7%), less variability, more consistent item-level patterns<br>- Strength: correct/appropriate strategies (69.7%)<br>- Weakness: visual graph interpretation (30.9%) |
| Revalde et al. [30] | - ChatGPT-3.5 accuracy: theoretical, practical, problem-solving questions<br>- Performance across four languages<br>- Comparison with human students | - English accuracy: 83% theoretical, 77% practical, 17% problem-solving<br>- By language: 83% English, 60% Latvian, 47% Russian, 33% Kazakh (major multilingual limits)<br>- Outperformed student average by ~8–9% on multiple-choice |
| Sirnoorkar et al. [32] | - Student vs AI responses: sensemaking & mechanistic reasoning features | - Sensemaking: AI showed all elements except noticing inconsistencies; students stronger on meta-cognitive reflection, iterative refinement<br>- Mechanistic reasoning: AI evidenced all seven codes; students stronger with diagrams/gestures (esp. when scaffolded) |
| Wang et al. [11] | - ChatGPT problem-solving variation by problem type<br>- Common failure modes<br>- Prompt engineering effects | - Success rates: 62.5% well-specified, 8.3% underspecified<br>- Failure modes: (a) incorrect physical modeling (forces, pivots), (b) unreasonable assumptions (density, friction), (c) calculation errors (arithmetic, trig)<br>- Prompt engineering: moderate but non-significant improvement; no impact on calculation errors |

The studies indicate that GenAI can positively impact students' performance by deepening their understanding. It is especially effective for complex physics concepts and provides individualized feedback that improves focus and learning. Students using ChatGPT often achieve higher test scores and develop higher-order thinking skills. This is aided by the tool's linguistic quality and ease of use. Nevertheless, GenAI exhibits notable limitations. It may occasionally provide inaccurate responses, particularly on more difficult questions, which may be overlooked by less knowledgeable students, leading them to absorb misinformation. Students may also become overly reliant on these tools and accept incorrect information without verification. In summary, while GenAI's integration can be beneficial for students, its accuracy and reliability need to improve to minimize the associated risks.

The studies in which GenAI tools function collectively as student tutors, teaching assistants, and problem solvers are summarized in Table 5. These studies examine both real-world applications and user perceptions across a wide range of educational functions.

These studies examined user perceptions, instructional utility, impact on student performance, and AI-supported grading. Overall, GenAI tools are regarded as accessible and useful for generating learning materials, facilitating inquiry-based learning, and reducing instructor workload. However, concerns were raised regarding reliability, overreliance, and the need for broader pedagogical adaptation and infrastructure to support effective integration.

## 4.2. Cross-study patterns and quantitative synthesis

Based on the descriptive summaries in Tables 1–5, we conducted a cross-study synthesis to identify broader patterns among the 21 included studies. This analysis aggregates study characteristics and reported outcomes to highlight convergences and divergences across the literature.

### *4.2.1. Pedagogical roles*

Three main functions of GenAI tools emerged: student's tutor, teacher's assistant, and problem solver. When counted inclusively (with studies coded under more than one role if applicable), 12 studies (57%) examined GenAI as a problem solver, 8 (38%) as a student's tutor, and 7 (33%) as a teacher's assistant. Four studies (Table 5) investigated multiple roles in combination, underscoring the versatility of GenAI tools in physics education. Solver roles were the most prevalent, reflecting interest in benchmarking the reasoning capabilities of

**Table 4**
**GenAI as student's tutor**

| Authors | Study's Research Questions/Objectives | Results |
|---|---|---|
| Alarbi et al. [19] | - ChatGPT impact on HS physics performance<br>- Comparison: ChatGPT vs traditional instruction<br>- Gender interaction effects | - Significant performance gains, esp. Newton's 2nd Law (large effect size)<br>- Higher post-test scores vs traditional methods<br>- Larger gains for female students |
| Coban et al. [7] | - LLM formative feedback → conceptual understanding in quantum cryptography lab<br>- ChatGPT feedback effects on visual attention patterns | - Better outcomes & response quality with feedback<br>- Eye-tracking: feedback directed attention to relevant experimental elements (virtual/physical components) |
| Dahlkemper et al. [20] | - Student ratings: ChatGPT vs expert solutions (scientific accuracy & linguistic quality)<br>- Impact of student expertise level<br>- Linguistic quality as confounding factor | - Linguistic quality rated similarly; scientific accuracy rated lower (esp. easier questions)<br>- Lower-knowledge students less able to detect inaccuracies<br>- Accuracy gap persists even controlling for linguistic quality |
| Ding et al. [9] | - ChatGPT accuracy on physics problems<br>- Student trust vs actual accuracy<br>- Trust influence on perceptions | - ~85% accuracy on exam questions<br>- ~50% of students trusted responses regardless of accuracy<br>- Higher trust → perceived ease of use, willingness to use |
| Fadillah et al. [21] | - Correlation: convenience/quality (CQ), motivation/engagement (ME), accuracy/trust (AT) vs higher-order thinking (HOTS)<br>- Relative predictive power | - Strong CQ-HOTS correlation (strongest predictor)<br>- Strong ME-HOTS correlation (2nd strongest)<br>- Positive AT-HOTS correlation (weakest predictor)<br>- Predictive hierarchy: CQ > ME > AT |

