E-ISSN: 3164-8940
https://doi.org/10.47945/create.v1i1.3355

Evaluating AI-Generated Physics Solutions: A Systematic Review of Solution Accuracy, Student Evaluation, and Implications for Critical Thinking in Higher Physics Education

Gladys Mahaut
[a]Departemen of Science Education, Institut Agama Islam Negeri Sorong, Indonesia

ABSTRACT

Generative artificial intelligence (AI) is now capable of producing coherent, step-by-step solutions to physics problems. However, its implications for physics education are determined more by students' ability to evaluate the validity of these solutions than by their availability. This systematic review, reported with reference to the PRISMA 2020 guidelines, synthesizes peer-reviewed physics education studies published between 2020 and 2026 to address three questions: (1) how accurately does generative AI solve or explain physics tasks, and what types of errors does it produce; (2) how do students evaluate AI-generated physics responses; and (3) what instructional opportunities and risks emerge for the development of critical thinking in higher physics education? Eleven studies were synthesized through deductive and inductive thematic analysis. AI accuracy was highly dependent on task type. Large language models performed at levels comparable to or above those of university students on several concept inventories and olympiad problems, but solved only 8.3% of underspecified problems (compared with 62.5% of well-specified problems), misinterpreted graphs and visual representations, and generated contradictory explanations. The dominant errors were related to physical modeling, assumptions, and representations rather than arithmetic. Students often failed to detect these errors: first- and second-year physics students (n = 102) rated incorrect AI responses to the most difficult problems as equivalent to correct worked solutions. Instructional studies reported positive attitudes and beneficial learning support, but none measured improvements in critical thinking using validated instruments. Thus, the claim that evaluating AI solutions develops critical thinking is theoretically plausible, grounded in research on learning from errors and critical-thinking skills, but has not yet been empirically demonstrated in physics education. This review proposes an Evaluate–Diagnose–Correct–Reflect framework and a research agenda that distinguishes procedural correctness from conceptual validity.

Keywords: Generative AI; Large Language Models; Physics Education; Critical Thinking

Correspondence: mahautgladys943@gmail.com | **Copyright:** This open access article is distributed under a CC BY-SA License
Article History: Received: 18 April 2026 | Accepted: 10 May 2026 | **Homepage**: https://e-jurnal.iainsorong.ac.id/index.php/great
Citation: Mahaut, G. (2026). Evaluating AI-Generated Physics Solutions: A Systematic Review of Solution Accuracy, Student Evaluation, and Implications for Critical Thinking in Higher Physics Education. CREATE: Contemporary Research in Education and Teaching, 1(1), 32–42. https://doi.org/10.47945/create.v1i1.3355

I. INTRODUCTION

Generative artificial intelligence (AI) has entered higher education faster than the research needed to guide its use. Reviews of AI in higher education have documented rapid growth of the field (Crompton & Burke, 2023), and an earlier systematic review warned that the literature showed little critical reflection on the challenges and risks of AI and only a weak connection to pedagogical theory (Zawacki-Richter et al., 2019). With the release of large language models (LLMs) such as ChatGPT, students gained immediate access to explanations and complete worked solutions. Commentaries and rapid reviews have highlighted the educational promise of these tools while stressing that their integration requires a clear pedagogical approach with a strong focus on critical thinking (Kasneci et al., 2023). Early reviews also noted that ChatGPT's performance varies across subject domains and raised concerns about accuracy and academic integrity (Lo, 2023). Case studies of ChatGPT in education reported enthusiasm alongside concerns about cheating, honesty, and truthfulness (Tlili et al., 2023).

Physics makes these concerns particularly acute. Expertise in physics problem solving is not a matter of producing a correct number. Experts classify problems by underlying principles, whereas novices tend to rely on surface features (Chi et al., 1981), and decades of physics education research show that conceptual understanding, representational competence, and problem solving are intertwined (Docktor & Mestre, 2014). Productive problem solving also involves blending conceptual reasoning with formal mathematics rather than executing algorithms mechanically (Kuo et al., 2013). A solution

can therefore be procedurally correct yet conceptually invalid, for example by reaching the right value through an inappropriate model, or it can be conceptually sound but numerically wrong. Judging an AI-generated physics solution requires exactly the expert-like capacities that introductory students are still developing.

