

Artificial Intelligence in Physics Education: A Critical Integrative Review

Joshua O. Okewale¹, Ibukunoluwa R. Omole², Abiola O. Ilori³, Rachel I. Obed⁴, N. Chetty⁵

^{1,2,3,4}*Department of Physics, University of Ibadan, Ibadan, Oyo State, Nigeria*

^{3,5}*School of Chemistry & Physics, University of KwaZulu-Natal, Pietermaritzburg, South Africa*

Corresponding email: ao.ilori@ui.edu.ng

ABSTRACT

Artificial intelligence (AI) is now used in physics education for tutoring, feedback, assessment, problem-solving, laboratory analysis, and teacher support, yet the literature often treats model capability, user satisfaction, and learning as equivalent. This review examines why similar technologies can lead to different educational consequences. A critical integrative design synthesized literature published from January 2018 to April 2026, while retaining earlier methodological sources where necessary. Physics-specific studies were prioritised, and science or STEM research was included only when it clarified mechanisms directly relevant to physics. A retrospective PRISMA-style audit identified 48 papers at the initial selection stage, of which 23 were included in the substantive critical synthesis. Six recurring functions were identified, but none were beneficial by definition. Positive effects were most credible when AI supported intermediate reasoning, operated within bounded tasks, drew on domain-relevant grounding, and remained subject to human verification. Evidence for affective and efficiency gains was more consistent than evidence for durable conceptual learning. The literature also revealed category errors in treating accurate answers as learning, conversational responsiveness as personalization, and fluent language as representational competence. Graphical, spatial, and multimodal tasks were especially vulnerable to reliability mismatch. The review proposes the Pedagogical Contingency Framework for AI in Physics Education, which explains outcomes through the interaction of pedagogical role, epistemic position, representational demand, and governance or grounding. Its central contribution is a move away from tool cataloguing towards an explanatory account of when AI augments disciplinary agency and when it substitutes for it.

ARTICLE INFO

Received: Aug. 04, 2026

Revised: Sept. 14, 2026

Accepted: Sept. 18, 2026

KEY WORDS

AI literacy, Automated feedback, Epistemic agency, Intelligent tutoring, Large language models

Suggested Citation (APA Style 7th Edition):

Okewale, J.O., Omole, I. R., Ilori, A. O., Obed, R. I., & Chetty N. (2026). Artificial Intelligence in Physics Education: A Critical Integrative Review. *International Research Journal of Science, Technology, Education, and Management*. 6(3), 45-60. <https://doi.org/10.5281/zenodo.23033819>.

INTRODUCTION

Physics provides a demanding test of educational artificial intelligence because disciplinary competence cannot be reduced to a correct verbal or numerical answer. Learners must coordinate concepts, equations, graphs, diagrams, vectors, measurements, idealisations, and empirical evidence. They must also judge whether a model applies, whether an approximation is defensible, and whether a result is physically plausible. An AI system may therefore perform well on a conventional benchmark while bypassing the representational coordination and epistemic judgement that physics education seeks to develop.

Generative AI and large language models (LLMs) have intensified the current debate, although AI in education has a longer history. Intelligent tutoring systems, adaptive learning environments, automated scoring, machine-learning classifiers, and learning analytics were established well before conversational generative systems became widely accessible. Reviews of STEM and science education identify applications in personalisation, assessment, prediction, feedback, and instructional decision-making (Zhai et al., 2020; Xu & Ouyang, 2022; Ismail & Umi, 2023; Chamola et al., 2026). What generative AI changes is the scale and accessibility of these functions. A general-purpose model can explain concepts, write code, analyse data, generate assessment items, and simulate tutorial dialogue without having been designed for a particular physics curriculum.

Pedagogically, this flexibility cuts both ways. Earlier domain-restricted systems were commonly organised around explicit knowledge structures and predetermined instructional sequences, whereas contemporary LLMs produce responses probabilistically across open domains. Their fluency can reduce barriers to help-seeking and widen access to sophisticated support. The same fluency, however, may conceal factual errors, conceptual instability, and unwarranted confidence (Ding et al., 2023; Kieser et al., 2023). In physics, where signs, diagrams, coordinate systems, and limiting cases carry substantial meaning, a coherent-sounding response may still be invalid in ways that a novice cannot readily detect.

The empirical record does not support simple claims of either benefit or harm. ChatGPT has reproduced familiar misconceptions in introductory physics (Kortemeyer, 2023a; Wheeler & Scherr, 2023), encountered difficulty with graph interpretation and spatially coordinated reasoning (Polverini & Gregorcic, 2024; Polverini et al., 2025), and produced polished responses that differ from students' visible sensemaking processes (Simmoorkar et al., 2024). However, AI-assisted feedback has proved useful in constrained tasks (Wan & Chen, 2024), customised chatbots have improved selected learning and affective outcomes (Lademann et al., 2025), and AI-supported data analysis has increased enjoyment and reduced stress even where cognitive gains did not exceed those achieved with conventional tools (Henze et al., 2026). These results become less contradictory when the pedagogical configuration of AI, rather than the technology alone, is treated as the object of analysis.

Review activity has expanded alongside the primary literature. Broad syntheses have mapped AI across science and STEM education (Xu & Ouyang, 2022; Almasri, 2024; Cheung et al., 2025; Lee et al., 2026), while Roziqin et al. (2026) recently provided a systematic account of content coverage, implementation models, learning effects, and challenges in physics education. The present article takes a different analytical position. Rather than adding another catalogue of tools, topics, or reported effects, it examines why apparently similar systems produce different educational outcomes and why positive findings may reflect accessibility, affect, or efficiency rather than deeper learning.