**Table 5**
**GenAI in multiple roles (student's tutor, teacher's assistant, or problem-solver)**

| Authors | Study's Research Questions/Objectives | Results |
|---|---|---|
| Fontao et al. [22] | - Prior ChatGPT knowledge among pre-service secondary teachers<br>- Student perceptions of the tool<br>- Perceptions as future teachers<br>- Humanities vs sciences comparison | - High awareness (98%), minimal training (2.8%)<br>- Perceived as easy/helpful; concerns: over-reliance, critical thinking erosion<br>- Useful for materials creation; concerns: plagiarism, authorship<br>- Science students more positive; humanities students more cautious |
| Jang and Choi [23] | - Physics educators' views on ChatGPT strengths/weaknesses for problem-solving<br>- Perceived educational applications<br>- Anticipated social/educational changes | - Strengths: high-level problem-solving, accessibility, interactivity<br>- Weaknesses: query-dependent, inconsistent, algorithmic/language limits<br>- Uses: self-directed learning, inquiry-based learning, teaching support, materials development, assessment, research/admin<br>- Concerns: need for curricular, methodological, infrastructure changes<br>- Predicted: paradigm shift, declining teacher authority, digital divide, AI literacy demands |
| Shamshin [31] | - OPS platform impact on problem-solving skills/performance<br>- Student perceptions of usefulness/effectiveness<br>- Instructor grading workload reduction | - 15–16% performance improvement (practical & lecture)<br>- 65% better understanding; 55% would recommend; 45% more motivated<br>- Grading time: 4–8 h/week → <1 h (automated evaluation) |
| Yeadon and Hardy [33] | - ChatGPT-3.5 accuracy across education levels (GCSE, A-Level, university)<br>- Prompting strategy effects (Zero Shot, ICL, Confirmatory)<br>- Self-evaluation reliability as grading assistant<br>- Mathematical calculation capabilities | - Performance by level: 83.4% (GCSE), 63.8% (A-Level), 37.4% (university); avg: 59.9%<br>- No significant differences between prompting methods<br>- Agreement with human markers: ~50%<br>- Adequate for basic operations; weaker on complex calculations |

LLMs. Tutoring roles, though fewer in number, provided the clearest evidence of learning benefits, even though several studies in this group emphasized perceptions rather than direct outcomes, while assistant roles concentrated on grading support and instructional material generation.

#### *4.2.2. Research design*

The field remains in an exploratory phase. Sixteen studies (76%) employed non-experimental designs, typically using performance-based benchmarking, perception surveys, or content analysis. Only four (19%) adopted experimental designs and one (5%) a quasi-experimental design. Most of these experimental investigations focused on tutoring applications, and those that assessed learning outcomes reported positive effects. Non-experimental studies, in contrast, provide breadth rather than causal evidence, mapping GenAI's accuracy, usability, and perceptions across a variety of contexts.

*4.2.3. Learning outcomes*

The studies reviewed found that measurable learning outcomes were primarily linked to the tutoring role of GenAI. Students who were supported by an LLM scored higher on post-tests, demonstrated better understanding, and their focus during lab activities improved. Of five experimental or quasi-experimental studies, two showed significant learning gains with large effect sizes [7, 19]. Another study [31] reported notable performance improvements without formal significance testing. Additional non-experimental studies found correlations between GenAI use and higher-order thinking skills, motivation, and engagement [21]. On the other hand, studies that examined GenAI as a problem solver focused on tool performance rather than student learning outcomes. Likewise, when GenAI was used as an assistant, it helped reduce grading time and provided practice tasks, but its impact on learning was not as well documented. Overall, evidence indicates that tutoring applications are most consistently effective in enhancing student performance.

*4.2.4. Performance accuracy*

Research on GenAI's ability to solve physics problems reports mixed outcomes. Routine, well-defined tasks yielded accuracy between 62% and 100%. These tasks included definitional questions, simple calculations, programming, and clicker-style assessments [12, 25]. However, accuracy dropped to between 0% and 41% on complex or context-rich problems, such as multistep calculations, conceptual reasoning, Fermi estimates, and tasks requiring visual or real-world modeling [11, 29]. GenAI commonly made errors such as selecting incorrect physical models, making unreasonable assumptions, or making arithmetic and trigonometric mistakes. Performance also varied across languages, ranging from 83% in English to 33% in Revalde et al. [30]. In summary, GenAI performs well on structured problems but struggles with complex, less clear, or multimodal tasks.