Evidence on how well generative AI solves physics problems is mixed and changes quickly. Kortemeyer (2023) found that ChatGPT would narrowly pass a calculus-based introductory physics course while displaying many preconceptions and errors of a beginning learner. Wang et al. (2024) showed that GPT-4 solved 62.5% of well-specified engineering physics problems but only 8.3% of under-specified ones, mostly because it failed to model the physical situation or to make reasonable assumptions. More recent models perform strongly on standardised instruments: ChatGPT-4o outperformed university cohorts on an electricity and magnetism assessment (Polverini et al., 2025), and reasoning-oriented models outperformed human participants on Physics Olympiad problems (Tschisgale et al., 2025). Yet the same studies document persistent difficulties with graphs, visual representations, and spatial reasoning (Polverini & Gregorcic, 2024b). Understanding why requires recognising that LLMs generate text probabilistically rather than by reasoning from physical principles, a point that has direct implications for how they should be used in teaching (Polverini & Gregorcic, 2024a). The broader natural-language-generation literature describes the associated risk of "hallucination," namely fluent output that is nonsensical or unfaithful to the source (Ji et al., 2023), and the developers of GPT-4 acknowledge that the model can still hallucinate and make reasoning errors (OpenAI, 2023).

The educational consequence of imperfect AI therefore depends on whether learners can recognise imperfection. When 102 first- and second-year physics students rated ChatGPT responses and sample solutions, they struggled to identify flawed responses, and on the most difficult question a ChatGPT response was rated as highly as the correct solution (Dahlkemper et al., 2023). Research on human–automation interaction offers an explanation: automation bias leads users to omit their own checks and to accept incorrect recommendations, and it affects experts as well as novices (Parasuraman & Manzey, 2010). Delegating cognitive work to external tools, known as cognitive offloading, can reduce the internal processing that learning requires (Risko & Gilbert, 2016). Survey evidence has associated frequent AI tool use with lower self-reported critical thinking via increased offloading, although the design is correlational (Gerlich, 2025). A large field experiment in mathematics further showed that unrestricted GPT-4 access improved practice performance but reduced performance once access was removed, whereas a tutor with learning safeguards largely mitigated this harm (Bastani et al., 2025).

The expert consensus on critical thinking describes it as purposeful, self-regulatory judgement involving interpretation, analysis, evaluation, inference, explanation, and self-regulation (Facione, 1990). Assessment of critical thinking requires tasks that actually elicit these processes (Ennis, 1993), and instruction that promotes transfer should include metacognitive monitoring, that is, checking for accuracy and progress (Halpern, 1998). Research on learning from errors indicates that errorful learning followed by corrective feedback can benefit learning (Metcalfe, 2017), and that finding and fixing errors in worked examples can foster far transfer, particularly for learners with sufficient prior knowledge (Große & Renkl, 2007). In physics specifically, Gregorcic and Pendrill (2023) suggested that ChatGPT's plausible but incorrect answers could be used to train students' error detection and critical assessment. Frameworks of AI literacy likewise treat evaluating AI output as a core competence (Ng et al., 2021; UNESCO, 2024).

Despite this convergence, the literature has three gaps. First, a *conceptual gap*: "evaluating AI-generated solutions" is often invoked as a route to critical thinking without specifying what evaluation involves or how it maps onto critical-thinking skills. Second, an *empirical gap*: it is unclear whether any physics study has demonstrated that such evaluation improves critical thinking, as opposed to attitudes

or task performance. Third, a *contextual gap*: general reviews address AI across education or other disciplines, where verification is also emphasised, for example in healthcare (Sallam, 2023), language teaching (Kohnke et al., 2023), and academic integrity (Cotton et al., 2024). A recent physics-education review mapped AI's role as a cognitive partner for critical thinking and recommended assessments that prioritise the critical evaluation of technological output (Lubis et al., 2025), but it did not synthesise the evidence on AI solution accuracy together with the evidence on students' evaluative behaviour.

