Editors' Suggestion

Trust-utility gap in introductory physics education: Students' adoption, domain-specific skepticism, and preferences for AI integration

Emadelden Fouad[*] and Ian Bentley[†]

Department of Physics, Florida Polytechnic University,
4700 Research Way, Lakeland, Florida 33805, USA

(Received 28 February 2026; accepted 26 June 2026; published 6 August 2026)

We report results from a mixed-methods survey study ($N = 81$) examining how undergraduate students in introductory physics courses use, trust, and prefer to integrate artificial intelligence (AI) tools into their learning. Among respondents, 91% (95% confidence interval (CI): [83%, 96%]) reported using AI for physics coursework, yet only 41% (95% CI: [30%, 52%]) trusted AI-generated physics explanations—a 50-percentage-point trust-utility gap consistent with domain-calibrated skepticism rather than uncritical adoption. Thematic analysis of open-ended responses ($n = 67$; Cohen's $\kappa = 0.78$) identified eight themes; the most distinctly physics-specific finding was that 40% of qualitative respondents spontaneously articulated *where*AI fails *in physics*—visual-spatial reasoning, circuit analysis, and abstract physical reasoning—aligning with physics education research on student difficulties with multirepresentational tasks. This skepticism is also consistent with Kortemeyer's benchmarking results, which show that AI performance is weakest on precisely the multirepresentational task types that students identified as failure-prone, suggesting that student trust calibration may track underlying AI competence boundaries in physics. A majority (65%) preferred optional over mandatory AI integration. Because high-performing students were overrepresented by 13.9 percentage points, the estimated preference for optional integration (and/or resistance to mandatory) may be inflated in this sample. These findings highlight a critical unresolved priority—the verification gap between self-reported and actual verification behavior—and inform recommendations for AI-integrated physics pedagogy.

DOI: 10.1103/bdl1-6bn4

I. INTRODUCTION

Understanding how students learn physics requires analyzing the tools they use and how those tools interact with the discipline's distinctive epistemic demands. Physics education presents a unique constellation of learning challenges: abstract conceptual reasoning, mathematical formalism, visual-spatial interpretation of diagrams and graphs, dimensional analysis, and the coordination of multiple representations [1,2]. Students who struggle to interpret a velocity-time graph, construct a free-body diagram, or trace current through a multiloop circuit face challenges that are fundamentally visual-spatial and representational—not merely mathematical.

Large language models (LLMs)—now widely adopted by undergraduates—enter this learning environment, yet their consequences for how students construct physical understanding remain poorly characterized. A central concern motivating this line of inquiry, and one that will require longitudinal, observational, and experimental studies to resolve, is whether students use AI to genuinely build physical intuition or use it to bypass the productive struggle that physics education research (PER) identifies as central to conceptual development [3]. The present study takes a first empirical step by characterizing usage patterns, trust calibration, and integration preferences, while treating the broader question as a horizon for a long-term research program rather than a claim the present study adjudicates.

Since ChatGPT's public release in November 2022, AI tools have achieved unprecedented adoption rates, reaching 100 million users within two months [4]. Recent surveys indicate that 86% of educational organizations now use generative AI [5]. Despite this prevalence, empirical research on student usage patterns *in physics education* remains limited. Most existing work benchmarks AI performance on physics problems [6,7] rather than examining how physics students integrate these tools into their learning practices that are shaped by the discipline's particular demands.

PER has established that persistent misconceptions, equation-hunting strategies, and weak conceptual-mathematical integration are hallmarks of novice physics

[*]Contact author: efouad@floridapoly.edu
[†]Contact author: ibentley@floridapoly.edu

learners [1,2]. The Resources Framework [1] and the construct of *epistemic framing* [8] suggest that how students engage with explanatory tools shapes the conceptual resources they activate and retain. If students use AI to bypass effortful sensemaking [3], widespread adoption could entrench surface-level learning at scale; if they use it as interactive scaffolding, AI could extend the benefits of active engagement [9,10].

Research on multirepresentational reasoning in physics [2] demonstrates that students must coordinate symbolic, graphical, and diagrammatic representations to develop expertlike physical thinking. AI's documented weakness in precisely these visual-spatial domains [6,11] creates a structural mismatch: the representations most central to deep physics learning are those AI handles least reliably. This mismatch provides the motivating tension for the present study.

We address this gap by examining how students at a science, technology, engineering, and mathematics (STEM)-focused technical university use AI tools for physics, which physics-domain limitations they recognize, and how they prefer these tools to be integrated into their courses. Our central finding is a pronounced *trust-utility gap*: 91% of students use AI for physics, yet only 41% trust AI-generated physics explanations—a 50-percentage-point divergence with nonoverlapping 95% confidence intervals, suggesting domain-calibrated pragmatism rather than uncritical dependence. This work contributes to the emerging AI-focused literature in PER, building on recent experimental and benchmark studies [12–14] by providing an empirical characterization of how introductory physics students navigate trust and autonomy when using LLMs to build physical understanding.

A. Theoretical framework

Our investigation draws on two complementary frameworks from PER and educational psychology, selected for their direct relevance to AI tool use in physics learning contexts.

Self-determination theory (SDT) [15,16] posits that learner autonomy—the sense of choice and self-direction—is central to intrinsic motivation and engagement. We selected SDT because it addresses how the *manner* of technology integration affects motivation, a question particularly salient in physics, where intrinsic motivation and epistemic curiosity are associated with deeper conceptual engagement [17,18]. Unlike standard technology-acceptance models [19], SDT predicts that externally mandated tool use can undermine intrinsic motivation even when objectively beneficial, thereby informing the counterintuitive finding that high-achieving students oppose mandatory AI integration more than struggling students.

Constructivist learning theory [3] provides a complementary lens: students who use AI as scaffolding engage in active knowledge construction, while students who extract answers passively may reinforce the novice equation-hunting patterns documented in PER [1–3,20].

B. Conceptual framework: Trust-autonomy-utility

Building on self-determination theory (SDT) and constructivist principles, we position physics AI tool integration at the intersection of three conceptually distinct dimensions:

1. **Utility**: Perceived practical value for physics learning tasks, including homework, exam preparation, conceptual understanding, and problem solving.
2. **Trust**: Confidence in the accuracy and reliability of AI-generated physics content, including quantitative solutions, conceptual explanations, and diagram-based reasoning.
3. **Autonomy**: Student agency over whether, when, and how to use AI tools in physics coursework.

The *trust-utility gap* is defined as the percentage-point difference between the proportion of students who use AI for physics (utility adoption) and the proportion who trust AI-generated physics explanations (trust endorsement). Operationalizations of each dimension—including specific survey items, Likert formats, and the precise trust-endorsement criterion—are detailed in Sec. III C.

Effective integration must balance all three dimensions. High utility without trust produces pragmatic but skeptical use, which aligns with the pattern observed in this study. Conversely, high trust without autonomy may undermine intrinsic motivation, while high autonomy without adequate utility or trust is likely to lead to nonadoption. Figure 1 illustrates the seven zones produced by these intersections. This framework guides our analysis and interpretation of the student survey data.

C. Research questions

1. **Utility and adoption:** What are the usage patterns and perceived effectiveness of AI tools among undergraduate physics students, and which physics-specific tasks do students view as best and least suited to AI assistance?
2. **Trust and skepticism:** How do students' trust levels in AI-generated physics content relate to their usage rates, and what does the divergence between these measures reveal about the trust-utility gap?
3. **Autonomy and integration:** What AI integration preferences do physics students express, and how do these preferences interact with autonomy considerations and prior academic performance?

