

Effects of generative AI-mediated socratic dialogue on preservice physics teachers' critical thinking in kinematics

Marcos Guerrero-Zambrano^{✉,*}, Bryan Valarezo-Chamba,[†] and Leonor Sanchez-Alvarado[‡]

*UNEMI Facultad de Educación, Universidad Estatal de Milagro,
Cda. Universitaria Km, 1.5 vía Km. 26, 091050, Milagro, Ecuador*

Erick Lamilla-Rubio[§]

*UNEMI Facultad de Educación, Universidad Estatal de Milagro, Cda. Universitaria Km,
1.5 vía Km. 26, 091050, Milagro, Ecuador
and ESPOL Departamento de Física, Escuela Superior Politécnica del Litoral,
Campus Gustavo Galindo, Km 30.5, 090150, Guayaquil, Ecuador*



(Received 12 February 2026; accepted 4 August 2026; published 25 September 2026)

Initial physics teacher education frequently approaches kinematics problem solving as a procedural activity, with limited engagement in conceptual reasoning, argumentation, and metacognitive regulation. This study examines the effects of Socratic dialogue mediated by generative artificial intelligence (GAI), compared with teacher-guided Socratic dialogue alone, on discipline-specific critical thinking (CT) in one-dimensional kinematics. A concurrent mixed-methods, within-group preexperimental design was implemented with 18 preservice physics teachers working in small groups across two sequential phases using parallel kinematics problem sets. Because the instructional conditions were implemented in a fixed order, the observed differences cannot be disentangled from possible practice, learning, and carryover effects. In the first stage, students engaged in instructor-guided Socratic dialogue; in the second stage, the same structure was maintained while incorporating GAI-mediated Socratic support under instructor supervision. Critical thinking was assessed through an analytic rubric applied to written group solutions and complemented by Likert-scale surveys and open-ended responses. Rubric scores showed consistent improvement during the GAI-mediated phase across all groups, with no observed declines. Students also reported higher self-efficacy and intrinsic motivation, while cognitive load exhibited heterogeneous responses associated with the need for verification and epistemic control when interacting with GAI. These results suggest that integrating generative AI into a teacher-mediated Socratic framework can support critical thinking in kinematics, although the sequential preexperimental design requires cautious interpretation.

DOI: [10.1103/x7vq-17rq](https://doi.org/10.1103/x7vq-17rq)

I. INTRODUCTION

The development of critical thinking (CT) in initial teacher training remains a persistent challenge, particularly in physics, where conceptual understanding must extend beyond the mechanical application of formulas. In much of Latin America, instruction in one-dimensional

kinematics has traditionally emphasized algorithmic procedures, thereby constraining the development of reflective, analytical, and metacognitive competencies required for interpreting physical phenomena beyond numerical computation [1,2]. This instructional pattern reveals a substantial gap between prevailing classroom practices and the epistemic demands of physics teacher preparation, where reasoning, model-based interpretation, and justification are central to disciplinary expertise.

International research has identified guided inquiry, collaborative argumentation, and conceptual analysis as relevant instructional approaches for supporting critical thinking in science education. [3,4]. In physics education, kinematics occupies a particularly demanding epistemic space, as it requires learners to construct mental models, interpret variable relationships, and engage in abstract reasoning—capacities that are central to the professional formation of physics teachers [5].

*Contact author: mguerreroz@unemi.edu.ec

†Contact author: bvalarezoc@unemi.edu.ec

‡Contact author: lsancheza4@unemi.edu.ec

§Contact author: elamill@espol.edu.ec, elamillar2@unemi.edu.ec

Published by the American Physical Society under the terms of the *Creative Commons Attribution 4.0 International license*. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Recent work further suggests that Socratic dialogue can be productively extended through the mediation of generative artificial intelligence (GAI), enabling richer metacognitive engagement and more explicit, evidence-based argumentation [6]. Such mediation opens new possibilities for personalized feedback, reflective dialogue, and guided reconstruction of conceptual errors. While emerging studies indicate that both human tutors and GAI-based systems can support critical thinking, empirical evidence within physics education remains limited. This gap is especially visible in domains such as one-dimensional kinematics, where reasoning about motion, representation, and causality is foundational yet persistently challenging [7,8].