This review addresses these gaps by asking three research questions (RQs). RQ1: How accurately does generative AI solve or explain physics tasks relevant to higher education, and what types of errors does it make? RQ2: What is known about how students (and prospective teachers) evaluate or use AI-generated physics responses, and what evidence links this evaluation to critical thinking? RQ3: What instructional opportunities and risks follow for higher physics education? The review contributes (a) an error taxonomy derived from physics-specific evidence, (b) a critical appraisal of the evidence connecting AI evaluation to critical thinking, and (c) an instructional framework and research agenda that separate procedural correctness from conceptual validity.

II. RESEARCH METHODS

The review followed the PRISMA 2020 reporting guideline (Page et al., 2021), which requires transparent reporting of the search, selection, appraisal, and synthesis so that the process can be scrutinised and replicated. Studies were eligible if they (1) were peer-reviewed journal articles published between 2020 and 2026; (2) were written in English; (3) were situated in physics education or reported physics-specific tasks; (4) examined generative AI (e.g., LLM-based chatbots) producing physics solutions or explanations, or examined students' or teachers' evaluation or use of such output; and (5) reported empirical data or a documented demonstration. Studies at the upper-secondary level were retained only when they examined students' critical engagement with AI output, because such studies directly inform the transition to higher education; this exception applied to one study. Conference proceedings, editorials, opinion pieces, technical reports, and studies outside physics were excluded from the synthesis corpus but could be cited as context.

Scopus, Web of Science, ERIC, and Google Scholar were searched on June-July 2026. The search combined three concept blocks: ("generative AI" OR "large language model∗" OR ChatGPT OR GPT-4) AND (physics) AND (education OR learning OR student∗ OR teaching). Records were de-duplicated and screened by title and abstract, and the full texts of potentially eligible reports were then assessed (Table 1). Screening was conducted by two reviewer.

Table 1. Study selection according to PRISMA 2020

PRISMA 2020 stage	Number	Notes
Records identified	24	Search date June-July 2026
Duplicates removed	7	Selected from reference manager data base
Records screened (title/abstract)	17	Excluded: 3 because in out of the scope
Full-text reports assessed	14	3 Excluded with reasons low paper quality
Studies included in the synthesis	11	Empirical: 9; conceptual–demonstrative: 2

Empirical studies were appraised with the Mixed Methods Appraisal Tool (MMAT, version 2018), which provides separate criteria for qualitative, quantitative randomised, quantitative non-randomised, quantitative descriptive, and mixed-methods studies (Hong et al., 2018). A standardised form captured bibliographic data, context and level, AI system and version, task type, design, sample, outcome measures, and main findings. Synthesis combined deductive coding with the frame in Table 2, which draws on Facione's critical-thinking skills and on error categories reported in the corpus, and inductive thematic analysis following the phases of familiarisation, coding, theme generation, review, and

definition (Braun & Clarke, 2006). Because the studies used heterogeneous tasks, models, and outcomes, no meta-analysis was attempted.

Table 2. Coding Framework

Code family	Operational definition	Source of the frame
Task type	Numerical problem; context-rich problem; conceptual question; visual representation; Olympiad-level problem	Inductive, refined against the corpus
AI error type	Physical-modelling error; unjustified or missing assumption; representational misreading; conceptualmisuse; internal contradiction; calculation error	Adapted from Wang et al. (2024) and Polverini and Gregorcic (2024b)
Evaluation behaviour	Detecting, diagnosing, correcting, and justifying judgements about AI output; accepting output without verification	Dahlkemper et al. (2023); Parasuraman and Manzey (2010)
Critical-thinking skill	Interpretation, analysis, evaluation, inference, explanation, self-regulation	Facione (1990)
Outcome and measurement	Performance, attitudes, perceived usefulness, critical thinking; type & validity of instrument	Ennis (1993)

III. RESULTS AND DISCUSSION

Eleven studies met the criteria (Table 3). Seven addressed the accuracy and error profile of AI-generated physics solutions (S1–S7), and four examined how students or prospective teachers evaluated or used AI output (S8–S11). All were published between 2023 and 2025, a timing consistent with the public release of ChatGPT in late 2022. Nine were empirical, and two were conceptual–demonstrative case analyses (S2 and S7). The studies were conducted mainly in Northern and Western European and North American institutions; none originated from Indonesia or South-East Asia.