OBJECTIVES OF THE STUDY

The main objective of this study is to critically examine the pedagogical conditions that determine the educational value of artificial intelligence in physics education. The study examines how different AI functions are pedagogically configured in physics teaching and learning and assesses the learning outcomes supported by available empirical evidence. It further identifies the task-related, contextual, governance, and grounding conditions that moderate the outcomes of AI-supported physics instruction and investigates why apparently similar AI systems may produce

different educational results. The study also determines whether reported positive outcomes represent deeper conceptual learning or merely improvements in accessibility, learner affect, and instructional efficiency. In addition, it distinguishes pedagogical augmentation from epistemic substitution in the use of AI for physics education. Ultimately, the study develops the Pedagogical Contingency Framework for AI in Physics Education (PCF-AIPE), a discipline-sensitive framework that explains how pedagogical role, epistemic position, representational demand, and governance or grounding interact to shape the educational value of AI.

CONCEPTUAL ORIENTATION

AI is not a Unitary Educational Intervention

Much of the imprecision in the literature stems from treating AI as a homogeneous independent variable. A machine-learning classifier used to code explanations, a domain-constrained tutor that delivers sequenced hints, and a general-purpose chatbot that supplies complete solutions differ in architecture, pedagogical role, and autonomy. Placing them under a single label obscures the mechanisms by which learning may be supported, redirected, or displaced.

The more useful analytical question is not simply which model or product was used, but what the system was authorised to do within the educational activity. The same LLM may prompt a learner to articulate assumptions, provide an answer that eliminates the need for reasoning, draft feedback for teacher review, write code for laboratory analysis, or classify responses for research purposes. Capability alone does not determine educational value; role matters. For this reason, the review treats AI as occupying a range of pedagogical positions rather than as a single intervention.

Physics Learning as Representational and Epistemic Practice

Physics expertise requires movement among verbal, symbolic, graphical, diagrammatic, numerical, and experimental representations. Such transformations are not secondary presentation skills. They are part of disciplinary reasoning itself. A learner who can substitute values into an equation but cannot connect that equation to a free-body diagram, a graph, or an empirical pattern has not developed the same competence as a learner who can coordinate these forms.

This distinction separates answer performance from learning performance. An AI system can produce a correct final response without undergoing the developmental process that a learner must master. By contrast, an incomplete or initially incorrect student attempt may be educationally productive when it exposes assumptions, prompts model revision, or supports translation between representations. Evaluating AI in physics education should therefore focus on whether an intervention strengthens learners' ability to construct, test, and revise explanations, rather than merely on whether it increases the proportion of correct answers.

Augmentation, Substitution, and Calibrated Reliance

AI augments learning by extending a learner's capacity to undertake meaningful disciplinary work while leaving central epistemic decisions in the learner's hands. A hint offered after an initial attempt, a critique that must be evaluated, coding support that preserves decisions about model selection and uncertainty, or feedback validated by a teacher all fit this description. Substitution occurs when AI performs the intellectual work that the task was designed to cultivate, such as constructing the model, selecting the governing principle, writing the interpretation, or producing the final derivation before the learner has engaged with the problem.

Augmentation and substitution lie on a continuum rather than forming a simple binary. Their educational significance also varies with expertise. An expert may use a generated solution efficiently because it can be tested against a well-developed knowledge base; a novice may accept the same output because fluency is mistaken for validity. The appropriate educational aim is calibrated reliance. This involves neither unconditional trust nor

categorical distrust, but an ability to adjust reliance according to the task, available evidence, the consequences of error, and the system's demonstrated limitations.

MATERIALS AND METHODS

Review Design and Scope

A critical integrative review design was selected because the evidence base varies widely in technology, research design, educational level, physics domain, and outcome measure. Integrative review methods are suited to bodies of literature in which experimental, quasi-experimental, qualitative, mixed-methods, capability-evaluation, and design-oriented studies must be brought together to develop an explanatory interpretation rather than a pooled effect estimate (Whittemore & Knafl, 2005). The review drew on PRISMA 2020 principles of transparent identification, selection, and reporting, although it was not structured as a conventional meta-analysis (Page et al., 2021).

The substantive publication window ran from January 2018 to April 2026, capturing the shift from machine-learning and automated-assessment applications to the rapid growth of generative AI after 2022. Earlier sources were retained only where required for the review methodology. The evidence boundary remained deliberately sensitive to the discipline. Physics-specific literature received priority, while broader science or STEM studies were included only when they clarified a mechanism directly relevant to physics learning, such as epistemic agency, automated evaluation of scientific explanations, teacher AI literacy, or human-centred governance.

Search Strategy and Evidence Identification

Searches were designed for Scopus, Web of Science Core Collection, ERIC, and IEEE Xplore, and were supplemented by publisher-platform searches, journal archive searches, backward reference checking, and forward citation chaining. The core Boolean structure paired terms for artificial intelligence, generative AI, LLMs, chatbots, intelligent tutoring, machine learning, learning analytics, automated feedback, and automated assessment with terms for physics education, physics teaching, physics learning, physics classrooms, and physics laboratories. A supplementary science and STEM string combined the same AI terms with assessment, reasoning, feedback, inquiry, and laboratory processes.

The uploaded source manuscript did not include database export files, duplicate-removal records, or a contemporaneous screening log, so conventional identification and exclusion totals could not be reconstructed without fabrication. A retrospective audit was therefore used to document the papers considered in the manuscript. Forty-eight papers were initially selected for closer assessment. After screening for topical relevance, source verifiability, and synthesis role, 23 reports formed the substantive critical synthesis. Two sources addressed review methodology and reporting standards, while three supplied conceptual or governance context without being treated as evidence of learning effects. Among the 23 substantive reports, 17 were physics-specific, and six were transferable science or STEM evidence syntheses. Figure 1 records this corpus composition and should not be interpreted as a prospective, exhaustive database flow.