From an epistemic standpoint, GAI can be understood not merely as an instructional tool but as a form of mediation that shapes how learners generate evidence, justify inferences, and regulate their own reasoning. When embedded within Socratic dialogue, GAI can function as an epistemic scaffold by prompting alternative explanations, counterexamples, and multiple solution pathways, thereby supporting metacognitive monitoring and sense-making. However, this form of mediation requires careful pedagogical design to ensure that epistemic agency remains with the learner rather than being transferred to the technology itself [9]. Recent scholarship has emphasized the importance of preserving responsibility, judgment, and reflective control when AI systems are used in educational settings [10–12]. Accordingly, the present study conceptualizes GAI as a supervised epistemic scaffold intended to support critical thinking and metacognitive regulation, rather than as a replacement for disciplinary understanding [13].

The preceding discussion situates this study within a view of physics problem solving that moves beyond procedural execution toward reasoning, interpretation, and argumentative sensemaking. Rather than treating kinematics as an exercise in formula substitution, this perspective emphasizes the examination of assumptions, the coordination of representations, and the use of dialogue to support the development of critical thinking. Within this framework, Socratic dialogue, when supported by generative artificial intelligence, offers a means of extending students' inferential and metacognitive activity by encouraging alternative explanations, conceptual contrasts, and reflective self-correction.

Guided by this perspective, the present study compares teacher-guided Socratic dialogue with a subsequent condition in which the same dialogical structure is supported by generative artificial intelligence. The study therefore asks what differences in discipline-specific critical-thinking performance are observed between these two instructional conditions; how participants' perceptions of kinematics self-efficacy (KSE), cognitive load (CL), and intrinsic motivation (IM) differ across the two phases; and how their qualitative reflections converge with or qualify

the quantitative findings. Because the two conditions were implemented sequentially in a fixed order, these questions concern within-sample differences and patterns of convergence rather than the independent causal effect of GAI mediation. Consequently, this study addresses the following research question: what differences emerge in rubric-assessed critical-thinking performance and qualitative reasoning when preservice physics teachers solve one-dimensional kinematics problems through teacher-guided Socratic dialogue alone, compared with the same dialogue structure supported by GAI?

The remainder of the paper is organized as follows. Section II presents the mixed-methods preexperimental design, participants, SOCRATIC-IA instructional framework, instruments, and data-analysis procedures. Section III reports the quantitative and qualitative findings concerning critical-thinking performance and participant perceptions under the two instructional conditions. Section IV discusses the findings in relation to previous research in physics education and AI-supported learning, and Sec. V presents the conclusions.

II. METHODOLOGY

A. Design study

The present study was conducted using a mixed-methods approach with a qualitative emphasis and an interpretive orientation, aimed at examining how critical thinking develops in initial physics teachers as they engage in Socratic dialogue while solving conceptual and quantitative problems in one-dimensional kinematics under two mediation conditions: (1) teacher-guided dialogue without GAI and (2) teacher-guided dialogue supported by GAI. This qualitative focus was complemented by a descriptive and inferential quantitative component that estimated the magnitude and direction of changes in performance and learning perceptions across conditions, consistent with concurrent mixed-methods integration designs described in the methodological literature [14,15].

An applied empirical design was used to examine a researcher-developed instructional intervention. The analysis focuses on both student performance and learning perceptions associated with a Socratic mediation strategy implemented either alone or in combination with generative artificial intelligence within an initial physics teacher training context. The research adopts a comparative and explanatory orientation with a practical focus, as it contrasts two instructional conditions within the same participant group and seeks to interpret observed changes using qualitative and quantitative evidence, without asserting the strong causal claims associated with randomized or fully controlled experimental designs. The study employed a within-group preexperimental repeated-measures design with two sequential stages using isomorphic tasks. In the first stage, participants solved kinematics problems

through teacher-guided Socratic dialogue. In the second stage, the same Socratic structure was preserved while GAI was introduced as an additional mediating resource. This design held the disciplinary content and procedural sequence constant, allowing changes in performance to be examined across two forms of mediation. However, because the phases followed a fixed order, potential practice and carryover effects are recognized as threats to internal validity. Consequently, differences between stages are interpreted as evidence of association compatible with the introduction of GAI rather than as definitive causal effects, consistent with classical analyses of causal inference limitations in uncontrolled designs.