Table 3. Characteristics and main findings of the included studies

No.	Study	Context	Design and sample	Main findings relevant to the review
A. Accuracy and error profile of AI-generated physics solutions				
S1	Kortemeyer (2023)	Calculus-based introductory physics	Case study; assessment content of an actual course graded	ChatGPT would "narrowly pass" the course while showing many preconceptions and errors of a beginning learner
S2	Gregorcic & Pendrill (2023)	Basic mechanics question	Case study of a Socratic dialogue	Responses were unreliable and contradictory; Socratic correction largely failed; authors suggest using plausible-but-wrong answers to train error detection
S3	Wang et al. (2024)	College engineering physics	40 problems (16 well-specified, 24 under-specified) solved by GPT-4	Accuracy 62.5% vs. 8.3% (Fisher's exact test, $p <$ .001); failures in modelling the physical world unreasonable assumptions (8), calculation (6)
S4	Polverini & Gregorcic (2024b)	Test of Understanding Graphs in Kinematics	Performance study of ChatGPT-4 (vision)	Average performance similar to high-school physics students; productive strategies and mostly correct reasoning
S5	Polverini et al. (2025)	Brief Electricity and Magnetism Assessment	Performance study of ChatGPT-4 and ChatGPT-4o	ChatGPT-4o outperformed university student cohorts, yet persistent problems remained with visual interpretation, application of laws, and spatial reasoning
S6	Tschisgale et al. (2025)	German Physics Olympiad problems	Comparison of GPT-4o and o1-preview with human participants	Both models on average outperformed human participants; o1-preview almost consistently outperformed GPT-4o & the human benchmark; prompting had little effect on GPT-4o
S7	Polverini & Gregorcic (2024a)	Conceptual introductory physics tasks	Conceptual–demonstrative paper on LLM functioning & prompting	Explains how LLMs work and shows that prompt engineering can change performance on conceptual tasks; discusses implications for teaching
B. Students and teachers evaluating or using AI output				

No.	Study	Context	Design and sample	Main findings relevant to the review
S8	Dahlkemper et al. (2023)	First- and second-year physics students	Survey experiment; n = 102	Students struggled to distinguish flawed ChatGPT responses from correct ones, especially on unfamiliar topics; on the hardest item a ChatGPT response was rated as good as the sample solution;
S9	Küchemann et al. (2023)	Prospective physics teachers; kinematics task development	Controlled experiment; n = 26 (ChatGPT vs. textbook)	No difference in task correctness; textbook group produced clearer, both groups struggled with task specificity; users reported output-quality difficulties
S10	Kregear et al. (2025)	Introductory physics laboratory	Analysis of student–GPT-4 interactions and expert assessment	Useful for answer verification and theoretical support, but in more than one of ten cases responses were incorrect, or misleading.
S11	Bitzenbauer (2023)	Secondary physics (upper-secondary level)	Pilot intervention; n = 53	Two easy-to-implement activities improved students' views of ChatGPT's usefulness and relevance; critical-thinking outcomes were not measured with a validated test

Accuracy and Error Profile of AI-Generated Physics Solutions

The corpus shows that the accuracy of AI-generated physics solutions is not a fixed property but depends strongly on the task. On standard assessments, performance ranged from a narrow pass in an introductory course (Kortemeyer, 2023) and a level comparable to high-school students on kinematics graphs (Polverini & Gregorcic, 2024b) to performance above university cohorts on an electricity and magnetism assessment (Polverini et al., 2025) and above human participants on Olympiad problems (Tschisgale et al., 2025). The contrast between well-specified and under-specified problems was the sharpest: 62.5% versus 8.3% accuracy (Wang et al., 2024). Because many authentic physics problems are under-specified, this finding is more consequential for teaching than headline scores on standardised tests.