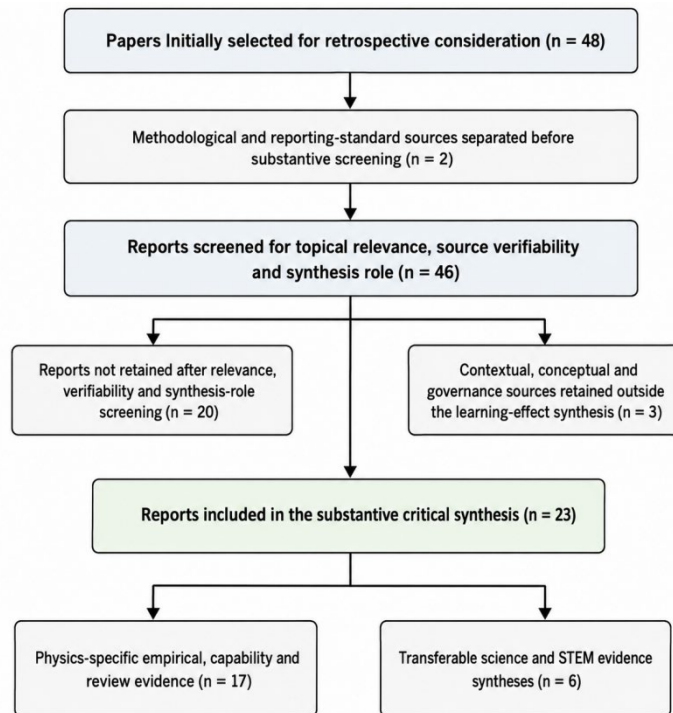


Figure 1. Retrospective PRISMA-style audit of the cited evidence corpus.

Eligibility, Appraisal, and Synthesis

Eligible reports addressed AI in physics teaching, learning, assessment, teacher education, or physics education research, or provided directly transferable evidence from science education. Empirical studies, model-capability evaluations, substantive peer-reviewed conference papers, and high-quality reviews were retained if they provided sufficient methodological detail to assess the relationship between the AI application and the educational claim. Reports were excluded from substantive consideration when AI was used solely for scientific discovery or computational physics, when an engineering or computing context offered no transferable mechanism for physics learning, when the report was promotional or speculative, or when the AI task and evaluation process were insufficiently described.

Given the diversity of study designs, a single numerical quality score would have created a misleading equivalence. Reports were instead appraised against the fit between the question and the design, the adequacy of the sample and context, the validity of outcome measures, transparency about the model and prompting configuration, the strength of the comparison or human benchmark, and acknowledgement of uncertainty. Observable learning outcomes, meaningful baselines, triangulation, and replication received greater interpretive weight. Self-reported satisfaction, behavioural intention, and one-off model performance were treated as useful evidence, but not as sufficient proof of durable learning.

Data extraction used a structured matrix covering educational level, country, physics domain, AI system, pedagogical role, learner or teacher task, representational demand, domain grounding, human oversight, study design, outcome type, reported risk, and methodological limitation. Studies were first coded by pedagogical function and outcome domain. Constant comparison was then used to examine contradictions across the evidence. Four explanatory dimensions emerged from this process: pedagogical role, epistemic position, representational demand, and governance or grounding. These dimensions underpin the PCF-AIPE developed in the discussion.

RESULTS: CRITICAL SYNTHESIS OF EVIDENCE

Expansion Without Commensurate Methodological Maturity

Research on AI in physics and science education is expanding more quickly than the evidence for many evidential claims is maturing. Earlier studies centred on intelligent tutoring, machine-learning assessment and analytics, whereas recent work is dominated by general-purpose and customised generative systems (Xu & Ouyang, 2022; Almasri, 2024; Lee et al., 2026; Roziqin et al., 2026). This growth has broadened the range of educational applications. However, it has also encouraged a tool-centred research culture in which model novelty can displace clarity about the learning process under investigation.

Three recurring category errors are especially consequential. Capability studies are often read as learning studies, although a model's ability to answer an inventory item says nothing by itself about whether students learn through using it. Perception studies are also used to imply effectiveness. Ease of use, confidence, and satisfaction matter, but they do not establish conceptual change or transfer. A third error occurs when responsiveness is labelled personalisation even though the system has not diagnosed the learner's conceptual state or selected an instructionally appropriate response. Together, these problems help explain how the literature can report impressive performance, positive attitudes, and uncertain learning gains simultaneously.

The evidence is narrow in both time and context. Many interventions are brief, rely on convenience samples, and assess immediate post-intervention achievement or self-report. Higher education is overrepresented, whereas sustained secondary-school implementation, multilingual classrooms, low-resource settings, and learners with disabilities receive comparatively little attention. Model volatility adds another difficulty. Findings associated with one version may not carry over to a later release, and greater accuracy does not, by itself, resolve dependency, assessment validity, or epistemic offloading.

Tutoring and Personalisation: Access Without Diagnosis

Conversational AI offers real advantages through availability, low-stakes help-seeking, and iterative dialogue. Students can ask for rephrasing, examples, or follow-up explanations without the social cost of repeatedly questioning an instructor. Ding et al. (2023) found that ChatGPT answered many selected introductory physics questions correctly and was valued for accessibility and interaction. The same study, however, showed that students sometimes trusted inaccurate responses and that continued dialogue could prompt the system to abandon an initially correct answer. The affordance was therefore accompanied by a reliability problem that users could not always recognise.