*4.2.5. Perceptions and attitudes*

Seven studies examined how GenAI tools are perceived by their users. Students frequently rated ChatGPT outputs as linguistically fluent and easy to use, with many reporting greater engagement and motivation. However, students with limited subject knowledge frequently failed to detect inaccuracies, which could lead to overreliance [20]. Teachers and prospective teachers saw GenAI as useful for generating materials and reducing workload, yet expressed concerns about plagiarism, loss of critical thinking, and inconsistent reliability [22, 23]. Researchers were more cautious, focusing on transparency, assessment integrity, and equity issues such as digital divides and multilingual performance gaps. Overall, users appreciated GenAI's usability but remained concerned about its accuracy and educational impact.

Taken together, this synthesis shows that GenAI is being tested across diverse pedagogical roles, with problem-solving functions receiving the greatest attention. Tutoring roles, though fewer, provide the clearest evidence of positive impact on learning, supported by the two experimental studies that reported statistically significant gains and one additional study showing performance improvements. Assistant roles are valued for efficiency but remain constrained by reliability concerns. Performance results highlight sharp contrasts between routine tasks, where accuracy ranged from about 62% to 100%, and complex problems, where accuracy often fell between 0% and 41%. Perceptions, reported in seven studies, reflect both enthusiasm for GenAI's usability and caution regarding its accuracy, reliability, and broader educational consequences. These quantitative patterns reinforce the exploratory but promising status of GenAI in physics education and supply a stronger empirical foundation for the interpretive analysis in Section 5.

## 5. Discussion

This section discusses the key findings of the systematic review through the lens of five refined research questions, based on the PRISMA framework and derived from the analysis of the selected empirical studies (see Table 1). In addition to the descriptive synthesis, the cross-study aggregation presented in Section 4.2 revealed several clear patterns: problem-solving functions were the most frequently examined (57% of studies), followed by student's tutor roles (38%) and teacher's assistant roles (33%). The vast majority of studies (76%) employed non-experimental designs, while among the five experimental or quasi-experimental investigations, two reported statistically significant learning gains and one additional study showed performance improvements when GenAI was used in a tutoring role. Accuracy outcomes also varied widely: routine, well-specified tasks were solved with success rates ranging from about 62% to 100%, but performance dropped to between 0% and 41% for multistep, under-specified, or context-rich problems. Perception findings, drawn from seven studies, consistently noted the linguistic fluency and usability of GenAI tools, while raising concerns about accuracy, reliability, and overreliance.

The observations indicate that GenAI is being implemented in three distinct pedagogical roles. Tutoring functions have the greatest impact. Assistant roles help with efficiency but are sometimes unreliable. Solver roles are less effective in complex problem contexts. The following discussion addresses thisevidence in detail through the five guiding research questions.

1) RQ1. What are the characteristics of empirical studies that examine the integration of GenAI tools in physics education since 2023?

As shown in Section 4.2, the included studies form a heterogeneous but converging body of early evidence, with clear patterns emerging in their scope, methods, and reported outcomes. These characteristics frame the more detailed examination of study designs, contexts, and pedagogical orientations that follows.

A total of 21 empirical studies on GenAI tools in physics education, published from 2023 to 2025, were included (Table 1). The majority of these are conducted in university-level settings and involve undergraduate students, pre-service teachers, or academic researchers. A smaller number focus on upper-secondary or high school contexts, examining students or teachers in those environments [19, 29]. Some studies are embedded in specific disciplinary areas, such as quantum cryptography [7] or thermodynamics [26], but most explore GenAI use in general or introductory physics contexts. In several instances, the studies do not specify a particular course, as their primary focus is on evaluating GenAI tools through performance-based tasks, content analysis, or perception studies rather than course-based interventions.

ChatGPT is the most widely used tool across the corpus, especially versions 3.5 and 4.0. Some studies use complementary systems, such as MathPix [26], or multimodal LLMs like GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro [27]. In one study, OPS, a platform that integrates multiple tools (ChatGPT, Gemini, Claude, and Copilot), is used to deliver a broader interface for learner interaction and instructor support [31].

Pedagogically, GenAI tools serve as student tutors, teaching assistants, and problem solvers. Some studies explore their ability to provide individualized guidance or feedback [7, 19]. Others examine their role in instructional planning, assessment support, or workload reduction [8, 31, 33]. Much of the literature focuses on GenAI's problem-solving, assessing its responses to physics questions of different difficulty levels [24, 32]. This focus highlights the emphasis on validating the tools' reasoning and the quality of their solutions.