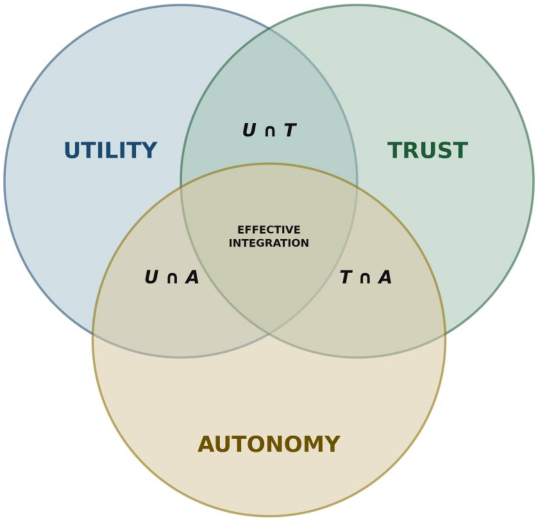

FIG. 1. Trust-autonomy-utility framework for AI integration in physics education. Each exclusive zone represents an imbalanced state: *Utility only* (passive, answer-seeking use without critical evaluation); *Trust only* (high confidence in AI accuracy but low practical engagement, trusting AI without adopting it); *Autonomy only* (principled optional use with minimal adoption). Pairwise intersections capture partial balance: *Utility* ∩ *Trust* without autonomy risks motivation loss under mandatory integration; *Trust* ∩ *Autonomy* without utility leads to nonadoption; and *Utility* ∩ *Autonomy* without trust, the pattern observed in this study (91% use, 41% trust), characterizes pragmatic, domain-calibrated skepticism with self-directed rather than institutionally mandated engagement. The center zone (all three) represents effective integration. Not all intersections are fully explored in the present study; the framework is offered as a conceptual tool for instructors and a scaffold for future AI-PER research.

II. BACKGROUND

A. Physics education research context

This study is situated within the PER tradition of examining how students develop physical understanding through interaction with representational and computational tools. McDermott and Redish [2] documented systematic difficulties with multirepresentational reasoning—velocity-time graphs, force diagrams, and circuit schematics—that persist even after instruction. The resources framework [1] provides a theoretical account of why: novice students activate fragmented "knowledge in pieces" rather than coherent physical models, leading to equation-hunting strategies that produce correct numerical answers without conceptual understanding. Hake's landmark study [20] established that interactive-engagement instruction yields learning gains roughly twice those of traditional instruction (mean normalized gain $\langle g \rangle = 0.48$ versus 0.23). These results make active engagement a widely used benchmark against which AI-tutoring interventions must be evaluated.

B. AI performance and student perceptions in physics

Research on LLM performance in physics reveals a pattern of competence on well-structured quantitative tasks and failure on tasks requiring visual-spatial reasoning or open-ended physical judgment. GPT-4 demonstrated 90% accuracy in categorizing introductory physics questions [7] and showed high accuracy on well-defined problems (62.5% success rate) but a sharp decline on underspecified scenarios (8.3% success rate) [6]. This asymmetry is particularly significant in physics, where problems frequently involve ambiguous physical situations, free-body diagrams, circuit schematics, and graphical data [11]—precisely the representations where current LLMs perform most poorly and where PER-identified student difficulties are most persistent [2].

Students who have used AI for physics coursework generally report positive perceptions [21,22], with limitations including difficulties with visual information and computational errors [11,23]. Sirnoorkar *et al.* [14] examined student and AI responses to physics problems through the lenses of sensemaking and mechanistic reasoning—both established PER constructs [1]—finding that students engaged more deeply in physical reasoning when AI responses were flawed and required critique. This suggests that AI errors in physics may carry pedagogical value if students are equipped to identify them, but this requires that students possess sufficient domain knowledge to detect those errors in the first place.

Kortemeyer [13] tested whether an AI agent could pass an introductory physics course, finding that while AI could handle structured calculations, it struggled with conceptual explanations, graph interpretation, and tasks requiring physical intuition—an AI performance profile that maps directly onto PER's taxonomy of novice-to-expert transitions in physics.

C. AI tutoring and scalability in physics

One-on-one human tutoring significantly improves physics learning outcomes (effect sizes approaching 2 standard deviations) [24], but it is logistically unscalable; intelligent tutoring systems show moderate positive effects [25,26]. In a particularly relevant recent study, Kestin *et al.* [12] reported that AI tutoring matched traditional active-learning outcomes (mean normalized gain comparable to peer instruction) in an introductory physics course. While this is a striking finding when evaluated against Hake's established benchmark [20], the authors noted that accuracy concerns and domain-specific pedagogical appropriateness—particularly for multirepresentational physics reasoning—remain unresolved.

D. AI literacy and critical evaluation in STEM

AI literacy is rapidly emerging as an essential educational competency [27–30]. While foundational frameworks defined AI literacy in terms of broad technological understanding [30], the advent of generative AI has shifted the focus toward applied competencies. Modern syntheses of STEM AI literacy [27,29] emphasize critical skills such as understanding AI limitations, effective prompting, critical evaluation of outputs, ethical use, and, crucially, *domain-specific calibration*—the ability to distinguish, on the basis of direct experience, where within a discipline AI performs reliably and where it fails. In physics, for example, domain-specific calibration means trusting AI for algebraic step-by-step solutions while distrusting it for circuit diagram analysis or free-body diagram interpretation, rather than holding a uniform skepticism across all task types. This last competency is particularly salient in physics, where disciplinary norms demand quantitative precision, conceptual coherence, and multirepresentational fluency [2]—all areas where current language models have documented weaknesses.

A critical methodological concern for all self-report studies in this domain—including the present one—is the gap between reported and actual verification behavior. Research on metacognitive monitoring consistently shows that students overestimate the frequency and quality of their own checking strategies, particularly under time pressure [31,32]. This "verification gap" is especially consequential in physics, where unchecked AI errors could crystallize specific misconceptions—such as conflating force with momentum or misapplying conservation laws—that PER has shown are highly resistant to correction [1,2]. Future work using conversation log analysis could provide the behavioral verification data that survey methods alone cannot.

E. Student autonomy and technology integration

Regarding equity, recent PER work suggests a "competency trap" in AI-assisted learning: effective use requires sufficient prior physics knowledge to distinguish accurate explanations from plausible hallucinations [33]. This creates a risk that lower-performing students—who most need scaffolding—are least equipped to verify AI outputs, leaving them vulnerable to specific misconceptions that may compound across sequential courses [13].

III. METHODOLOGY

A. Context and participants

This study was conducted at Florida Polytechnic University, a STEM-focused public institution, following approval by the Institutional Review Board (IRB). During the Fall 2025 semester, students enrolled in three introductory physics courses were invited to participate in an anonymous online survey. Informed consent was obtained from all participants prior to data collection. Of the 447 eligible students, 81 responded, yielding an 18.12% response rate.

TABLE I. Demographic characteristics of the sample ($N = 81$).

Characteristic	n	%
Physics course		
Calculus-based physics I (PHY 2048)	36	44
Calculus-based physics II (PHY 2049)	32	40
Algebra-based physics I (PHY 2053)	13	16
Self-reported grade range		
A (90%–100%)	39	48
B (80%–89%)	28	35
C (70%–79%)	11	14
D (60%–69%)	3	4
Physics interest level		
High interest	26	32
Moderate interest	46	57
Low interest	9	11
AI tool usage		
Yes	74	91.4
No	6	7.4
Maybe	1	1.2

The anonymous design protected student privacy and encouraged candid responses regarding AI usage, but it precluded direct nonresponse follow-up. An analysis of institutional final grade data ($N = 447$) indicated a 13.9-percentage-point overrepresentation of high-performing students in our sample (83.0% self-reported A/B versus 69.1% institutional A/B), an effect discussed further in Sec. V H.

Participants were enrolled in calculus-based Physics I (PHY 2048, $n = 36$), calculus-based Physics II (PHY 2049, $n = 32$), or algebra-based Physics I (PHY 2053, $n = 13$). Detailed demographic characteristics of the sample are provided in Table I.

B. Survey instrument

The survey comprised 29 items distributed across six categories: demographics, usage patterns, effectiveness assessment, comparison with traditional resources, integration preferences, and qualitative feedback [34]. The survey design was based on previous works, specifically, Refs. [35,36]. Questions employed a mix of Likert scales, multiple-choice formats, and open-ended responses. The integration preferences question allowed multiple selections, enabling students to indicate support for some modalities while opposing others (e.g., supporting optional study help while opposing mandatory lecture integration). Survey completion time was approximately 15 min.