Operationally, the independent variable was the form of Socratic mediation, contrasting teacher-guided dialogue without GAI with teacher-guided dialogue supported by GAI. The dependent variable was the level of critical thinking demonstrated in the solution of one-dimensional kinematics problems, operationalized through a single analytic rubric designed and applied consistently across both phases. The rubric assessed problem and principle comprehension, logical reasoning and argumentative structure [16,17], counterexamples [18], recognition and correction of conceptual errors [16,19], and clarity and organization of the written response [18]. The quantitative component analyzed score variations between stages using descriptive statistics and inferential tests for related samples, selected according to the distributional properties of the difference scores and accompanied by effect size estimates, as recommended for within-subject designs [20,21].

As a complementary measure, a perception survey was administered at the end of each phase. In the teacher-guided condition, the survey assessed kinematics self-efficacy, cognitive load, intrinsic motivation, Socratic interaction (SI), and perceived critical thinking, together with an open-ended item designed to elicit qualitative evidence about learning derived from the dialogue. In the teacher-GAI condition, the survey measured kinematics self-efficacy, cognitive load, and intrinsic motivation, along with perceived usefulness of GAI (PU-GAI), trust in educational AI (TE-AI), and resource integration (RI), and included an open-ended item focused on perceived usefulness, clarity, and potential sources of confusion associated with AI-supported problem solving. Because not all dimensions were common across stages, inferential comparisons between conditions were restricted to the shared constructs, whereas phase-specific dimensions were reported descriptively and interpreted qualitatively, in accordance with principles of measurement comparability and the cautious treatment of repeated-measures scales [20,22].

Finally, the qualitative component examined the nature of the reasoning displayed by participants through their open-ended responses and, when applicable, through

evidence drawn from dialogue and written problem solutions. This analysis was aimed at identifying patterns of justification, error monitoring, use of analogies or counterexamples, and argumentative organization. The qualitative findings were then integrated with the quantitative results through a convergent mixed-methods strategy, in which numerical evidence indicates the direction and magnitude of change, while qualitative evidence elucidates the form and conditions under which those changes occurred, in line with established principles for mixed-methods integration [14].

Given the small number of participants and the use of six collaborative groups as the unit of analysis for rubric-based performance, the quantitative component was treated as exploratory and complementary. Its purpose was not to generate population-level estimates or establish causal effects but to describe the direction, consistency, and approximate magnitude of within-sample changes across the two instructional conditions. Inferential tests were therefore interpreted cautiously and were integrated with qualitative evidence through the mixed-methods convergence analysis.

Because the process produced a single consensus score for each product instead of independent ratings from each evaluator, an intraclass correlation coefficient (ICC) could not be calculated. Regarding the content validity index (CVI), we clarified that the final CVI = 1.00 was obtained only after an iterative two-round expert review. In the first round, the analytic rubric obtained a global CVI of 0.944, and the problem-Socratic question set obtained a global CVI of 0.886. After revisions, the second validation round produced a global CVI of 1.00.

B. Participants and teacher education context

The study was conducted within an initial physics teacher training program at a public university in Ecuador. The potential population consisted of 97 students who had previously completed coursework in one-dimensional kinematics. Participation was based on voluntary enrollment and instructional relevance, yielding a group that was not merely selected but represented an appropriate community for engaging in Socratic dialogue and examining the development of critical thinking in an authentic teacher education setting. A total of 18 students completed all in-person sessions of the intervention, corresponding to a participation rate of 18.6%. Participants were enrolled between the third and eighth semesters of their program and ranged in age from 18 to 25 years. To support collaborative work and enable comparative analysis, students were organized into six groups of three. Group-level performance served as the unit of analysis for critical-thinking scores, while individual contributions were preserved through the dialogic and collaborative writing processes.