Research foci align with these roles. Numerous studies evaluate the performance of GenAI tools on physics-related tasks [23]. Other investigations examine learner and instructor perceptions, focusing on usability, trust, and perceived pedagogical value [22]. Learning outcomes are also addressed, though by a smaller subset of studies [19, 31], often in conjunction with perceptions and tool functionality. Some research combines focus areas, evaluating both system accuracy and student attitudes. This reflects the field's experimental and exploratory orientation.

Regarding research design, most studies are non-experimental ($n$ = 16) and frequently use performance-based evaluations (e.g., problem-solving accuracy), survey instruments, or content analysis. Mixed-methods that combine quantitative accuracy data with qualitative reflections or perception ratings are widely used [9, 30]. Only a few studies follow experimental ($n$ = 4) or quasi-experimental ($n$ = 1) approaches. These usually examine the effects of GenAI tools on student learning or task performance [7, 19]. Often, they use pre- and post-assessments or compare control and intervention groups. This trend indicates a shift toward more rigorous evaluation as the field develops.

The reviewed studies show a growing, yet still exploratory, interest in GenAI's instructional potential in physics education, with ChatGPT serving as the primary platform for early trials. Most evidence comes from university settings, which may not reflect the unique challenges and opportunities present in secondary education, where conceptual and motivational barriers are often greater. The focus on performance benchmarking and perception surveys suggests the field is still assessing technical feasibility rather than integrating GenAI into ongoing instructional practice. This trend is also evident in the keyword cloud in Figure 5, where terms such as "ChatGPT," "problem solving," "physics education," and "students" are most prominent, highlighting the narrow scope of current research. The absence of longitudinal studies and curriculum-level integration reveals a significant gap: without moving beyond short-term trials, it is uncertain whether current positive results will lead to lasting learning gains. The field appears to be at a transitional stage, shifting from proof-of-concept studies to the need for systematic, classroom-based evaluation.

2) RQ2. In what roles are GenAI tools employed, and how are these roles operationalized across empirical studies in physics education?

As outlined in Section 4.2, three recurring pedagogical functions of GenAI were identified: student's tutor, teacher's assistant, and problem solver. RQ2 examines how these roles are operationalized across individual studies, as detailed in Tables 2–5.

As student tutors, GenAI tools can guide learners through complex problems, provide step-by-step feedback, and support personalized learning. These applications are designed to enhance conceptual understanding, facilitate reasoning, and foster metacognitive awareness. For instance, Alarbi et al. [19] and Coban et al. [7] used ChatGPT to help students solve physics problems, which led to measurable improvements in their performance. Table 4 shows how these tutoring functions are embedded in structured tasks. GenAI typically delivers just-in-time support, often as explanatory feedback or scaffolded dialogue. Additional studies assess this tutoring function based on students' perceptions of helpfulness, engagement, and conceptual understanding [21, 22], although they do not always measure direct learning gains.

In the capacity of teaching assistants, GenAI tools are used to generate instructional content and support assessment and grading practices. In studies such as Mok et al. [27] and Yeadon and Hardy [33], GenAI served as an automated grader or question generator. In other studies, like Küchemann et al. [8], tools supported prospective teachers in planning and reflecting on instructional tasks. Research on this role often notes the potential to reduce workload, but also raises concerns about the accuracy, consistency, and pedagogical alignment of AI-generated content (Tables 2 and 5).

The problem-solver role, outlined in Table 3, is the most frequently investigated among the reviewed studies. In this role, GenAI tools are tasked with independently solving physics problems or producing reasoning sequences intended to emulate expert problem-solving. Studies such as Kieser et al. [24], Wang et al. [11], and López-Simó and Rezende [12] assessed the accuracy, logical coherence, and instructional usefulness of these outputs. Results generally show high performance on algorithmic or well-structured questions but reveal persistent weaknesses in tasks requiring modeling, contextual reasoning, or interpretation of abstract representations. Some studies use benchmark problem sets to compare AI-generated solutions with those of human experts or students [29]. Others assess the conceptual validity and linguistic clarity of AI-generated explanations [30] help measure performance and show the strengths and weaknesses of GenAI reasoning in physics.

In several studies, these roles overlap (Table 5). For example, Shamshin [31] describes a platform where GenAI acts as tutor, assistant, and evaluator. This reflects a broader trend toward hybrid applications, where GenAI tools fulfill multiple instructional roles within a single environment, instead of being limited to one function.