1. Instrument validity

Content validity was established through iterative expert review: both authors independently mapped each item to the trust-autonomy-utility framework and to PER-identified physics learning challenges, revising or removing items lacking clear construct correspondence. Trust was operationalized via a primary item ("I trust the accuracy of AI-generated physics explanations") paired with a supplementary confidence item; directional agreement was 91%, supporting the single-item measure for this exploratory study [37]. We acknowledge this as a limitation relative to validated multi-item scales (e.g., trust in automation [38]); future studies should adopt such instruments.

C. Data collection and analysis

Data were collected during weeks 10–13 of the Fall 2025 semester via an anonymous online platform, allowing students to reflect on substantial AI experience while avoiding end-of-semester pressures.

1. Quantitative analysis

Descriptive statistics summarized closed-ended responses. We report 95% confidence intervals (CIs) for key proportions using the Wilson score interval for a binomial proportion ($\alpha = 0.05$). Chi-square tests of independence examined associations between categorical variables, with Cramér's V reported as the effect size [39]. In reporting results, we prioritize effect sizes and CIs to support interpretation of practical significance, consistent with contemporary statistical guidance that cautions against sole reliance on p values [40].

Operational definitions. *AI adoption* is defined as selecting "Yes" to the AI use item (Table I). *Trust endorsement* is defined as selecting "Agree" or "Strongly agree" in response to the item "I trust the accuracy of AI-generated physics explanations" (Neutral is not counted as a trust endorsement). The *trust-utility gap* is reported as the percentage-point difference between the adoption proportion and the trust-endorsement proportion. For interpretive purposes, we treat a gap exceeding 30 percentage points as a prima facie indicator of domain-calibrated skepticism, while emphasizing that the gap is a continuous measure; our observed 50-point gap is qualitatively large under any reasonable heuristic threshold. *Domain-calibrated skepticism* refers to a state in which pragmatic use persists despite explicit reservations about AI accuracy, grounded in the student's direct experience of domain-specific AI failures rather than generalized distrust of technology.

Sample size and statistical power. *Post hoc* power analysis (G * Power 3.1 [41]; $N = 81$, $\alpha = 0.05$) indicated 80% power to detect medium effects (Cramér's $V \geq 0.30$) and 45% power to detect small effects ($V = 0.10$) in chi-square tests of independence. Subgroup comparisons, particularly those involving C/D students ($n = 13$) or algebra-based students ($n = 13$), were substantially underpowered (approximately 25% power for $V = 0.30$) and are therefore reported descriptively. All underpowered comparisons are explicitly identified in the Results section.

2. Qualitative analysis

Of the 81 respondents, 67 (83%; mean of 2.3 responses per student) provided written answers to at least one open-ended question. Thematic analysis [42,43] proceeded through four phases: (i) familiarization through repeated reading; (ii) independent code generation by both authors; (iii) code comparison and refinement, with an interrater reliability of Cohen's $\kappa = 0.78$ indicating substantial agreement [44]; and (iv) iterative identification of themes. Disagreements were resolved through a consensus review of the original responses.

We employed a hybrid approach: initial codes were generated inductively from student responses, then organized deductively using the trust-autonomy-utility framework. The interrater reliability of $\kappa = 0.78$ supports the trustworthiness of the findings.

D. Ethical approval

This study was approved as exempt by the Florida Polytechnic University Institutional Review Board (IRB Approval No. 25013, approved October 30, 2025, valid through October 30, 2026). Participation was voluntary and anonymous; no identifying information was collected.

IV. RESULTS

A. Academic performance and interest profile

The sample was skewed toward high performers: 83% self-reported A or B grades versus 69.1% institutional A/B final grades—a 13.9-percentage-point overrepresentation (see Sec. V H for full implications). This overrepresentation compounds a well-documented systematic bias in PER samples: Kanim and Cid [45] demonstrated that calculus-based introductory physics courses disproportionately enroll students who are demographically and academically unrepresentative of the broader undergraduate population. Our sample thus lies at the intersection of two compounding selection effects—disciplinary and survey response—both of which warrant caution when generalizing the findings. The institutional distribution by course was as follows: PHY 2048 ($n = 222$): 61.7% A/B, 34.2% C/D, 4.1% F; PHY 2049 ($n = 193$): 76.7% A/B, 21.8% C/D, 1.6% F; PHY 2053 ($n = 32$): 75.0% A/B, 12.5% C/D, 12.5% F.

Student interest in physics was predominantly moderate (57%), with high interest (32%) substantially exceeding low interest (11%). The high AI adoption across all interest levels suggests that AI usage is not limited to highly motivated physics students. A descriptive examination of the trust-utility gap by interest level reveals that the gap is present across all groups (high interest: use 96.2%, trust 38.5%, gap = 57.7 pp, $n = 26$; moderate: use 91.3%, trust 47.8%, gap = 43.5 pp, $n = 46$; low: use 77.8%, trust 11.1%, gap = 66.7 pp, $n = 9$). The large gap among low-interest students reflects pragmatic use with little domain

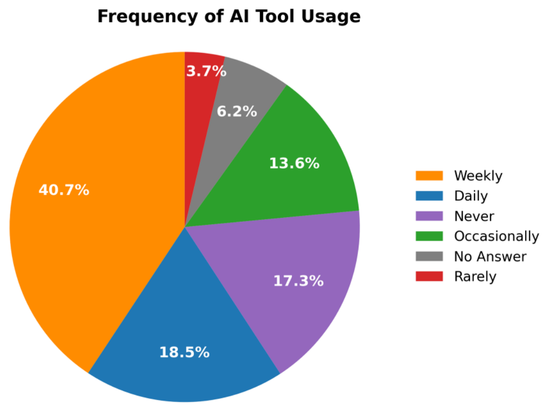

FIG. 2. Frequency of AI tool usage for physics coursework ($N = 81$). Weekly usage was most common (40.7%), with 59.2% using AI tools daily or weekly—indicating AI has become a routine rather than an occasional resource for most physics students.

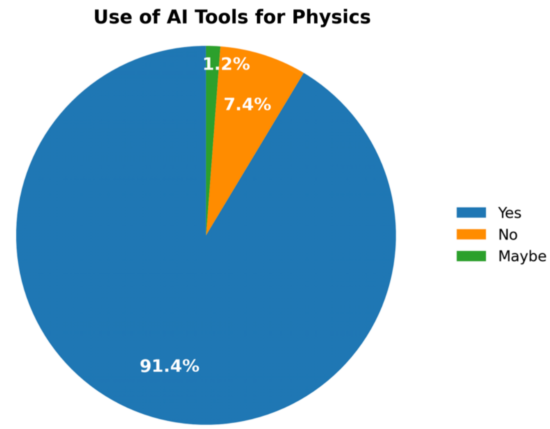

FIG. 3. AI tool usage among physics students ($N = 81$): 91.4% reported using AI despite only 41% trusting its accuracy—the trust-utility gap.

knowledge to anchor trust judgments; however, the small cell size ($n = 9$) precludes inferential analysis.

B. Usage patterns and physics-specific applications

Among the 81 respondents, 74 (91%; 95% CI: [83%, 96%]) reported using AI for physics coursework. Usage frequency (Fig. 2) varied: daily (19%, $n = 15$), weekly (41%, $n = 33$), occasionally (14%, $n = 11$), rarely (4%, $n = 3$), and never (17%, $n = 14$), with 5 students (6%) not answering the frequency question despite reporting AI use. Weekly usage was most common (40.7%), and 59.2% of students used AI daily or weekly.

TABLE II. AI usage patterns and trust indicators among physics students. Percentages for use cases are of the 74 AI users. 95% binomial confidence intervals are provided for primary measures.

Measure	n	%	95% CI
Primary use cases (multiple selections)			
Exam preparation	56	76	[64%, 85%]
Homework problem solving	53	72	[60%, 81%]
General physics questions	50	68	[56%, 78%]
Understanding lectures	46	62	[50%, 73%]
Trust in AI accuracy			
Strongly agree/agree	30	41	[30%, 52%]
Neutral	30	41	[30%, 52%]
Disagree/strongly disagree	14	19	[11%, 29%]
Perceived homework helpfulness			
Very/somewhat helpful	68	92	[84%, 97%]
Neutral	4	5	[1%, 13%]
Not helpful	2	3	[0%, 10%]
Exam preparation helpfulness			
Very/somewhat helpful	57	77	[66%, 86%]
Neutral	11	15	[8%, 25%]
Not helpful	6	8	[3%, 17%]

Students applied AI across multiple physics-related tasks (Table II). The most common applications were exam preparation (76%, $n = 56$), homework problem solving (72%, $n = 53$), general physics questions (68%, $n = 50$), and understanding lecture content (62%, $n = 46$).