Across studies, the errors that matter most were not computational. In Wang et al. (2024), 14 of 28 incorrect solutions failed because the model did not construct an accurate model of the physical situation, eight failed because of unreasonable assumptions about missing data, and only six involved calculation errors. Studies of visual tasks found that models often proposed productive strategies and verbalised correct reasoning but misread graphs, fields, or circuit diagrams and misapplied the right-hand rule (Polverini & Gregorcic, 2024b; Polverini et al., 2025). Case analyses documented learner-like preconceptions (Kortemeyer, 2023) and internally contradictory answers that did not converge under Socratic questioning (Gregorcic & Pendrill, 2023). Polverini and Gregorcic (2024a) also showed that performance on conceptual tasks can shift with prompt engineering, which implies that the same task may yield different AI output quality for different users. Table 4 summarises these error categories and the corresponding detection demands on students.

Table 4. Error taxonomy of AI-generated physics solutions and implied detection demands

Error category	Description and illustrative evidence	What students must do to detect it
Physical-modelling error	Inaccurate representation of spatial relations or interactions; the most frequent failure mode in Wang et al. (2024)	Construct an independent qualitative model (diagram, free-body diagram) and compare
Assumption error	Unreasonable estimates or unstated assumptions when data are missing, driving the collapse of accuracy on under-specified problems (S3)	Identify missing information and judge the plausibility of assumptions
Representational error	Misreading graphs, vector fields, circuit diagrams, or applying the right-hand rule incorrectly (Polverini & Gregorcic, 2024b; Polverini et al., 2025)	Re-read the representation and translate it into another representation
Conceptual/ principle error	Learner-like preconceptions and misapplied principles (Kortemeyer, 2023)	Check each step against governing principles, not only the final number

Internal inconsistency	Contradictory statements across turns that persist under Socratic questioning (Gregorcic & Pendrill, 2023)	Track claims across the explanation and test them for coherence
Calculation error	Arithmetic or algebraic slips; comparatively infrequent (S3)	Verify units, orders of magnitude, and limiting cases

How Students Evaluate AI-Generated Physics Responses

Only one included study directly measured students' evaluation of AI-generated physics responses against correct solutions. Dahlkemper et al. (2023) found that first- and second-year physics students had difficulty identifying incorrect ChatGPT responses, especially on less familiar topics. On the most difficult item, which concerned fluid dynamics, a flawed ChatGPT response was rated as highly as the sample solution, and students perceived little difference in linguistic quality between the two. The fluency of AI output therefore did not help students separate correct from incorrect physics. Evidence from a peer-reviewed conference study, which is outside the synthesis corpus but consistent with this finding, reported that nearly half of the solutions produced with ChatGPT support were wrongly judged correct by physics students, and that students copied and pasted tasks into ChatGPT far more often than into a search engine (42% vs. 4%) (Krupp et al., 2024).

Studies in which AI was used as a learning or teaching resource add nuance. In a laboratory course, students used a GPT-4 assistant for answer verification and theoretical support, but more than one in ten responses were incorrect, misleading, or confusing (Kregear et al., 2025). Among prospective teachers developing kinematics tasks, ChatGPT support did not change task correctness, but textbook users produced clearer and better-contextualised tasks, and ChatGPT users reported difficulties with output quality (Küchemann et al., 2023). In a secondary-school pilot, two short activities designed to engage students critically with ChatGPT improved their views of its usefulness and relevance (Bitzenbauer, 2023).

Crucially, no included study measured change in critical thinking with a validated instrument after students evaluated AI-generated physics solutions. The outcomes reported were error-detection judgements, perceptions, attitudes, interaction patterns, or task quality. The proposition that evaluating AI output develops critical thinking, which the original motivation of this review assumed, is therefore supported by theory and by adjacent research on learning from errors, but not yet by direct evidence in physics education.