These findings clarify the difference between adaptive language and pedagogical personalisation. A system may vary its wording without maintaining a valid account of the learner's misconception, prior knowledge, or progress. Genuine personalisation requires diagnosis, selection of an appropriate instructional move, monitoring of subsequent understanding, and adjustment of support. General-purpose chatbots often reproduce only the surface features of this process. Their conversational flexibility can create an impression of individualisation that is stronger than the underlying pedagogical evidence would suggest.

Adoption is also shaped by context. In a survey of 804 physics students across Ghana, Jordan and the United States, Agyare et al. (2025) showed that perceived usefulness and ease of use were embedded in national and institutional conditions. Guo et al. (2025) likewise found that students valued ease and engagement but were less persuaded that ChatGPT met individual needs or could replace instructors; physics discussion activities were also rated less favourably than more procedural coding activities. Such findings challenge universal adoption narratives. Access, curriculum, language, disciplinary tasks, and institutional norms all influence what students experience as useful.

Conversational AI is therefore most defensible as a dialogic resource rather than an authority. Its capacity to augment learning is stronger when students attempt the task first, the system provides bounded prompts or alternative explanations, and verification is built explicitly into the activity. Without these conditions, fluent dialogue may create only the appearance of personalisation, increasing exposure to unnoticed error while reducing learners' responsibility for conceptual progression.

Assessment and Feedback: Scalable Assistance, Non-Delegable Judgement

Assessment and feedback are among the most practical areas of application because teachers often manage large volumes of student work under conditions of delayed feedback. AI can draft comments, identify recurring patterns, classify responses, and flag submissions for review. The central issue is not whether AI may participate in assessment, but which elements of judgement can be redistributed without weakening validity or accountability.

Wan and Chen (2024) examined generative-AI-assisted feedback on written responses to a physics conceptual question. Students rated human and AI feedback similarly for correctness and found the AI feedback useful, while instructor review indicated that a substantial proportion of submissions required minor or no modification. The findings support AI as a first-pass co-producer of feedback. They do not support autonomous feedback across unconstrained tasks, since the reported success depended on a specific question, engineered prompting, and human evaluation.

The same distinction applies to grading. Kortemeyer (2023b) demonstrated the feasibility of GPT-based evaluation of handwritten physics solutions, but feasibility should not be confused with validity. Physics solutions may use unconventional yet defensible methods, contain partial reasoning, employ ambiguous notation, or carry essential information in diagrams. A general-purpose system may misread these features or reward polished language over sound physical reasoning. A layered architecture is more defensible: AI can classify, draft, or flag, while humans validate, contextualise, and remain accountable for consequential decisions.

AI-generated assessment items introduce a related validity risk. Küchemann et al. (2024) showed that plausible concept-inventory items still require empirical evidence of reliability and construct validity. Producing more items at greater speed does not solve the psychometric problem; it may simply multiply weak items. The available literature, therefore, favours the redistribution of workload rather than the replacement of professional judgement. AI can reduce repetitive labour, but disciplinary expertise remains necessary to interpret misconceptions, protect fairness, and connect assessment evidence with instruction.

Problem Solving: The Final-Answer Fallacy

Physics problem-solving is where model capability and educational value are most frequently conflated. Kortemeyer (2023a) reported that an early version of ChatGPT could narrowly pass an introductory physics assessment while making errors that resembled novice misconceptions. Wheeler and Scherr (2023) similarly showed that the model could reproduce familiar incorrect ideas. These were not merely random mistakes. Their pedagogical danger lay in the coherent form through which intuitive but scientifically inappropriate reasoning was expressed.

Sirnoorkar et al. (2024) offer a more fundamental critique. Comparing student and AI responses through the lenses of sensemaking and mechanistic reasoning, they found that AI output could display formal coherence and appropriate terminology. In contrast, student work more clearly showed iterative refinement and active use of representations. The comparison suggests that AI is commonly judged by the quality of completed discourse, even though learning depends on the quality of the disciplinary activity that produced it.

This pattern constitutes the final-answer fallacy: the assumption that an improved final product signals an improved learning process. A polished derivation can conceal the fact that the learner did not select the model, coordinate representations, test assumptions, or evaluate the result. Even a highly accurate answer generator may be

educationally counterproductive when the objective is to construct the answer. Improved model accuracy does not eliminate this problem; it may make the substitution harder to detect because the output is more persuasive. Problem-solving support is more defensible when AI is positioned away from final-answer production and closer to intermediate reasoning. It may ask diagnostic questions, provide progressively stronger hints, identify an assumption, critique a proposed model, or generate alternative explanations for comparison. Designs of this kind preserve the student as the principal epistemic agent and leave observable traces of reasoning that can be assessed, discussed, and revised.

Representational and Multimodal Reasoning: Fluency is Not Coordination

The emerging evidence base shows that AI competence in physics is sensitive to the choice of representation. Polverini and Gregorcic (2024) found that ChatGPT-4 used productive strategies on the Test of Understanding Graphs in Kinematics and achieved overall performance comparable to that of high school students, although graphical interpretation remained difficult. Polverini et al. (2025) reported substantial improvement by ChatGPT-4o on the Brief Electricity and Magnetism Assessment, yet persistent problems remained in visually coordinated, spatial, and right-hand-rule tasks.

These results point beyond the limitations of any single model version. They reveal a broader evaluation problem in which text-heavy benchmarks may overstate disciplinary competence. A model can describe the relationship between velocity and acceleration accurately while misreading a graph, or explain an electromagnetic rule while applying it incorrectly to a diagram. Physics knowledge is distributed across representational forms, so success in one form does not establish competence in another.