C. The trust-utility gap in physics

The central quantitative finding is a pronounced mismatch between AI usage and trust, operationally defined in Sec. I B. As shown in Fig. 3, 91% of students reported

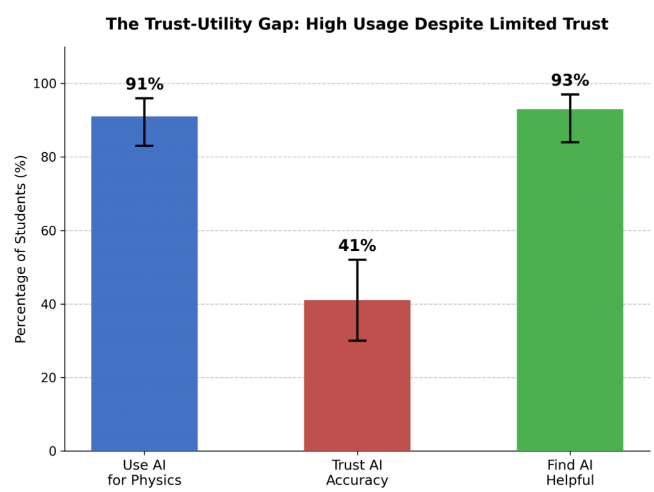

FIG. 4. The trust-utility gap in physics AI use: 91% adoption, 92% perceived helpfulness, yet only 41% trust in AI-generated physics explanations—a 50-percentage-point gap between use and trust. Bars represent sample proportions with 95% Wilson score confidence interval error bars: adoption [83%, 96%], trust [30%, 52%], helpfulness [84%, 97%]. The nonoverlapping CIs for adoption and trust directly confirm that the gap is not attributable to sampling variability. The pattern is consistent with domain-calibrated skepticism shaped by students' direct experience of AI failures in physics-specific tasks (circuit analysis, diagram interpretation, and symbolic derivations).

TABLE III. Trust level in AI-generated physics explanations by grade range and usage frequency ($N = 74$ AI users). Cell values are counts; cells with $n < 5$ (marked †) are underpowered and should be interpreted as descriptive trends only.

	Trust level		
Subgroup	High(Agree)	Neutral	Low(Disagree)
Grade Range			
A/B grades ($n = 64$)	27	26	11
C/D grades ($n = 10$)†	4	4	2
Usage frequency			
Daily ($n = 15$)	8	5	2†
Weekly ($n = 33$)	11	16	6
Occasional/never ($n = 26$)	11	9	6

using AI for physics coursework. Yet despite this high adoption rate (Fig. 4), only 41% ($n = 30$; 95% CI: [30%, 52%]) agreed or strongly agreed that they trusted AI-generated physics explanations; 41% ($n = 30$) were neutral, and 19% ($n = 14$) expressed explicit distrust. Additionally, 92% found AI helpful for homework—further widening the contrast between utility and trust. The trust-utility gap is thus 50 percentage points (91% use minus 41% trust), far exceeding the 30-point threshold defined in Sec. III C as evidence of domain-calibrated skepticism. The nonoverlapping 95% CIs for adoption [83%, 96%] and trust [30%, 52%] support the interpretation that the gap is substantial in this sample and unlikely to be explained by sampling variability alone. Confidence levels followed the same pattern: only 12% ($n = 9$) reported being extremely confident in AI explanations, while 43% ($n = 32$) described themselves as only somewhat confident.

This pattern is further explored in Table III, which cross-tabulates trust level against grade range and usage frequency. High-trust students (agree/strongly agree) were concentrated among daily users regardless of grade, while weekly users showed the most neutral trust.

Computing the trust-utility gap separately by grade: A/B students (use 94.1%, trust 41.2%, gap = 52.9 pp, $n = 68$) versus C/D students (use 76.9%, trust 38.5%, gap = 38.5 pp, $n = 13$)—consistent with high-performing students using AI more intensively while maintaining domain-calibrated skepticism. The C/D subgroup is underpowered; these are descriptive observations requiring replication.

D. Perceived learning outcomes

Students reported nuanced perceptions of whether AI supports physics understanding or merely provides quick answers (Fig. 5, Table IV). Nearly half (49%, $n = 39$) indicated AI provides both understanding and quick answers, while 35% ($n = 28$) emphasized deep conceptual understanding. Only 9% ($n = 7$) characterized AI as primarily providing quick answers without understanding, and 6% ($n = 5$) reported neither benefit. Taken together, 84% (95% CI: [74%, 91%]) of respondents perceived AI as contributing to understanding—a pattern inconsistent with the concern that AI serves primarily as an answer key in physics courses (see Sec. V C).

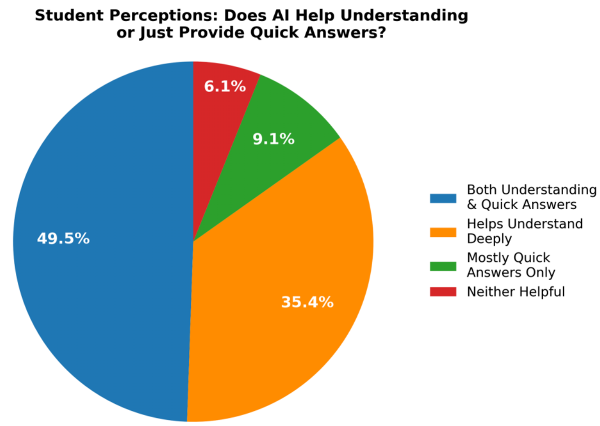

FIG. 5. Student perceptions of whether AI helps with understanding or merely provides quick answers ($N = 79$). The majority (84%; 95% CI: [74%, 91%]) report that AI contributes to physics understanding, with only 9% characterizing it as primarily an answer source.

More than half of respondents (57%, $n = 46$) reported that AI changed their approach to learning physics, with 22% ($n = 18$) indicating significant changes and 35% ($n = 28$) reporting slight changes.

E. Comparison with traditional physics resources

Compared to textbooks (Fig. 6), 92% of students indicated AI helped them understand physics concepts better: 51% responded "Yes" definitively and 41% "Somewhat"; only 8% preferred textbooks. This preference likely reflects AI's interactive, conversational approach—students can request alternative explanations, ask why a formula is derived, or probe limiting cases—versus the static presentation of most physics texts. Preference does not establish superior learning outcomes, however; a student who perceives AI explanations as clearer may nonetheless construct less coherent physical models if the interaction is passive rather than effortfully sensemaking [3].

Descriptive characterization of the six textbook-preferring students reveals that this preference is associated

TABLE IV. Student perceptions of AI's effect on physics understanding ($N = 79$).

Response category	n	%
Both understanding and quick answers	39	49
Helps understand concepts deeply	28	35
Mostly just gives quick answers	7	9
Neither helpful	5	6

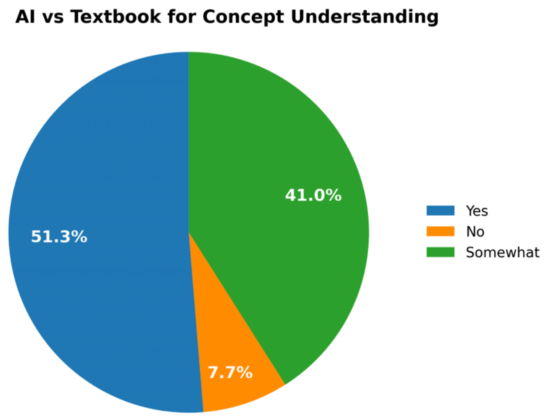

FIG. 6. Student perceptions of AI versus textbook explanations ($N = 74$ AI users). The overwhelming majority (92%) found AI explanations equally or more helpful than textbooks for physics concepts, with only 8% preferring textbook explanations exclusively.