This operational diversity is evident in the thematic cluster map in Figure 6, where "ChatGPT" and "problem solving" are central, high-density nodes linked to educational constructs such as "physics education," "curricula," and "teaching." These links illustrate GenAI's broad integration across its three main roles. Separate clusters on "ethical aspects" and "educational perceptions" indicate that these functions are part of wider socio-technical concerns. The evidence shows that GenAI's roles in physics education are still changing and often overlap, with more hybrid uses appearing. The field is moving toward a more integrated view of GenAI as a multi-purpose mediator of learning and instruction. However, its educational impact will depend on how well technical capabilities align with pedagogical and ethical needs.

3) RQ3. How are GenAI tools employed in the role of a student tutor in physics education, and what effects do they have on students' learning outcomes and perceptions?

Studies summarized in Table 4 examine cases where GenAI tools serve as student tutors, offering explanations, stepwise problem guidance, and individualized feedback to support physics learning. These interventions consistently yield positive effects on student achievement. Alarbi et al. [19] found that students who used ChatGPT to learn Newton's Second Law achieved significantly higher post-test scores than those in traditional classrooms, with female students showing particularly strong gains. This suggests the potential to reduce performance disparities. In a university quantum cryptography lab, Coban et al. [7] reported that personalized ChatGPT feedback improved conceptual understanding, response quality, and visual attention to key experimental features, as shown by eye-tracking data. Ding et al. [9] noted that ChatGPT correctly answered about 85% of exam-style questions. However, many students relied on its outputs regardless of accuracy, suggesting a risk of overreliance. Similarly, Dahlkemper et al. [20] observed that students rated ChatGPT's linguistic quality as similar to expert solutions but considered its scientific accuracy lower, particularly among those with stronger physics backgrounds. Fadillah et al. [21] identified perceptions of convenience, motivation, and trust as predictors of higher-order thinking skills, with convenience being the most influential factor. Additional insights are provided by the studies in Table 5, which integrate tutoring functions with roles such as teaching assistant or problem solver. Shamshin [31] documented a 15–16% improvement in both practical and lecture-based performance among students using a multi-tool AI platform that combined tutoring with automated evaluation and content generation.

Overall, the findings show that using GenAI as a tutor, either alone or with other supports, can result in measurable improvements in physics learning, including higher post-test scores, enhanced understanding, and greater student engagement. However, effectiveness depends on students' prior knowledge, their ability to critically evaluate GenAI's outputs, and the presence of structured guidance. Therefore, GenAI tutoring should complement teacher-led instruction by providing personalized support, while educators remain responsible for ensuring accuracy and fostering critical thinking.

4) RQ4. How are GenAI tools employed in the role of a teaching assistant in physics education, and what effects do they have on assessment, instructional support, and perceptions of their use?

Teaching assistant roles focus more on efficiency and reducing workload than on direct student learning. Table 2 highlights how GenAI helps educators with assessment design, grading, and content creation. Kortemeyer et al. [26] studied automated grading of handwritten physics exams using GPT-4 and MathPix OCR. They found the system was highly precise, correctly identifying about 95% of answers, but it often missed valid responses, with a recall of about 47%. The best results came from part-based grading, while rubric-based grading was limited by the model's memory. Mok et al. [27] also examined AI-assisted grading of undergraduate physics problems. They found that GPT-4 was more accurate when using a mark scheme, but it still struggled with mathematical equivalence and missed major errors. Küchemann et al. [8] explored how prospective physics teachers used AI to generate tasks. They found that ChatGPT-created problems were usually correct and challenging enough, but they were less clear and lacked the context found in textbook questions. Most of these AI-generated tasks were used with only minor changes, showing they were easy to use but not often critically revised.

Table 5 presents perspectives where teaching assistant functions are combined with tutoring or problem-solving ones. Shamshin [31] reported that the OPS platform, which uses automated AI grading and other supports, reduced instructor grading time from 4–8 h per week to less than 1 h. In the study of Yeadon and Hardy [33], ChatGPT-3.5 was used as an automated grading assistant for GCSE, A-Level, and university physics exams. They found that it agreed with human markers about half the time, with no major differences between prompting strategies. Fontao et al. [22] observed that future secondary teachers appreciated ChatGPT for creating materials and supporting tasks, but were concerned about plagiarism and reduced critical thinking. Jang and Choi [23] outlined GenAI uses proposed by physics educators, such as process-based assessment, research, and administrative work, as well as their caution against overreliance and the need for careful curriculum planning.

Overall, the studies indicate that using GenAI tools as teaching assistants can reduce instructor workload and make assessment and content creation more scalable. Still, these benefits depend on careful implementation, as several factors, such as rubric clarity, task complexity, and input data quality, can affect the tools' reliability. Educators express both enthusiasm for increased efficiency and concerns about transparency, reliability, and the potential weakening of professional judgment. Based on these findings, it is recommended that GenAI be used to support, not replace, teacher expertise, and that institutions create workflows that combine GenAI's efficiency with safeguards to ensure validity, fairness, and professional oversight.