Opportunities and Risks for Higher Physics Education

Table 5 organises the opportunities and risks identified in the corpus and in the theoretical literature. The opportunities centre on AI as a source of feedback, alternative strategies, and, most distinctively, plausible errors that can be used for critique. The risks centre on undetected errors, automation bias, dependence that harms independent performance, and threats to assessment integrity as models outperform students on standard problems.

Table 5. Opportunities and risks of AI-generated physics solutions, with supporting evidence

Aspect	Opportunity (evidence)	Risk (evidence)
Learning support	Answer verification and theoretical support in laboratory work (Kregear et al., 2025); tutoring-like support has historically (VanLehn, 2011)	Problematic responses in more than one in ten laboratory interactions (Kregear et al., 2025)
Error-based learning	Plausible-but-wrong AI answers as material for error-detection training (Gregorcic & Pendrill, 2023); errorful learning with corrective feedback benefits learning (Metcalfe, 2017)	Students may not detect errors, particularly on unfamiliar topics (Dahlkemper et al., 2023)

Problem-solving practice	Exposure to alternative strategies; productive strategies even when execution fails (Polverini & Gregorcic, 2024b)	Performance gains with unrestricted AI access can turn into losses when access is removed (Bastani et al., 2025)
Teacher preparation	Rapid generation of draft tasks for prospective teachers (Küchemann et al., 2023)	Low clarity & contextualisation than textbook supported tasks (Küchemann et al., 2023)
Assessment	Opportunity to redesign assessment towards critique and justification (Tschisgale et al., 2025)	Threats to assessment integrity as models surpass students on standard problems (Tschisgale et al., 2025; Cotton et al., 2024)
AI literacy	Frameworks that include evaluating AI output (Ng et al., 2021; UNESCO, 2024)	Automation bias and cognitive offloading (Parasuraman & Manzey, 2010)

Accuracy Is Task-Dependent, and Its Errors Lie Where Novices Are Weakest

The main finding for RQ1 is that AI accuracy is both improving and uneven, and that the errors that remain are concentrated in physical modelling, assumptions, and representations. These are precisely the dimensions that distinguish experts from novices (Chi et al., 1981; Docktor & Mestre, 2014). A procedurally tidy AI solution may therefore contain a conceptual error that a novice cannot see, because detecting it requires the principle-based representation the novice has not yet built. This interpretation also explains why headline scores can mislead: a model that outperforms students on a multiple-choice inventory (Polverini et al., 2025) may still fail on an under-specified problem that requires modelling decisions (Wang et al., 2024). For instruction, the relevant question is not whether AI is "accurate enough," but whether its errors are the kind students can learn to detect.

Why Students Struggle to Evaluate AI Output

The evidence for RQ2 suggests that students' evaluation difficulties arise from at least three sources. First, fluent language is not a reliable cue for correctness, yet students appear to use it; the similar linguistic ratings for correct and flawed responses reported by Dahlkemper et al. (2023) are consistent with this interpretation. Second, automation bias predicts that users will under-check machine output, and this bias is not eliminated by expertise or simple instruction (Parasuraman & Manzey, 2010). The copy-and-paste behaviour reported by Krupp et al. (2024) is consistent with this prediction. Third, evaluating a solution requires domain knowledge. Error-based learning appears to benefit learners with sufficient prior knowledge (Große & Renkl, 2007), which implies that AI-critique tasks may be least effective for the students who most need support unless the tasks are carefully scaffolded. These explanations are inferences from adjacent literature and should be tested directly in physics contexts.

From Evaluation to Critical Thinking: Mechanism and Conditions

Although direct evidence is lacking, the theoretical case that evaluating AI solutions can develop critical thinking is coherent. Each step of a well-designed critique maps onto the skills identified in the expert consensus (Facione, 1990): interpreting the problem and the AI's representation, analysing its argument, evaluating the validity of its assumptions, inferring where it went wrong, explaining a corrected solution, and regulating one's own judgement. Four conditions appear necessary for this potential to be realised. The first is an *independent attempt* before seeing AI output, which counters offloading (Risko & Gilbert, 2016). The second is *corrective feedback* on students' judgements, because errors are productive when they are followed by correction and analysis of the reasoning that produced them (Metcalfe, 2017). The third is *explicit metacognitive monitoring*, which supports transfer (Halpern, 1998). The fourth is *design guardrails* that prevent AI from becoming a crutch (Bastani et al., 2025). Blending conceptual and formal reasoning (Kuo et al., 2013) offers a concrete criterion for these tasks: students should judge not only whether the mathematics is executed correctly, but also whether the equations express a valid physical model.