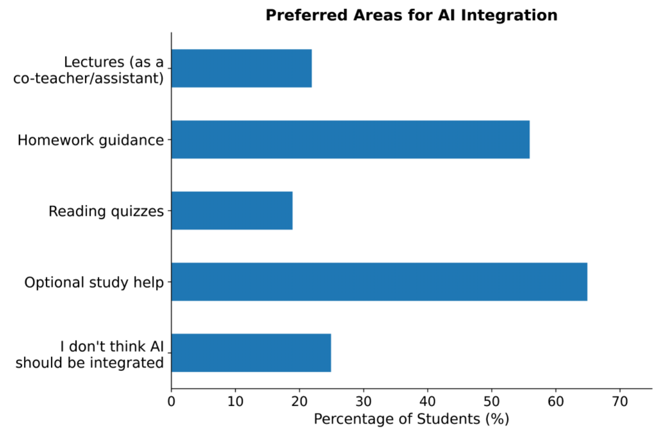

FIG. 7. Student preferences for AI integration in different physics course components ($N = 81$, multiple selections allowed). Optional study help (65%; 95% CI: [54%, 75%]) and homework guidance (56%; 95% CI: [45%, 66%]) received strong support, while lecture integration (22%; 95% CI: [14%, 32%]) faced substantial resistance. The large gap between these estimates, alongside their respective confidence intervals, indicates a clear student preference for optional over mandatory integration.

with low AI trust rather than with academic performance level: five were A/B students and one was a C/D student, while trust was distributed as low ($n = 3$), neutral ($n = 2$), and high ($n = 1$). Physics interest was roughly even across levels. The very small cell size ($n = 6$) precludes inferential analysis; however, the pattern suggests that textbook preference reflects principled skepticism about AI accuracy rather than demographic factors.

Regarding peer tutoring—itself a physics education intervention with established PER efficacy [9]—utilization was relatively low: 59% of respondents did not use peer tutoring during the course. Among the 41% who did use it, 35% found AI and peer tutoring equally helpful, 24% preferred AI, and 20% preferred peer tutoring. Lower peer tutoring utilization most likely reflects accessibility constraints (limited hours, scheduling, and location)—highlighting AI's advantage in availability, not pedagogical quality [9,10].

Cross-tabulating peer-tutoring preference by trust level reveals a substantively meaningful pattern consistent with the trust-utility framework: high-trust students strongly favored AI over peer tutoring (56% preferred AI; only 4% preferred peer tutoring, $n = 27$), whereas neutral-trust students showed the inverse pattern (11% preferred AI versus 41% preferred peer tutoring, $n = 27$). By grade, A/B students were more likely than C/D students to prefer AI over peer tutoring (34% versus 0%), though cell sizes again preclude inferential conclusions. These patterns suggest that trust in AI is a meaningful predictor of preference for AI over human tutoring and that the complementary roles of AI and peer tutoring in physics are mediated by students' domain-calibrated trust judgments.

F. Integration preferences and autonomy

Student preferences strongly favored optional over mandatory integration (Fig. 7). Support was highest for optional study help (65%, $n = 53$; 95% CI: [54%, 75%]) and homework guidance (56%, $n = 45$; 95% CI: [45%, 66%]), but only 22% ($n = 18$; 95% CI: [14%, 32%]) supported lecture integration. Notably, 25% ($n = 20$) selected "I don't think AI should be integrated," with 85% of these students ($n = 17$) choosing this as their sole response—a strong signal of principled resistance to institutional mandates.

Opposition rates showed counterintuitive patterns by student characteristics (Fig. 8, Table V): higher-performing physics students (A/B grades) showed greater opposition (26.5%) than lower-performing students

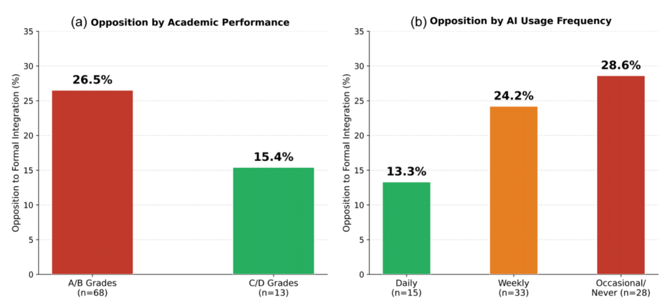

FIG. 8. Opposition to formal AI integration by student characteristics. (a) Higher-performing physics students (A/B grades) show greater opposition than lower-performing students (26.5% versus 15.4%), consistent with SDT predictions that students with established learning strategies are most sensitive to autonomy threats; this comparison is descriptive given the small C/D cell ($n = 13$). (b) Opposition decreases with AI usage frequency, suggesting familiarity reduces resistance to formal integration.

TABLE V. Opposition to formal AI integration by student characteristics. Cells marked † are underpowered for chi-square inference.

Characteristic	Total n	Oppose n	Oppose %
Overall	81	20	24.7
Self-reported grade range			
A/B grades	68	18	26.5
C/D grades†	13	2	15.4
Usage frequency			
Daily	15	2	13.3
Weekly	33	8	24.2
Occasional/never	28	8	28.6

(15.4%), while daily AI users showed the lowest opposition (13.3%) compared to occasional or nonusers (28.6%). These grade-based differences are observed patterns; given the small C/D cell ($n = 13$), confirmatory interpretation is not warranted.

Chi-square analysis revealed a meaningful association between usage frequency and integration preferences ($\chi^2 = 12.84$, $df = 4$, Cramér's $V = 0.28$, small-to-medium effect; 80% power at $N = 81$) (Table VI): daily users showed strongest support for optional integration (80%) but caution about lectures (33%); infrequent users showed highest opposition (43%).

Table VII examining integration preferences by trust level reveals a pattern that directly tests the trust-utility framework: opposition to formal integration increases monotonically as trust decreases. Among high-trust students ($n = 33$), 0% opposed formal integration and 90.9% supported optional study help. Among neutral-trust students ($n = 33$), 36.4% opposed formal integration and 48.5% supported optional help. Among low-trust students ($n = 15$), 53.3% opposed formal integration and 46.7% supported optional help. This monotonic pattern—with no high-trust student opposing formal integration—provides direct empirical support for the prediction that trust calibration moderates integration preferences. The low-trust cell ($n = 15$) is underpowered for chi-square inference; nonetheless, the effect is large and theoretically coherent. These findings are further discussed in Sec. V D.

G. Qualitative themes

Thematic analysis of open-ended responses ($n = 67$) yielded eight conceptually distinct patterns (Table VIII). We present all eight themes but emphasize those of the most direct relevance to physics-specific learning and PER constructs. Descriptive examination of whether themes cluster differentially across subgroups revealed several patterns worth noting: technically precise domain critiques (theme 1) appeared most commonly in open-ended responses from daily, A/B students, consistent with the interpretation that frequent high-performing users accumulate enough failure experience to articulate specific error modes; verification language (theme 2) appeared more often among weekly and daily users than occasional users; and all 17 students expressing principled opposition (theme 8) reported neutral or low trust scores—a convergence with the quantitative trust-integration finding reported in Sec. IV F.

1. Theme 1: Physics domain-specific limitations

This theme is the most distinctively physics-specific finding of the study and the one most directly grounded in PER constructs. Qualitative respondents spontaneously identified *where in physics* AI fails, with 40% doing so through a calibrated, domain-specific critique that maps directly onto PER's established taxonomy of student difficulties with multirepresentational reasoning [2].

Visual reasoning was a consistent weakness. One student observed: "It cannot significantly see the graph pictures clearly so it makes things up and they almost all the time come out wrong answer unless I write out what the graph looks like." Another noted: "The biggest limitation is with diagrams, since AI often gets confused or does not fully understand drawings or graphs." These responses map directly onto PER-identified competencies where physics students already struggle—reading and interpreting kinematic graphs, force diagrams, and field line representations [2], the same representational domains where AI performance is documented to be weakest [6,11].