Participants had previous classroom experience using generative AI tools. Their prior use was guided by the instructor through structured prompts designed for academic activities. In previous classes, students had been instructed to formulate prompts using four components: role, context, instruction, and expected output format. For the present study, this structure was reviewed before the GAI-mediated phase and applied specifically to the solution of one-dimensional kinematics problems.

C. The SOCRATIC-IA socratic learning model

The instructional intervention was implemented through the SOCRATIC-IA Socratic learning model, a framework developed by the authors to promote discipline-specific critical thinking during the solution of conceptual and quantitative problems in one-dimensional kinematics. The model integrates three instructional components: (i) conceptual activation, (ii) teacher-guided Socratic dialogue, and (iii) teacher-guided Socratic dialogue mediated by GAI. This structure allows the evolution of students' argumentative and metacognitive reasoning to be examined under two levels of epistemic mediation while keeping the disciplinary content and problem-solving sequence constant.

The complete instructional sequence was organized into three stages and implemented over 10 h. A 2-h initial stage was devoted to activating foundational kinematics concepts and aligning conceptual criteria across participants. This was followed by two 4-h stages in which students engaged in structured Socratic problem solving, first with teacher mediation alone and then with teacher mediation supported by GAI. Table I summarizes the instructional structure of the SOCRATIC-IA model, outlining the three stages of the intervention, their pedagogical purposes, the core learning activities, and the corresponding assessment strategies used in each stage.

The problem worksheets included conceptual and quantitative tasks designed to elicit interpretation, justification, comparison of alternative explanations, and explicit verification of the physical principles involved. As a representative example, participants were asked to solve the task, which required students to identify the relevant kinematic principles, explain the relationships among

the variables, justify the selected solution pathway, and evaluate the conceptual consistency of the final response. The complete problem statement, accompanying figure, and Socratic prompts are provided in Appendix in Supplementary Material.

As shown in Table I, in both Stage 2 and Stage 3, students engaged in the same sequence of activities—individual activation of prior knowledge, collaborative dialogue, Socratic questioning, and consensus construction—followed by analytic assessment of their reasoning. The only difference between the two stages was the presence or absence of GAI within the Socratic interaction, allowing the effects of human-only versus hybrid human-AI mediation to be examined under controlled procedural conditions.

Figure 1 provides a graphical representation of the SOCRATIC-IA instructional flow. The upper portion of the diagram shows the progression from conceptual activation to critical-thinking assessment, while the lower portion depicts the two parallel mediation pathways implemented in the study: teacher-guided Socratic dialogue (Stage 2) and teacher-guided Socratic dialogue supported by generative artificial intelligence (Stage 3). By superimposing these two pathways within the same structural framework, the figure illustrates how the intervention maintained identical epistemic steps while varying only the form of mediation.

D. Ethical considerations and validity of the SOCRATIC-IA model

The present study was conducted in accordance with the ethical principles of the Declaration of Helsinki [23]. All participants provided written informed consent after being informed about the objectives of the study, the voluntary nature of participation, their right to withdraw at any time, and the procedures used to ensure confidentiality and data protection. These safeguards are particularly important in educational research involving emerging technologies, where transparency and the protection of participants' rights are essential [24].

To ensure the methodological quality of the instruments, their development followed a two-stage validation process. In the first stage, the research team designed the problem

TABLE I. SOCRATIC-IA instructional stages.

Stage	Purpose	Core activities	Assessment
Stage 1: Conceptual activation	Establish a shared conceptual baseline in kinematics	Review questions targeting velocity, acceleration, and motion representations	None
Stage 2: Teacher-guided Socratic dialogue	Develop critical thinking through human-guided Socratic questioning	Individual hypothesis generation, group discussion, teacher Socratic questioning, and consensus writing	Rubric; survey
Stage 3: GAI-mediated Socratic dialogue	Extend reasoning through hybrid human-AI epistemic mediation	Same Socratic questions explored with GAI support under teacher supervision	Rubric; survey