An Evaluate–Diagnose–Correct–Reflect Framework

Building on this synthesis, Table 6 proposes an Evaluate–Diagnose–Correct–Reflect (EDCR) framework for AI-critique tasks in physics courses. The framework is a design proposal derived from the literature, not an empirically validated model. Its distinctive features are the requirement for an independent attempt, the use of the error taxonomy in Table 4 as a diagnostic scaffold, and a reflection phase that asks students to distinguish procedural correctness from conceptual validity. Because prompt wording and model version affect AI output (Polverini & Gregorcic, 2024a), instructors should pre-test and archive the AI solutions used in class so that all students critique the same artefact.

Table 6. Proposed Evaluate–Diagnose–Correct–Reflect (EDCR) framework (for empirical testing)

Phase	Student activity with an AI-generated	Critical-thinking skills engaged	Design safeguard
Evaluate	Solve the problem independently first; then judge whether the AI solution is plausible	Interpretation, evaluation	Require an independent attempt before AI output is revealed
Diagnose	Classify the error (modelling, assumption, representation, inconsistency, calculation)	Analysis, inference	Provide the error taxonomy (Table 4) as a scaffold
Correct	Produce a corrected solution and justify each change with physical principles	Explanation	Corrective feedback from instructor or peers
Reflect	Explain why the error was detected and what cue to use next time; distinguish procedural correctness from conceptual validity	Self-regulation	Structured reflection prompts; confidence ratings before and after

AI-Assisted Laboratory and Video Analysis as a Future Direction

The original draft of this review proposed AI-assisted video analysis as a promising application. Video analysis already has an established place in physics teaching, where students track motion and model kinematic data (Brown & Cox, 2009), and LLM assistants are beginning to appear in laboratory courses (Kregear et al., 2025). However, the corpus contains no study of AI-based automatic tracking or interpretation in physics video analysis. This application should therefore be treated as a research direction. The same EDCR logic applies: AI-generated tracks, fits, and interpretations should be checked against theoretical expectations, measurement uncertainty, and alternative models before being accepted.

Implications for Teaching, Assessment, and AI Literacy

Three implications follow. First, assessment should reward critique and justification rather than final answers alone, since models now match or exceed students on many standard problems (Tschisgale et al., 2025); clear guidance on responsible use remains necessary to protect integrity (Cotton et al., 2024). Second, AI literacy in physics courses should include domain-specific evaluation, meaning the ability to recognise modelling, assumption, and representational errors, in line with frameworks that treat evaluating AI as a core competence (Ng et al., 2021; UNESCO, 2024). Third, AI support should be designed as a tutor with safeguards rather than an answer engine. Step-based tutoring systems have historically approached the effectiveness of human tutoring (VanLehn, 2011), but unrestricted generative AI can undermine independent performance (Bastani et al., 2025).

Table 7 summarises the research needed to move from plausible claims to established evidence. The most urgent need is for controlled studies that measure critical thinking or physics reasoning with validated instruments, examine the moderating role of prior knowledge, and include delayed, AI-free measures. Because models change rapidly, studies should report model versions, dates, and prompts.

Table 7. Research agenda

Research need	Recommended design feature
Causal evidence that AI-critique tasks develop critical thinking	Randomised or quasi-experimental designs comparing AI-critique tasks with critique of human-generated erroneous examples

Research need	Recommended design feature
Valid measurement	Validated critical-thinking or physics-reasoning instruments, plus process data (think-aloud, error-detection accuracy, confidence calibration)
Prior-knowledge moderation	Stratify by prior physics knowledge to test whether error-based tasks benefit only stronger students
Durability and dependence	Delayed and AI-free post-tests to detect "crutch" effects
Model drift	Report model version, date, and prompts; re-test tasks as models change
Context	Studies in under-represented higher-education systems.