5) RQ5. How are GenAI tools employed in the role of a problem solver in physics education, and what patterns of performance, reasoning, and limitations emerge from their application across different types of physics tasks?

Studies in Table 3 position GenAI as an autonomous problem solver, tasked with generating complete solutions and reasoning sequences for a variety of physics problems. Tasks range from traditional exercises to context-rich scenarios, conceptual reasoning, and specialized assessments. Performance outcomes significantly depend on task type and complexity. Horchani [10] found that GPT-4 performed strongly on traditional problems (77% correct) but less successfully on context-rich ones (41%), with most failures occurring at the modeling and solution-planning stages. Similarly, Wang et al. [11] reported an accuracy rate of 62.5% for well-specified problems but only 8.3% for underspecified problems, and noted as frequent issues the incorrect physical modeling, unreasonable assumptions, and calculation errors.

Other studies explored domain-specific capabilities. Kieser et al. [24] showed ChatGPT could achieve scores exceeding those of actual student cohorts and simulate common student preconceptions, though it did not fully replicate the variability of real learners. Polverini and Gregorcic [29] found ChatGPT performed comparably to high school students in interpreting kinematics graphs but struggled with visual interpretation tasks, even when it could propose correct solution strategies. López-Simó and Rezende [12] demonstrated high accuracy on definitional and simple calculation items but a complete breakdown on multistep and conceptual reasoning questions. Performance on real-course assessments also varied. Kortemeyer [25] reported ChatGPT's aggregate university course score was above the pass level but inconsistent across assessment types. Pimbblet and Morrell [28] showed GPT-4 achieved an upper-second degree average in UK physics but failed practical and oral components. Revalde et al. [30] added a multilingual perspective, finding accuracy dropped sharply outside English, especially for complex problem-solving tasks.

Table 5 includes studies where problem-solving roles are part of broader implementations. Jang and Choi [23] noted that physics educators valued ChatGPT's ability to tackle advanced problems and support inquiry-based learning, but expressed concerns about its dependence on prompt quality, inconsistent reliability, and language limitations. Yeadon and Hardy [33] assessed ChatGPT-3.5's performance on physics exam questions at various educational levels. They found strong results in GCSEs, moderate performance in A-Levels, and weaker outcomes at the university level. They also reported no significant improvement from alternative prompting strategies and noted that mathematical precision decreased as problem complexity increased.

Overall, the findings indicate that GenAI problem-solving tools are effective as assistants for routine tasks but remain unreliable for complex reasoning. Their speed, consistency with structured problems, and ability to mimic student reasoning make them useful for practice, diagnostics, and generating alternative solutions. However, their frequent failures in modeling, visual interpretation, and contextual reasoning highlight a major limitation: current systems cannot replicate the reasoning processes of expert physicists. Therefore, these tools should be used as supplementary resources for exploration and feedback, not as autonomous solvers. Successful integration will require embedding them in human-guided tasks that emphasize critical thinking, model evaluation, and conceptual scaffolding.

## 6. Conclusions

This PRISMA-guided systematic review synthesized 21 empirical studies (2023–2025) examining the integration of GenAI tools in physics education. Three primary pedagogical roles emerged: student tutor, teacher assistant, and problem solver. Problem-solving functions were the most frequently investigated (57% of studies), followed by student tutor roles (38%) and teacher assistant roles (33%). This distribution indicates a research focus on evaluating reasoning capabilities and developing new applications for tutoring and instructional support. As student tutors, GenAI systems can improve conceptual understanding and engagement by providing adaptive explanations, personalized feedback, and real-time scaffolding. As teacher assistants, these systems can streamline formative assessment and content creation, thereby reducing workload and supporting instructional design. As problem solvers, GenAI tools often perform competitively on well-structured tasks, sometimes matching or surpassing average student performance, but they continue to face challenges with complex, context-rich problems.

A key finding of the synthesis is that tutoring applications show the strongest evidence of positive impact on learning outcomes. Among the five experimental or quasi-experimental studies, two reported statistically significant learning gains with large effect sizes, while an additional study observed performance improvements without formal significance testing. In the case of problem-solving roles, studies predominantly assessed tool performance: accuracy on routine, well-specified tasks ranged from about 62% to 100%, but fell to between 0% and 41% on multistep, under-specified, or context-rich problems requiring modeling assumptions, diagram interpretation, or real-world reasoning. Teacher's assistant roles demonstrated efficiency gains in grading, task generation, and workload reduction, but their contribution to learning outcomes was indirect and less systematically examined.