Circuit analysis emerged as a particularly failure-prone domain: "For circuit analysis, AI bots almost always make mistakes." and "Sometimes they switch values around (i.e., they use all of the same numbers as the problem, but they switch them around)." The latter describes a specific error pattern—numerical transposition—that would be difficult to detect without independent calculation, precisely the verification behavior that the trust-utility gap predicts students may not consistently perform.

Students also identified limitations in abstract physical reasoning: "Good with computations if given numbers and equations but not abstract understanding of applications. Doesn't always perform derivations competently. Always have to double-check anything it does."

This pattern of domain-calibrated skepticism—distinguishing AI competence in structured quantitative tasks from its failures in visual-spatial and abstract-physical reasoning—represents exactly the physics-specific AI literacy that the framework of Long *et al.* [30] identifies as essential and maps closely onto the multi-representational challenges PER has long identified as central to physics learning [1,2].

2. Theme 2: Strategic verification and critical evaluation

Respondents described verification strategies reflecting metacognitive awareness of AI limitations in physics, with

TABLE VI. AI integration preferences by usage frequency ($N = 81$, multiple selections allowed).

Usage frequency	Support optional ($n/\%$)	Support lectures ($n/\%$)	Oppose any ($n/\%$)
Daily ($n = 15$)	12/80	5/33	1/7
Weekly ($n = 33$)	24/73	8/24	5/15
Occasional/never ($n = 28$)	17/61	5/18	12/43

TABLE VII. AI integration preferences by trust level ($N = 81$, multiple selections allowed). The low-trust cell ($n = 15$) is underpowered for chi-square inference; patterns are descriptive.

Trust level	n	Oppose any n (%)	Optional n (%)	Lectures n (%)
High	33	0 (0.0)	30 (90.9)	10 (30.3)
Neutral	33	12 (36.4)	16 (48.5)	6 (18.2)
Low†	15	8 (53.3)	7 (46.7)	2 (13.3)

TABLE VIII. Prevalence of qualitative themes among respondents providing open-ended responses ($N = 67$). Respondents could exhibit multiple themes. Interrater reliability: Cohen's $\kappa = 0.78$ (substantial agreement).

Theme	n	%
Strategic verification and critical evaluation	31	46
Scaffolding versus direct answers	28	42
Accessibility and time constraints	35	52
Personalized and adaptive explanations	19	28
Support for exam preparation	41	61
Physics domain-specific limitations	27	40
AI as supplement not replacement	44	66
Principled opposition to formal integration	17	25

46% providing such responses. One student explained: "I use AI to help guide me through topics since I struggle with how they all work together and how to solve for questions, but I do the actual solving on my own, and check the final answer with AI." Another cross-referenced with course notes, recognizing that AI "sometimes does things differently or wrong."

Some students compared multiple AI platforms for physics problems: "I use multiple chatbots to see if it's accurate. For circuit analysis, AI bots almost always make mistakes."

Critically, these are entirely self-reported strategies. The verification gap documented by the work of Kasneci *et al.* [31] suggests that actual verification rates in physics—particularly under the time pressure of pre-exam problem sets—may be substantially lower than these self-reports suggest (see Sec. V C). This represents the most important unresolved limitation of all self-report studies in this domain, including the present work.

3. Theme 3: Scaffolding rather than direct answers

Students (42%) distinguished using AI as scaffolding for physics understanding versus extracting direct solutions—a distinction central to PER's concern about surface versus deep learning [3]. One student explained: "My preferred method of using AI is to explain what a question is asking me to do when I don't understand what a question is asking." Another used AI to understand the physical reasoning behind formulas: "Alternatives when time is restricting one to get someone to explain the how's and whys behind developing a formula."

4. Theme 4: Accessibility and physics study timing

AI's constant availability was highlighted by 52% of respondents, particularly when traditional physics support structures are absent: "There are many times in which I am doing the homework by myself and don't know where to even start a problem. I could spend half an hour scanning through a textbook for a concept that might be there, or I can ask ChatGPT and get a good explanation immediately." This theme points to a structural gap in physics support: office hours, tutoring centers, and peer study groups operate on fixed schedules that poorly match students' actual late-night physics study patterns, a mismatch that AI's availability advantage addresses regardless of pedagogical quality.

5. Theme 5: Personalized and adaptive explanations

Students used AI to rephrase or tailor physics explanations to individual learning preferences, with 28% of respondents reporting this use, reflecting an advantage over static texts in providing personalized conceptual scaffolding—though whether this personalization produces superior conceptual outcomes remains untested.

6. Theme 6: Support for exam preparation

The most common quantitative use case (76%) was also prominent qualitatively (61%): students used AI to generate practice problems and diagnose conceptual gaps before high-stakes assessments. This application aligns with formative assessment principles well established in PER [20].

7. Theme 7: AI as supplement rather than replacement

The most prevalent theme (66%) was students explicitly framing AI as supplementary to traditional physics instruction. Even daily users emphasized maintaining personal problem-solving effort and independent reasoning—consistent with physics education norms that deep engagement is intrinsic to developing expertise [1,3]. One student captured the boundary clearly: "I never let AI solve the problem for me. I use it to check my understanding of a concept, and then I close it and work through the problem myself. If I just copy what it gives me I will not be able to do it on the exam."

8. Theme 8: Principled opposition to formal integration

Respondents articulated opposition to formal course integration on principled educational grounds, with 25% emphasizing the preservation of instructor-centered pedagogy, resistance to normalization of AI in academic contexts, and protection of self-regulated learning strategies. Our survey did not ask about other grounds for opposition (e.g., academic integrity, environmental costs, or workforce displacement), and we do not interpret results as bearing on those debates. Importantly, many who opposed formal integration still valued AI as an optional resource—exemplifying SDT's autonomy principle [15,16]: "AI is a good tool, but I don't think that it needs to be integrated into the course. It is, first and foremost, an optional study tool to be used to deepen a student's understanding outside of class."

V. DISCUSSION

A. Contributions to the emerging AI-PER literature

This study contributes to a rapidly developing empirical literature on AI in physics education. Kortemeyer [13] demonstrated that an AI agent could navigate structured physics calculations but failed on tasks requiring conceptual judgment and graphical reasoning. Our student-side data mirrors this AI-side finding: students trust AI precisely where Kortemeyer found it competent, and distrust it where Kortemeyer found it failing. This convergence between AI performance benchmarks and student trust calibration is a new result that would not emerge from either study alone.

Kestin *et al.* [12] showed that AI tutoring can match active-learning outcomes under controlled conditions. Our data situate this result: in authentic self-directed study, students use AI with only moderate trust and self-reported (but unvalidated) verification habits—a context very different from that of a structured tutoring trial.

B. Interpreting the trust-utility gap in physics

The central finding—91% usage but only 41% trust (nonoverlapping 95% CIs)—is best understood as domain-calibrated pragmatism specific to physics and grounded in the discipline's multirepresentational demands [2]. Students use AI for tasks where experience tells them it performs adequately (e.g., structured quantitative problem solving, conceptual explanations, and practice problem generation) while withholding trust for tasks where they have encountered specific failures (e.g., circuit analysis, diagram interpretation, and derivations). This interpretation is supported by the qualitative finding that 40% of respondents spontaneously identified concrete, physics-specific AI failure modes that map directly onto the representational hierarchies that PER has documented as most challenging [1,2]. Alternative explanations—pragmatic availability trade-offs or failures of self-regulation under time pressure [46,47]—require observational studies to distinguish, which stands as the most critical research priority from this study.

C. The verification gap: A critical priority for physics education research

The *verification gap*—the divergence between students' self-reported AI checking behaviors and their actual practices—is the most critical unresolved question for physics education and the most significant limitation of the present self-report study [31,32]. This gap carries particularly acute implications in physics:

1. **Physics misconceptions may crystallize undetected**. AI errors that perpetuate common misconceptions—such as conflating velocity with acceleration, misapplying Newton's third law, incorrectly applying conservation laws—may look plausible to students who lack the conceptual foundation to detect them. PER has shown such misconceptions are highly resistant to subsequent correction [1,2], meaning that unverified AI errors during introductory courses could compound across sequential physics courses.
2. **Quantitative errors in physics are often invisible without verification.** A wrong sign in an electromagnetic field calculation, an incorrect limiting case, or a misapplied boundary condition may produce a numerically reasonable-looking answer. Physics-specific verification habits—dimensional analysis, limiting cases, and energy and momentum

conservation cross-checks—are essential but require deliberate instruction and are unlikely to develop spontaneously [1].