Multimodal models are likely to improve, but technical progress does not settle the pedagogical question. Even a system that correctly interprets a diagram may remove the learner's need to practise interpreting diagrams. Evaluation must therefore address two issues: whether the system can perform the representational task reliably, and whether its position in the activity preserves the learner's responsibility for mastering that task. Representation-specific validation and pedagogical design need to be considered together.

Laboratory Work and Data Analysis: Selective Delegation

Laboratory and data-analysis settings show how AI may reduce unnecessary friction without necessarily improving conceptual learning. Students often devote substantial time to spreadsheet operations, plotting, coding syntax, and data cleaning. Some of this work has educational value, but an excessive procedural burden can crowd out decisions about variables, models, uncertainty, and evidence.

Henze et al. (2026) compared an AI-supported data-analysis environment with a conventional spreadsheet approach. Both groups recorded learning gains, but the AI condition did not produce statistically superior cognitive outcomes. It increased enjoyment, engagement, and perceived effectiveness while reducing stress. Lademann et al. (2025) similarly showed that customised chatbots can influence learning performance, cognitive load, and affective variables in design-dependent ways. Read together, these studies reveal an affective-cognitive dissociation: AI may improve the learning experience without demonstrating superior conceptual achievement.

The distinction matters because reduced difficulty can have different educational meanings. Removing syntax errors or repetitive plotting may free attention for interpretation. Removing the need to select a model, justify an exclusion, or evaluate uncertainty weakens scientific judgment. Selective delegation is therefore the more defensible design. AI may assist with low-level operations, but learners should retain responsibility for defining variables, explaining transformations, selecting and defending models, interpreting residuals, evaluating uncertainty, and linking results to physical claims.

Laboratory use also brings a black-box risk into view. When AI generates code, removes outliers, and recommends a model without transparent justification, efficiency may be gained at the expense of methodological understanding. Responsible use requires provenance, inspectable reasoning, and opportunities for students to reconstruct the analysis independently.

Teacher Education, Educational Research, and The Limits of Scale

AI alters the work of teachers and researchers as well as that of students. Abdulayeva et al. (2025) found that structured training improved pre-service physics teachers' AI literacy and influenced their intention to use AI, with perceived usefulness acting as a mediator. Adoption, then, is not simply a matter of access. Teachers need knowledge of system limitations, evaluation practices, ethics, privacy, assessment redesign, and pedagogical orchestration. Sorge et al. (2025) used a language model to provide individualised feedback on pre-service physics teachers' written reflections and found that context specification improved classification accuracy. The study demonstrates the value of domain-specific configuration while also cautioning against making easy assumptions about transfer. A model calibrated to a structured reflection framework may not generalise across courses, languages, or pedagogical purposes. Local validation is part of the intervention itself, not an optional technical step.

AI may also extend the scale of physics education research. Kieser et al. (2023) used ChatGPT for educational data augmentation with the Force Concept Inventory. The model could solve the inventory and generate responses oriented towards misconceptions, but the synthetic data did not automatically reproduce the variability found in human populations. Synthetic learners are therefore useful for simulation and methodological development, not as substitutes for actual learners. Their apparent coherence may reflect assumptions embedded in prompts or training data.

The wider science-education literature reaches a comparable judgement. Human-centred analytics research stresses stakeholder participation, trustworthiness, and alignment between technical design and educational purpose (Alfredo et al., 2024). Reviews in science education also identify unresolved questions concerning epistemic agency, ethics, and AI's tendency to extend existing practices rather than transform their underlying logic (Cheung et al., 2025; Lee et al., 2026). Scale becomes educationally valuable only when transparency and accountability expand with it.

Table 1. Critical synthesis of major AI functions in physics education

AI function	Evidence-supported contribution	Central evidential weakness	Critical interpretation
1. Conversational tutoring	Accessible, iterative clarification and low-stakes help	Inaccuracy, overtrust, and superficial personalisation	Most defensible as verified dialogue after an initial learner attempt
2. Automated feedback	Rapid first-pass comments and misconception flagging	Generalisability beyond constrained prompts and tasks	A co-production model is better supported than autonomous feedback
3. Grading and assessment	Scalable classification, item drafting, and preliminary scoring	Construct validity, fairness, and multimodal misreading	Efficiency does not remove human accountability for consequential judgment
4. Problem-solving support	Hints, critique, and comparison of alternative methods	Final-answer substitution and cognitive offloading	Value increases when AI supports intermediate reasoning rather than completes it
5. Graphical and multimodal support	Potential assistance across diagrams, graphs, and vectors	Representation-specific and spatial errors	Text fluency should not be treated as evidence of multimodal competence

6. Laboratory and data analysis	Reduced procedural friction in coding, plotting, and workflow	Black-box modelling and weakened uncertainty judgement	Delegate operations selectively while preserving interpretive responsibility
7. Teacher support	Drafting, planning, feedback triage, and response-pattern analysis	Automation bias and generic pedagogy	Teacher expertise must localise, validate, and orchestrate AI use
8. Educational research	Coding, data augmentation, and large-scale analysis	Synthetic-data validity, model volatility, and reproducibility	Transparent reporting and local human validation are indispensable

Cross-Cutting Outcome Pattern

Across the reviewed functions, cognitive outcomes are promising but inconsistent. AI can widen access to explanations, support practice, and improve performance in structured activities, yet evidence of durable superiority over well-designed conventional instruction remains limited. Educational value is predicted less consistently by the presence of AI than by the design of the learning activity and the degree to which students remain responsible for disciplinary reasoning.