This review has several limitations. The corpus is small and recent, reflecting the short history of generative AI in education. Most studies evaluated specific model versions that have since been superseded, so accuracy findings may date quickly. Only one study directly measured students' evaluation of AI responses. Excluding conference papers may have omitted relevant evidence, which is why one such study was used only as context. Finally, the corpus is concentrated in European and North American institutions, which limits transferability to other higher-education systems.

IV. CONCLUSION

This review examined the accuracy of AI-generated physics solutions, students' evaluation of such solutions, and the implications for critical thinking in higher physics education. For RQ1, generative AI performs strongly on many standard physics tasks, but its accuracy drops sharply on under-specified problems and visual representations, and its remaining errors are concentrated in physical modelling, assumptions, and representations rather than calculation. For RQ2, students frequently fail to detect these errors, and fluent language offers no reliable cue. No physics study has yet shown that evaluating AI output improves critical thinking when measured with a validated instrument. For RQ3, AI offers genuine opportunities, especially plausible errors that can be used for critique, but also risks of undetected misconceptions, automation bias, dependence, and threats to assessment integrity. The educational value of AI-generated physics solutions therefore lies less in the answers they provide than in the structured evaluation they can provoke. Whether that evaluation actually builds critical thinking remains an open empirical question. The proposed EDCR framework and research agenda offer a starting point for answering it rigorously.

V. REFFERENCE

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

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

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, D., & Cox, A. J. (2009). Innovative uses of video analysis. *The Physics Teacher, 47*(3), 145–150. https://pubs.aip.org/aapt/pte/article-abstract/47/3/145/275154/Innovative-Uses-of-Video-Analysis

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

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

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics—Physics Education Research, 10*(2), Article 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. https://doi.org/10.1080/00405849309543594

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (ERIC Document No. ED315423). American Philosophical Association. https://eric.ed.gov/?id=ED315423

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

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

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction, 17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

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

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

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

Kregear, T., Babayeva, M., & Widenhorn, R. (2025). Analysis of student interactions with a large language model in an introductory physics lab setting. *International Journal of Artificial Intelligence in Education, 35*, 2993–3016. https://doi.org/10.1007/s40593-025-00489-3

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance—Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In *HHAI 2024: Hybrid human AI systems for the social good* (Frontiers in Artificial Intelligence and Applications). IOS Press. https://doi.org/10.3233/FAIA240195

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

Kuo, E., Hull, M. M., Gupta, A., & Elby, A. (2013). How students blend conceptual and formal mathematical reasoning in solving physics problems. *Science Education, 97*(1), 32–57. https://doi.org/10.1002/sce.21043

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Lubis, H., Van Harling, V. N., & Panunggul, V. B. (2025). Enhancing critical thinking in physics education through AI: A systematic literature review of trends and pedagogical implications. *Jurnal Pendidikan dan Ilmu Fisika, 5*(2), 343–354. https://doi.org/10.52434/jpif.v5i2.43353

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

OpenAI. (2023). *GPT-4 technical report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Polverini, G., & Gregorcic, B. (2024a). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics, 45*(2), Article 025701. https://doi.org/10.1088/1361-6404/ad1420

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

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), Article 010154. https://doi.org/10.1103/PhysRevPhysEducRes.21.010154

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare, 11*(6), Article 887. https://doi.org/10.3390/healthcare11060887

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*, Article 15. https://doi.org/10.1186/s40561-023-00237-x

Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S., & Wulff, P. (2025). Evaluating GPT- and reasoning-based large language models on Physics Olympiad problems: Surpassing human performance and implications for educational assessment. *Physical Review Physics Education Research*. https://doi.org/10.1103/6fmx-bsnl

UNESCO. (2024). *AI competency framework for students*. UNESCO. https://www.unesco.org/en/articles/ai-competency-framework-students

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

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*, Article 1330486. https://doi.org/10.3389/feduc.2023.1330486

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0