3. **Assessment redesign becomes essential.** If take-home physics assignments involve substantial unverified AI use, proctored assessments of independent problem solving become the primary valid measurement of physics understanding, and assessment reform must prioritize tasks requiring physical argument beyond current AI capabilities.
4. **AI literacy instruction must target habits, not just awareness.** Students may know that circuit-analysis outputs require verification without applying those verification habits under time pressure [32]. Explicit training in physics-specific verification strategies is necessary to bridge the knowing-doing gap that the verification gap represents.

Future research should log real student-AI physics interactions, code actual verification behaviors, and compare observed versus self-reported rates across task types and time-pressure conditions.

D. Autonomy and integration in physics courses

The strong preference for optional over mandatory integration (65%; 95% CI: [54%, 75%] versus 22%; 95% CI: [14%, 32%]) directly supports predictions derived from SDT [15,16]. The descriptive finding that higher-performing physics students oppose mandatory integration more (26.5%) than lower-performing students (15.4%) is consistent with SDT: students who have successfully developed self-regulated physics learning strategies—working through derivations, consulting multiple textbook treatments, attending office hours—are particularly sensitive to perceived threats to their autonomy over those strategies [15,17].

The trust-stratified integration analysis reported in Sec. IV F provides additional mechanistic support for the SDT account: opposition to formal integration is concentrated entirely among neutral- and low-trust students (36.4% and 53.3% opposition, respectively), while no high-trust student opposed formal integration. This convergence between trust calibration and autonomy preference suggests that resistance to mandatory integration is not a blanket rejection of AI but rather a domain-informed judgment that compulsory adoption is inappropriate given current AI limitations—a response that is more epistemically sophisticated than simple technology resistance.

Conversely, lower opposition among struggling students may reflect greater receptivity to structured support or greater deference to institutional guidance—consistent with the "trust-performance paradox" [48]. This cautions against interpreting lower opposition as enthusiasm for mandatory integration; it may reflect uncritical deference that could harm vulnerable students if AI is mandated without adequate verification instruction.

E. AI as complementary physics learning resource

Students' comparative evaluations reveal complementary rather than competitive roles for AI and traditional physics support systems. AI's primary advantages lie in its 24/7 availability, scalability, and capacity for step-by-step algebraic explanation. In contrast, peer tutoring and office hours provide physics-specific pedagogical expertise, diagram-based explanation at a whiteboard, and the collaborative problem-solving dynamics that PER has shown support deeper conceptual engagement [9,10].

A related open question is whether a gap exists between students' *self-identified* learning goals and those for which they actually seek AI assistance. Theme 7 suggests that many students aspire to use AI for understanding rather than answers, but whether higher-order goals are genuinely prioritized when AI provides an immediately satisfying answer is an open question best addressed through behavioral studies using conversation logs and think-aloud protocols [50].

F. Equity considerations for physics AI integration

Access barriers to physics AI use extend beyond device access to include direct financial costs, as many capable AI models remain behind premium subscription paywalls. More critically, the trust-performance paradox [48] carries particular severity in physics: students who cannot independently detect AI errors in force diagrams or circuit analysis risk internalizing misconceptions that compound across sequential courses. Because physics curricula are highly sequential, foundational errors during kinematics or dynamics will propagate into electromagnetism and beyond—precisely where lower-performing students, who most need scaffolding, are least equipped to catch them. Equity-minded integration must therefore ensure that AI access is accompanied by explicit instruction in physics-specific verification, not merely tool availability.

G. Implications for physics instruction

Our findings generate several evidence-based recommendations for instructors, departments, and institutions navigating the AI transition (Table IX).

H. Limitations and generalizability

Several limitations constrain interpretation. First, the sample is restricted to a STEM-focused technical university; findings may not transfer to comprehensive universities, liberal arts colleges, or community colleges where student populations and physics course structures differ substantially.

TABLE IX. Evidence-based recommendations for AI integration in physics education, derived from study findings.

Stakeholder	Recommendation	Empirical rationale
Instructors	Acknowledge AI as an existing resource	91% adoption rate indicates prohibition is ineffective.
	Provide physics-specific AI literacy	40% of students identify domain failures (circuits and diagrams).
	Redesign out-of-class assessments	92% find AI helpful for homework; proctored exams are needed.
	Offer AI guidance as an optional resource	65% prefer optional support, preserving student autonomy.
Department/curriculum	Develop tiered integration models	Usage and opposition gradients suggest differentiated needs.
	Formalize verification habits	Verification requires instruction in dimensional analysis.
	Maintain peer tutoring alongside AI	Tutors provide physical intuition; AI provides procedural steps.
Institutional policy	Avoid mandatory AI integration	25% opposition; mandatory use risks undermining motivation.
	Ensure equitable AI access	Trust-performance paradox places vulnerable students at risk.
	Invest in observational research	Self-reported verification likely overstates actual behavior.

Second, the 18.12% response rate introduces potential self-selection bias. Using institutional data ($N = 447$), our sample overrepresented high-performing students by 13.9 percentage points (83.0% self-reported A/B versus 69.1% institutional A/B). In addition, repeated reminders in a subset of sections may have created uneven awareness of the survey. These patterns suggest possible nonresponse-related bias; however, we did not model nonresponse quantitatively and therefore cannot estimate the magnitude or direction of any such bias. Replication with higher response rates is needed to assess generalizability.

Third, all data are cross-sectional and self-reported; causal claims are not warranted. Social desirability bias and metacognitive overestimation [46,49] likely inflate reported verification rigor. The well-documented gap between how students report checking their work and actual behavior under time pressure [32] is the primary threat to all self-report findings here. Screen-capture studies, think-aloud protocols [50], and conversation log analysis are needed to validate whether the scaffolding orientation students report translates into practice.

Fourth, the survey used a single-item trust measurement rather than a validated multi-item scale [38]. Fifth through eighth: data collection in weeks 10–13 precludes end-of-semester correlations; AI platform was not recorded (a deliberate design choice given our focus on usage patterns rather than platform comparisons, though future work should address this); a larger coding team would strengthen qualitative credibility; and the C/D ($n = 13$) and algebra-based ($n = 13$) subgroups are underpowered (∼25% for medium effects) for all subgroup comparisons.

VI. FUTURE DIRECTIONS

The most critical priority is the observational validation of verification behaviors: log analysis and think-aloud protocols [50] examining verification rates by physics task type, time pressure, and performance level. Randomized controlled trials using the Force Concept Inventory [20] and Conceptual Survey of Electricity and Magnetism could establish causal effects of integration strategies on conceptual outcomes. Psychometrically validated instruments for trust, AI literacy, and verification behavior in physics contexts are needed, alongside longitudinal tracking to assess whether skepticism and error-detection accuracy increase with content knowledge. Finally, cross-disciplinary and cross-institutional comparisons would clarify which behaviors are physics-specific versus broadly characteristic of STEM AI use.

VII. SUMMARY AND CONCLUSION

This study examined how undergraduate physics students at a STEM-focused university use, trust, and prefer to integrate AI tools ($N = 81$). Among respondents, 91% (95% CI: [83%, 96%]) used AI for physics coursework—with 59.2% using it daily or weekly—yet only 41% (95% CI: [30%, 52%]) trusted AI-generated physics explanations. This 50-percentage-point *trust-utility gap* reflects domain-calibrated pragmatism grounded in students' direct experience with AI failures in multirepresentational physics tasks.

Physics-specific findings distinguish this study from general AI adoption surveys: 40% of qualitative respondents spontaneously identified physics-domain AI limitations (circuit analysis, diagram and graph interpretation, and symbolic derivations) that map directly onto PER's established taxonomy of multirepresentational reasoning difficulties [2]. Furthermore, students consistently framed their AI use as scaffolding for building physical intuition rather than merely extracting answers, a mindset consistent with constructivist accounts of active physics learning [3].