Affective outcomes show a more consistent positive pattern. Students commonly report engagement, convenience, confidence, enjoyment, and lower stress (Guo et al., 2025; Henze et al., 2026). These outcomes may support participation and persistence, but they should not be treated as cognitive achievement. Learning can feel easier because a tool removes an extraneous burden or productive struggle. The literature rarely separates these possibilities with enough care.

Efficiency gains are likewise more robust than learning gains. AI accelerates feedback, content generation, coding, analysis, and access to information. Speed is educationally useful only when the accelerated process preserves the intended construct. A faster final answer offers no learning advantage when the purpose of the task is to develop the reasoning required to produce it.

Epistemic outcomes are the least directly measured. Few studies ask whether students verify claims, tolerate uncertainty, persist with difficult reasoning, coordinate representations independently, or become dependent on the explanations generated. This gap is consequential because some of the biggest educational risks may emerge gradually through altered habits of judgement rather than through an immediate decline in test scores.

DISCUSSIONS

AI Effectiveness is Contingent Rather Than Intrinsic

The main finding is that the question, 'Does AI improve physics education?', is too coarse for analytical use. Outcomes depend on the pedagogical role assigned to the system, the point at which it enters the learner's reasoning, the representations required by the task, the learner's prior knowledge, and the degree of grounding, verification, and institutional accountability. A single model may strengthen one activity while weakening another.

This contingency perspective helps reconcile several apparent contradictions. Strong benchmark performance may coexist with fragile conceptual reasoning because capability tests assess products rather than learning processes. Positive perceptions may accompany unchanged achievement because affect and cognition are distinct outcomes. Greater efficiency may also coexist with weaker epistemic agency when a task is completed more quickly by outsourcing the judgment it was intended to develop.

The Epistemic Proximity Principle

The synthesis supports an epistemic proximity principle. As AI moves closer to producing the final intellectual product that the learner is expected to construct, the risk of substitution increases. This principle is probabilistic, not absolute. Experts may compare AI-generated solutions productively because they possess the knowledge needed to interrogate them. Novices face a different risk because subtle errors and missing assumptions may pass unnoticed.

For instruction, the principle favours delayed and intermediate use. AI is better introduced after an initial attempt, during comparison of alternative representations, as a source of hints or critique, or through feedback that the learner must evaluate. Assessment should similarly move away from final-answer production and towards visible reasoning, model choice, oral defence, uncertainty analysis, interpretation of unfamiliar evidence, and reflection on revision. These constructs remain educationally meaningful even when AI is available.

The Representation-Sensitivity Principle

The second principle holds that AI reliability must be established within the representational form required by the task. Evidence from kinematics graphs and electricity-and-magnetism assessments shows that strong linguistic or overall performance does not guarantee dependable visual and spatial reasoning (Polverini & Gregorcic, 2024; Polverini et al., 2025). A physics benchmark built mainly around text is, therefore, incomplete.

This principle has technical and pedagogical consequences. Models should be tested across equations, graphs, diagrams, vectors, spatial arrangements, experimental data, and transformations among these forms. Teachers should not infer broad competence from polished prose. Even as multimodal accuracy improves, students still need to practise coordinating representations rather than simply receiving a correct multimodal interpretation.

The Affective-Cognitive Dissociation

The third principle distinguishes the experience of learning from what is actually learned. AI may increase enjoyment, confidence, and engagement, and reduce stress, while producing cognitive outcomes comparable to those of conventional tools (Henze et al., 2026). These affective benefits have educational value, but they do not demonstrate superior conceptual learning.

Interpretation turns on the kind of effort that AI removes. Reducing repetitive syntax and interface friction may release cognitive resources for model-based reasoning. Removing the need to formulate a model or test an assumption may eliminate productive struggle. Future evaluations should distinguish extraneous, procedural, and germane effort, then examine whether students can later perform independently and transfer their learning to unfamiliar contexts.

Calibrated Trust, Governance, and Equity

AI literacy in physics is better defined through calibrated reliance than through generic confidence. Students need established routines for checking units, signs, limiting cases, orders of magnitude, source provenance, and consistency across representations. Structured adversarial verification is particularly promising. Learners can analyse an AI response, identify its assumptions, test it against physical constraints, reconstruct the derivation, and revise the answer. In this arrangement, AI error becomes visible material for learning rather than a concealed risk.

Governance forms part of pedagogy rather than sitting outside it. The systems an institution permits, the data it retains, the access it provides, and the accountability it assigns all shape the conditions of learning. UNESCO's guidance and teacher competency framework emphasise human agency, inclusion, safety, and responsible use (UNESCO, 2023; UNESCO, 2024). These concerns become especially pressing when student prompts contain personal information, assessment content, or identifiable learning records.

Equity extends beyond connectivity. Paid models may offer stronger capabilities than free versions, language support remains uneven, and culturally specific curricula may be poorly represented. Cross-national evidence confirms that perceptions and patterns of use depend on context (Agyare et al., 2025). Responsible integration, therefore, requires local validation, equitable access, and a clear account of who bears the costs of error, verification, and exclusion.

The Pedagogical Contingency Framework For AI in Physics Education

The PCF-AIPE organises the synthesis around four interacting dimensions. The pedagogical role identifies what AI is asked to do, including tutoring, feedback, assessment, problem-solving support, representational assistance, laboratory analysis, teacher support, or research coding. No role is beneficial by definition. Each carries its own combination of educational opportunity and risk.