Several recurring constraints affect different roles. Performance drops when tasks require higher-order modeling or multimodal interpretation. Multilingual research highlights significant inequities in non-English-dominant contexts. Overreliance on fluent but sometimes inaccurate outputs highlights the need for explicit AI literacy for students and teachers. Educators appreciate GenAI's time-saving benefits but raise concerns about reliability, plagiarism, and diminished critical thinking. Researchers stress the importance of transparency, fair assessments, and making sure all students have equal opportunities.

This review makes three key contributions. First, it offers the first discipline-specific, PRISMA-guided synthesis of GenAI in physics education, consolidating a fragmented and rapidly growing body of evidence. Second, it establishes a three-role pedagogical framework (tutor, assistant, solver), supported by bibliometric mapping, that can orient both future research and instructional design. Third, it contributes a quantitative synthesis layer by aggregating role distributions, research designs, learning outcomes, and performance accuracies, thereby strengthening the evidence base for subsequent investigations and providing practical guidance for educators considering the integration of GenAI into physics teaching and assessment.

Several limitations of this review should be noted. The evidence base is still small and emergent, with only 21 studies meeting the inclusion criteria. Most of these are exploratory, small-scale, and concentrated in university contexts, with a strong focus on ChatGPT-based tools—limiting the generalizability of findings across educational levels, tool types, and instructional formats. The heterogeneity of research designs, outcome measures, and reporting standards precluded formal meta-analysis and complicates direct comparison across studies; descriptive aggregation of role distributions, accuracy rates, and outcome patterns was therefore employed instead of pooled effect sizes. A formal risk-of-bias assessment was not conducted, as the methodological diversity of the included studies made the application of a single standard tool impractical; however, such variability likely introduces biases that should be considered when interpreting the findings. Notably, the majority of studies (76%) employed non-experimental designs, predominantly using performance-based evaluations of GenAI tools rather than controlled comparisons with human learners or alternative instructional methods. This prevalence of benchmarking studies—while valuable for establishing technical feasibility—limits the strength of causal claims regarding learning outcomes and pedagogical effectiveness. The search was restricted to the Scopus database, which may have excluded relevant work in other indexed sources or grey literature. Finally, the time frame (2023–2025) captures only the earliest phase of GenAI adoption in physics education, meaning that practices and impacts are likely to evolve rapidly.

Future research needs to move beyond early trials and focus on long-term classroom studies, especially in secondary education and multilingual contexts. Priority areas include integrating AI literacy into physics curricula, developing equitable access strategies, and designing assessment practices that address GenAI use. With clear pedagogical goals and strong teacher oversight, GenAI can be a helpful partner in physics education. It can improve students' understanding and make teaching more efficient, while still valuing the important role of human expertise.

## Recommendations

The findings of this review suggest that GenAI tools can enhance conceptual understanding, student engagement, and instructional efficiency in physics education. This is most effective when integration occurs within well-structured pedagogical frameworks. To optimize benefits and reduce risks, several measures are recommended. Educators and students should receive targeted training in AI literacy, prompt engineering, and critical evaluation of AI-generated content. GenAI tools are most effective in tutoring, assessment, and problem-solving when supported by rigorous human oversight to ensure accuracy and fairness. Tutoring applications consistently yield the most positive learning outcomes. Institutions are encouraged to implement pilot programs. They should adjust GenAI use to align with curricular objectives, ensure equitable access, and address current challenges with complex or multimodal tasks.

Beyond these considerations, integrating GenAI in physics education raises questions about the evolving role of teachers and the long-term development of student competencies. As GenAI manages more routine explanations and problem-solving, educators may focus more on critical evaluation, conceptual synthesis, and designing learning experiences that remain beyond GenAI's capabilities, such as laboratory investigations, collaborative problem-solving, and epistemic reflection. However, continued reliance on AI-generated solutions could undermine students' independent problem-solving and procedural fluency. Providing opportunities for unassisted practice and metacognitive reflection helps mitigate this risk.

## Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

## Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Author Contribution Statement

**Maria Moundridou:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Nikolaos Moutis:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **Konstantinos Charalampopoulos:** Investigation, Writing – review & editing. **Nikolaos Matzakos:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization.

## References

[1] Chiu, T. K. F. (2024). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, *32*(10), 6187–6203. https://doi.org/10.1080/10494820.2023.2253861

[2] Moundridou, M., Matzakos, N., & Doukakis, S. (2024). Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans. *Computers and Education: Artificial Intelligence*, *7*, 100277. https://doi.org/10.1016/j.caeai.2024.100277

[3] Redish, E. F., & Burciaga, J. R. (2004). Teaching physics with the Physics Suite. *American Journal of Physics*, *72*(3), 414. https://doi.org/10.1119/1.1691552

[4] Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

[5] Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, *23*(7), 5–12. https://doi.org/10.3102/0013189X023007005

[6] McDermott, L. C., & Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics*, *67*(9), 755–767. https://doi.org/10.1119/1.19122