Students strongly preferred optional (65%; 95% CI: [54%, 75%]) over mandatory (22%; 95% CI: [14%, 32%]) AI integration, a divergence consistent with the autonomy predictions of self-determination theory [15]. The descriptive pattern that higher-performing physics students showed greater opposition to formal integration (26.5%) than lower-performing peers (15.4%) suggests elevated autonomy sensitivity among students who already possess established physics learning strategies. As noted, this subgroup comparison is descriptive pending replication in larger samples.

The most critical limitation of this work—and the most urgent future research direction for the field—is the reliance on self-reported verification data. Because the educational psychology literature demonstrates that students frequently overestimate their metacognitive checking behaviors under time pressure [31,32], actual physics verification rates are likely substantially lower than our self-reported data indicate. Observational validation of physics-specific verification behaviors is therefore the most pressing research priority identified by this study.

Ultimately, physics education can benefit from acknowledging widespread AI use while building physics-specific verification habits into formal instruction. Training students in dimensional analysis, limiting cases, and conservation-law cross-checks [1] will prepare them for professional contexts where AI is standard—while preserving the authentic reasoning and human expertise that define physics education.

ACKNOWLEDGMENTS

We thank the students who participated in this study and the physics faculty at Florida Polytechnic University for their support.

E. F.: Conceptualization, Methodology, Investigation, Formal analysis, Writing—original draft, Writing—review and editing. I. B.: Conceptualization, Formal analysis, Writing—editing.

The authors declare no competing financial interests.

DATA AVAILABILITY

The survey instrument and de-identified response data that support the findings of this article are publicly available on ResearchGate: Survey instrument [34]. De-identified survey response data underlying the findings of this study are publicly available at Ref. [34].

[1] E. F. Redish, *Teaching Physics with the Physics Suite* (Wiley, Hoboken, NJ, 2003).

[2] L. C. McDermott and E. F. Redish, Resource letter PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

[3] M. T. H. Chi, Active-constructive-interactive: A conceptual framework for differentiating learning activities, Top. Cogn. Sci. **1**, 73 (2009).

[4] C. K. Lo, What is the impact of ChatGPT on education? A rapid review of the literature, Educ. Sci. **13**, 410 (2023).

[5] Microsoft Corporation, *2025 AI in Education Report: Trends and Insights* (Microsoft Corporation, Redmond, 2025).

[6] K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving, Front. Educ. **8**, 1330486 (2024).

[7] I. Bentley and M. Gebran, *Neural Networks in the Physical Sciences* (American Chemical Society, Washington, DC, 2025).

[8] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, *Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. Mestre (Information Age Publishing, Greenwich, CT, 2005), p. 120.

[9] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[10] S. Freeman *et al.*, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A **111**, 8410 (2014).

[11] D. Tong, B. Jin, Y. Tao, H. Ren, A Y M. Atiquil Islam, and L. Bao, Exploring the role of human-AI collaboration in solving scientific problems, Phys. Rev. Phys. Educ. Res. **21**, 010149 (2025).

[12] G. Kestin, K. Miller, A. Klales, T. Milbourne, and G. Ponti, AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting, Sci. Rep. **15**, 17458 (2025).

[13] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).

[14] A. Sirnoorkar, D. Zollman, J. T. Laverty, A. J. Magana, N. S. Rebello, and L. A. Bryan, Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning, Comput. Educ. Artif. Intell. **7**, 100318 (2024).
[15] R. M. Ryan and E. L. Deci, Intrinsic and extrinsic motivations: Classic definitions and new directions, Contemp. Educ. Psychol. **25**, 54 (2000).
[16] E. L. Deci, R. M. Ryan, and F. Guay, *Theory Driving Research: New Wave Perspectives on Self-Processes and Human Development*, edited by D. M. McInerney and E. Talt (Information Age Publishing, Charlotte, 2013), p. 133.
[17] F. Guay, Applying self-determination theory to education: Regulation types, psychological needs, and autonomy supporting behaviors, Can. J. Sch. Psychol. **37**, 75 (2022).
[18] J. S. Bureau, J. L. Howard, J. X. Y. Chong, and Fédéric Guay, Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations, Rev. Educ. Res. **92**, 46 (2022).
[19] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, User acceptance of information technology: Toward a unified view, MIS Q. **27**, 425 (2003).
[20] R. R. Hake, Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[21] L. Ding, T. Li, S. Jiang, and A. Gapud, Students' perceptions of using ChatGPT in a physics class as a virtual tutor, Int. J. Educ. Technol. High. Educ. **20**, 63 (2023).
[22] B. G. Acosta-Enriquez, E. V.ónica Ramos Farroñan, L. I. Villena Zapata, F. S. Mogollon Garcia, H. C. Rabanal-León, J. E. M. Angaspilco, and J. C. S. Bocanegra, Acceptance of artificial intelligence in university contexts: A conceptual analysis based on UTAUT2 theory, Heliyon **10**, e38315 (2024).
[23] Y. Liang, D. Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, Smart Learn. Environ. **10**, 52 (2023).
[24] B. S. Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, Educ. Res. **13**, 4 (1984).
[25] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, Intelligent tutoring systems and learning outcomes: A meta-analysis, J. Educ. Psychol. **106**, 901 (2014).
[26] J. A. Kulik and J. D. Fletcher, Effectiveness of intelligent tutoring systems: A meta-analytic review, Rev. Educ. Res. **86**, 42 (2016).
[27] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, Conceptualizing AI literacy: An exploratory review, Comput. Educ. Artif. Intell. **2**, 100041 (2021).
[28] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, AI literacy: Definition, teaching, evaluation and ethical issues, Proc. Assoc. Inf. Sci. Technol. **58**, 504 (2021).
[29] Y. Dong, W. Xu, J. Huang, and K. Yann, Validating and refining a multi-dimensional scale for measuring AI literacy in education using the Rasch Model, Humanit. Soc. Sci. Commun. **12**, 1317 (2025).
[30] D. Long and B. Magerko, What is AI literacy? Competencies and design considerations, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), Vol. 1.
[31] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, G. Kasneci, T. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. **103**, 102274 (2023).
[32] J. Dunlosky and K. A. Rawson, Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention, Learn. Instr. **22**, 271 (2012).
[33] P. Bitzenbauer, ChatGPT in physics education: A pilot study on easy-to-implement activities, Contemp. Educ. Technol. **15**, ep430 (2023).
[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/bdl1-6bn4 for the survey instrument and de-identified response data.
[35] F. J. Fowler Jr., *Survey Research Methods*, 5th ed. (Sage Publications, Thousand Oaks, 2014).
[36] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. (Sage Publications, Thousand Oaks, 2018).
[37] K. Gogol, M. Brunner, T. Goetz, R. Martin, S. Ugen, U. Keller, A. Fischbach, and F. Preckel, “My Questionnaire is Too Long!” The assessments of motivational-affective constructs with three-item and single-item measures, Contemp. Educ. Psychol. **39**, 188 (2014).
[38] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, Int. J. Cogn. Ergon. **4**, 53 (2000).
[39] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, 1988).
[40] R. L. Wasserstein and N. A. Lazar, The ASA’s statement on *p*-values: Context, process, and purpose, Am. Stat. **70**, 129 (2016).
[41] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behav. Res. Methods **39**, 175 (2007).
[42] V. Braun and V. Clarke, Using thematic analysis in psychology, Qual. Res. Psychol. **3**, 77 (2006).
[43] M. Lungu, Book review: The coding manual for qualitative researchers, Am. J. Qual. Res. **6**, 232 (2022).
[44] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).
[45] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).
[46] G. Schraw and R. S. Dennison, Assessing metacognitive awareness, Contemp. Educ. Psychol. **19**, 460 (1994).

[47] B. J. Zimmerman, *Handbook of Self-Regulation*, edited by M. Boekaerts, P. R. Pintrich, and M. Zeidner (Academic Press, San Diego, 2000), p. 39.

[48] F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, and I. Shklovski, in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (ACM, New York, 2024), pp. 1–18.

[49] J. H. Flavell, Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry, Am. Psychol. **34**, 906 (1979).

[50] K. R. Koedinger, A. T. Corbett, and C. Perfetti, The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning, Cognit. Sci. **36**, 757 (2012).