Epistemic position locates AI in relation to the learner's intellectual work, along a continuum from process scaffold to collaborative critic, co-producer, and answer substitute. Movement towards substitution increases the need for learner expertise, transparency, and verification. Representational demand refers to the forms that must be coordinated, ranging from structured textual or symbolic tasks to novel graphical, diagrammatic, spatial, and multimodal problems. Governance and grounding capture the extent to which a system is constrained, source-linked, domain-specific, and subject to human review.

The framework makes an interactive claim. A generic open model used by a novice to generate the final answer to an unvalidated spatial task represents a high-risk configuration. By contrast, a domain-grounded system that offers progressively stronger hints after a learner's attempt, with teacher review and explicit verification, represents a configuration of higher educational value. The framework shifts attention from whether a model works in general to the conditions under which it supports or displaces disciplinary agency.

Table 2. PCF-AIPE decision matrix for responsible integration

Pedagogical configuration	Expected educational value	Primary risk	Design response
1. Structured task; AI supplies hints after an initial attempt; outputs verified	High	Over-scaffolding	Use progressive hints and require learner explanation
2. Formative feedback drafted by AI and reviewed by the teacher and the student	High	Generic or inaccurate advice	Retain human validation and require a response to feedback
3. Data-analysis assistance with learner-controlled model choice and interpretation	Moderate to high	Black-box methods	Require methodological justification, uncertainty analysis, and provenance.
4. Open problem; final solution generated for a novice before engagement	Low to uncertain	Cognitive offloading and epistemic substitution	Delay access and assess the reasoning process

5. Multimodal or spatial task used without representation-specific validation	Uncertain	Reliability mismatch	Benchmark locally and require independent checking
6. High-stakes autonomous grading	Low acceptability	Validity, fairness, and accountability	Retain human judgement and auditable decision pathways
7. Teacher planning with local curricular adaptation	Moderate to high	Automation bias and generic content	Require professional editing and alignment to learning goals
8. Research coding or synthetic data with transparent local validation	High for augmentation	Reproducibility and construct distortion	Report model configuration, prompts, repeated runs, and validation metrics

Better Models Will Not Remove The Pedagogical Problem

One common response to present limitations is that better models will eventually solve them. Improved accuracy will reduce some technical failures, but it cannot remove the central educational problem. Even a perfectly accurate answer generator may undermine learning when the objective is for students to construct and justify their own answers.

Greater capability may intensify other difficulties. More persuasive output can discourage verification, make cognitive outsourcing less visible, and weaken the validity of conventional take-home assessments. The future of AI in physics education, therefore, rests not only on better models but also on stronger learning design, redesigned assessment, explicit AI literacy, and institutions prepared to preserve human accountability.

IMPLICATIONS FOR PRACTICE AND POLICY

Physics teachers should begin with the disciplinary process that students are expected to learn, not with the features offered by an AI system. AI is best placed where it removes unnecessary friction without performing the central intellectual work. In mechanics, for instance, a system may critique a free-body diagram after a student has constructed it; generating the diagram before any attempt changes the learning objective. In laboratory work, coding assistance may be appropriate, while model selection and interpretation of uncertainty remain the learner's responsibility.

Assessment designers need not focus on creating tasks that AI can never perform, since that strategy will become increasingly unstable. A more durable assessment foregrounds constructs that remain educationally meaningful when AI is available: visible reasoning, oral defence, interpretation of unfamiliar graphs, comparison of competing models, uncertainty analysis, provenance, and reflection on revision. Critiquing AI output can itself serve as an assessment of disciplinary judgement.

Teacher education must extend beyond generic prompt training. Pre-service and practising teachers need a working understanding of how generative systems operate, how hallucination and automation bias appear, how representational failures can be identified, how privacy and disclosure should be managed, and how assessment may be redesigned. Authentic physics tasks should anchor this training because generic demonstrations understate the discipline's representational demands.

Institutions should avoid both unrestricted adoption and blanket prohibition. A defensible policy combines equitable access, approved data-protection practices, clear expectations for disclosure, professional development,

procedures for local evaluation, and human accountability for consequential decisions. Developers of physics-oriented systems should also benchmark performance across textual explanation, symbolic manipulation, graphs, diagrams, vectors, experimental data, uncertainty, and movement between representations.

RESEARCH AGENDA

The field has moved beyond the point where additional surveys of general willingness to use ChatGPT are sufficient. Research should now prioritise longitudinal learning, delayed retention, and transfer to unfamiliar tasks. A system that improves immediate performance while leaving students unable to work independently does not provide the same educational value as one that supports durable competence.

Process measures deserve equal attention. Researchers should examine how AI changes time spent on reasoning, help-seeking, representational activity, revision, verification, persistence, uncertainty monitoring, and the distribution of talk among the learner, the teacher, and the system. Achievement scores alone cannot show whether the mechanism of learning has changed.

Representation-sensitive experiments should hold learners and models constant while varying text, equations, graphs, diagrams, vectors, and multimodal tasks. Comparisons should also include novices, intermediate learners, and experts, as the capacity to evaluate AI output is likely to moderate both the benefits and harms. Trust-calibration studies need to test whether verification routines transfer across topics and whether learners can distinguish dependable output from unreliable output.

Teacher-focused research should establish whether AI reduces workload or merely shifts it from drafting to verification, correction, and monitoring. Equity research needs wider coverage of low- and middle-income countries, multilingual classrooms, secondary schools, resource-constrained institutions, and settings where access to premium systems is unequal. Research on high-stakes assessment should move beyond detection to examine how constructs, evidence, and accountability change when AI is routinely available.

Reproducibility standards also need revision. Studies should report the model provider and version, access date, system instructions, prompts, relevant parameters, number of repeated runs, output selection procedures, human validation criteria, and sensitivity across models or prompts. Without such detail, an apparent finding may reflect an undocumented configuration rather than reproducible educational knowledge.