[7] Coban, A., Dzsotjan, D., Küchemann, S., Durst, J., Kuhn, J., & Hoyer, C. (2025). AI support meets AR visualization for Alice and Bob: Personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses. *EPJ Quantum Technology*, *12*(1), 15. https://doi.org/10.1140/epjqt/s40507-025-00310-z

[8] 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), 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

[9] 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

[10] Horchani, R. (2025). ChatGPT's problem solving abilities in context-rich and traditional physics problems. *Physics Education*, *60*(2), 025019. https://doi.org/10.1088/1361-6552/adb473

[11] Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education*, *8*, 1330486. https://doi.org/10.3389/feduc.2023.1330486

[12] López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher*, *62*(4), 290–294. https://doi.org/10.1119/5.0160160

[13] Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *International Journal of STEM Education*, *9*(1), 59. https://doi.org/10.1186/s40594-022-00377-5

[14] 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

[15] Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and higher education: Trends, challenges and future directions from a systematic literature review. *Information*, *15*(11), 676. https://doi.org/10.3390/info15110676

[16] Aydin-Günbatar, S., Durukan, A., & Günbatar, M. S. (2025). Generative AI as the new frontier in science education: A systematic review of Web of Science articles. *Science & Education*. Advance online publication. https://doi.org/10.1007/s11191-025-00677-6

[17] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ..., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

[18] Aria, M., & Cuccurullo, C. (2017). *bibliometrix*: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

[19] Alarbi, K., Halaweh, M., Tairab, H., Alsalhi, N. R., Annamalai, N., & Aldarmaki, F. (2024). Making a revolution in physics learning in high schools with ChatGPT: A case study in UAE. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(9), em2499. https://doi.org/10.29333/ejmste/14983

[20] 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), 010142. https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

[21] Fadillah, M. A., Usmeldi, U., & Asrizal, A. (2024). The role of ChatGPT and higher-order thinking skills as predictors of physics inquiry. *Journal of Baltic Science Education*, *23*(6), 1178–1192. https://doi.org/10.33225/jbse/24.23.1178

[22] Fontao, C. B., Santos, M. L., & Lozano, A. (2024). ChatGPT's role in the education system: Insights from the future secondary teachers. *International Journal of Information and Education Technology*, *14*(8), 1035–1043. https://doi.org/10.18178/ijiet.2024.14.8.2131

[23] Jang, H., & Choi, H. (2025). A double-edged sword: Physics educators' perspectives on utilizing ChatGPT and its future in classrooms. *Journal of Science Education and Technology*, *34*(2), 267–283. https://doi.org/10.1007/s10956-024-10173-1

[24] Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, *19*(2), 020150. https://doi.org/10.1103/PhysRevPhysEducRes.19.020150

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

[26] Kortemeyer, G., Nöhl, J., & Onishchuk, D. (2024). Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. *Physical Review Physics Education Research*, *20*(2), 020144. https://doi.org/10.1103/PhysRevPhysEducRes.20.020144

[27] Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., & Campanelli, M. (2025). Using large language models for grading

in education: An applied test for physics. *Physics Education*, *60*(3), 035006. https://doi.org/10.1088/1361-6552/adb92b

[28] Pimbblet, K. A., & Morrell, L. J. (2025). Can ChatGPT pass a physics degree? Making a case for reformation of assessment of undergraduate degrees. *European Journal of Physics*, *46*(1), 015702. https://doi.org/10.1088/1361-6404/ad9874

[29] Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the test of understanding graphs in kinematics. *Physical Review Physics Education Research*, *20*(1), 010109. https://doi.org/10.1103/PhysRevPhysEducRes.20.010109

[30] Revalde, G., Zholdakhmet, M., Abola, A., & Murzagaliyeva, A. (2025). Can ChatGPT pass a physics test? *Technology, Knowledge and Learning*, *30*, 2459–2478. https://doi.org/10.1007/s10758-025-09814-0

[31] Shamshin, A. (2024). Implementation of artificial intelligence in an online platform for solving and checking the solutions of physical problems. In *Information Technology for Education, Science, and Technics: Proceedings of ITEST 2024*, *Volume 2*, 400–417. https://doi.org/10.1007/978-3-031-71804-5_27

[32] Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, N. S., & Bryan, L. A. (2024). Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning. *Computers and Education: Artificial Intelligence*, *7*, 100318. https://doi.org/10.1016/j.caeai.2024.100318

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

**How to Cite:** Moundridou, M., Moutis, N., Charalampopoulos, K., & Matzakos, N. (2026). Generative AI in Physics Education: A PRISMA-Guided Systematic Review of Empirical Studies. *Artificial Intelligence and Applications*, *4*(2), 149–164. https://doi.org/10.47852/bonviewAIA62027183