LIMITATIONS

Several limitations shape this review. AI systems change rapidly, so conclusions tied to a specific model may become technologically outdated. The analysis consequently emphasised pedagogical mechanisms, although those mechanisms will still require continued testing as multimodal capability and grounding improve.

The evidence base is heterogeneous across technology, topic, sample, and outcome, which precludes a meaningful meta-analysis. English-language literature received priority, and broader science and STEM evidence was used selectively, so relevant work in other languages or adjacent disciplines may be missing. The corpus also overrepresents higher education, short interventions, and self-reported outcomes.

CONCLUSION

Artificial intelligence is neither inherently transformative nor inherently destructive in physics education. Its consequences depend on where and how it is positioned within the learning process. The strongest evidence supports its use as a scaffold, critic, first-pass feedback assistant, selective data-analysis aid, and means of reducing unnecessary procedural friction. The greatest risks arise when AI is treated as an unquestioned authority, performs

the reasoning that students are expected to develop, or is used for representationally demanding, high-stakes tasks without suitable validation.

The PCF-AIPE accounts for these differences through four interacting dimensions: pedagogical role, epistemic position, representational demand, and governance or grounding. It extends beyond existing tool maps by explaining why similar technologies may yield different outcomes and by identifying the central educational concern as the preservation of learner epistemic agency.

The future of AI in physics education should not be judged by the fluency with which machines answer physics questions. The stronger test is whether human-AI arrangements deepen learners' capacity to construct, test, represent, criticise, and revise physical explanations. The productive aim is not an AI system that performs physics on students' behalf, but a pedagogy that improves the conditions under which students learn to think and judge as physicists.

REFERENCES

- Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., & Boribekova, F. (2025). Fostering AI literacy in pre-service physics teachers: Inputs from training and co-variables. *Frontiers in Education*, 10, 1505420.
- Agyare, B., Asare, J., Kraishan, A., Nkrumah, I., & Adjekum, D. K. (2025). A cross-national assessment of artificial intelligence (AI) chatbot user perceptions in collegiate physics education. *Computers and Education: Artificial Intelligence*, 8, 100365.
- Alfredo, R., Echeverría, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215.
- Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education*, 54, 977–997.
- Chamola, V., Dave, D., Goyal, I., & Sharma, S. (2026). Generative Artificial Intelligence in STEM Education: A Review of Applications, Benefits, and Challenges. *Expert Systems*, 43(2).
- Cheung, K. K. C., Long, Y., Liu, Q., & Chan, H.-Y. (2025). Unpacking epistemic insights of artificial intelligence (AI) in science education: A systematic review. *Science & Education*, 34, 747–777.
- Ding, L., Li, T., Jiang, S., & A.A. Gapud. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, 20(1).
- Guo, F., Li, T., & Cunningham, C. J. L. (2025). One year in the classroom with ChatGPT: Empirical insights and transformative impacts. *Frontiers in Education*, 10, 1574477.
- Henze, J., Lademann, J., Becker-Genschow, S., & André Bresges. (2026). AI-supported data analysis boosts student motivation and reduces stress in physics education. *Frontiers in Education*, 11.
- Ismail, N., & Umi, K. Y. (2023). A systematic literature review: Recent techniques of predicting STEM stream students. *Computers & Education: Artificial Intelligence*, 100141–100141.
- Kieser, F., Wulff, P., Jochen Kühn, & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review*, 19(2).
- Kortemeyer, G. (2023a). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1).
- Kortemeyer, G. (2023b). Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research*, 19, 020163.
- Küchemann, S., Rau, M., Schmidt, A., & Kuhn, J. (2024). ChatGPT's quality: Reliability and validity of concept inventory items. *Frontiers in Psychology*, 15, 1426209.
- Lademann, J., Henze, J., & Becker-Genschow, S. (2025). Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables. *Physical Review Physics Education Research*, 21(1).

- Lee, G., Yun, M., Zhai, X., & Crippen, K. (2026). Artificial intelligence in science education research: Current states and challenges. *Journal of Science Education and Technology*, 35, 110–127.
- 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, n71.
- Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the Test of Understanding Graphs in Kinematics. *Physical Review Physics Education Research*, 20(1).
- Polverini, G., Melin, J., Önerud, E., & Gregorcic, B. (2025). Performance of ChatGPT on tasks involving physics visual representations: The case of the brief electricity and magnetism assessment. *Physical Review Physics Education Research*, 21(1).
- Roziqin, R., Samsudin, A., Saepuzaman, D., Sari, H. N., & Iryanti, M. (2026). Artificial intelligence in physics education: A systematic review of content coverage, implementation models, learning impact, and pedagogical challenges. *STEM Education*, 6, 539–583.
- Sirnoorkar, A., Zollman, D., Lavery, 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.
- Sorge, S., Wulff, P., & Kubsch, M. (2025). Using a large language model to provide individualized feedback for pre-service physics teachers' written reflections. *Disciplinary and Interdisciplinary Science Education Research*, 7, 25.
- UNESCO. (2023). Guidance for generative AI in education and research. In *UNESCO.org*. <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- UNESCO. (2024). AI competency framework for teachers. In *UNESCO.org*.
- Wan, T., & Chen, Z. (2024). Exploring generative AI-assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review. Physics Education Research*, 20(1).
- Wheeler, S., & Scherr, R. E. (2023). ChatGPT reflects student misconceptions in physics. *Per-Central.Org*, 386–390.
- Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, 52, 546–553.
- 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), 1–20. .
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, 56(1), 1–41. .