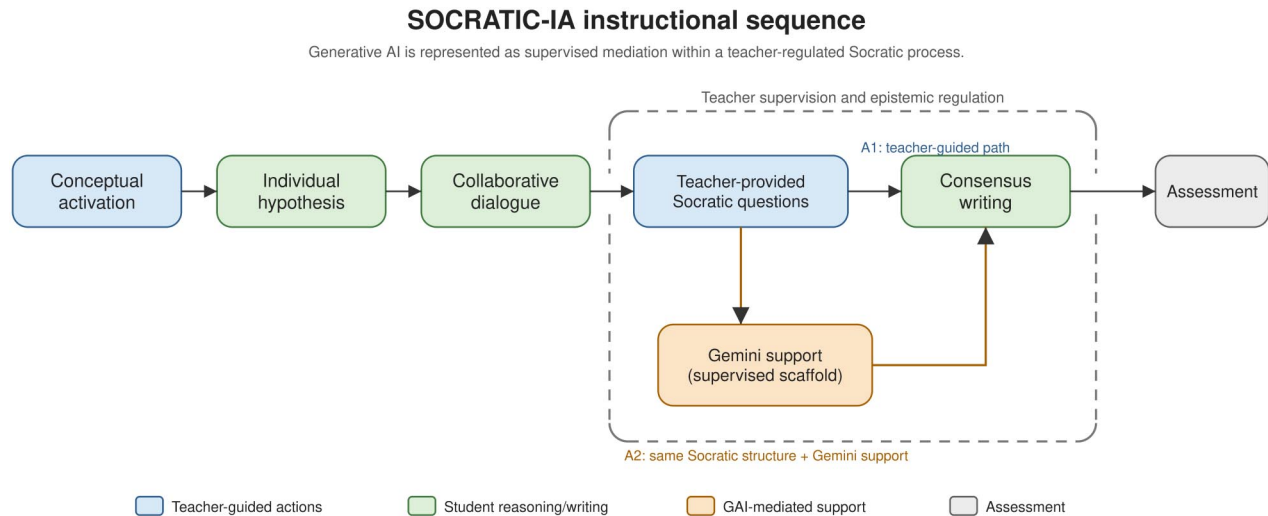


FIG. 1. Process diagram of the SOCRATIC-IA instructional stages. The lower portion of the figure shows the two parallel mediation pathways implemented in the study: teacher-guided Socratic dialogue (Stage 2) and teacher-guided Socratic dialogue supported by generative artificial intelligence (Stage 3). In both stages, the same procedural sequence is preserved, allowing the effects of human-only and hybrid human-AI mediation to be compared under controlled conditions.

worksheets, Socratic question sets, analytic rubric, and the two perception surveys. In the second stage, these instruments were reviewed through expert judgment: three specialists in physics education and pedagogy evaluated the problem sets, Socratic questions, and analytic rubric, while three experts in statistics reviewed the survey instruments. Based on their evaluations, content validity was assessed using the CVI. All instruments achieved a CVI of 1.00, indicating complete agreement regarding the clarity, relevance, and representativeness of the items, which corresponds to an excellent level of content validity according to established methodological standards.

III. EFFECTS OF SOCRATIC AND AI-MEDIATED DIALOGUE ON CRITICAL THINKING IN KINEMATICS

A. Performance on conceptual and quantitative kinematics problems

The comparative analysis of the six groups revealed a consistent pattern: across all evaluated situations, Socratic mediation supported by GAI was associated with systematic improvements in response quality for both conceptual and quantitative kinematics problems. The use of the same five problems under both instructional conditions enabled a direct and conceptually aligned comparison of performance. Nevertheless, because identical items were reused, potential practice effects are acknowledged, and the results are therefore interpreted as evidence of improvement under the second condition rather than as strong causal effects.

Before conducting inferential comparisons, the distributional assumptions of the score differences were examined. A Shapiro–Wilk test, recommended for small

samples, indicated that the overall difference scores were not compatible with normality ($W = 0.736$, $p = 0.0148$, $\alpha = 0.05$) [25]. This result, together with the ordinal nature of the analytic rubric, supported the use of non-parametric methods in a small-sample context, consistent with methodological guidance in educational research and recommendations for cautious interpretation of normality tests with limited data [26]. Accordingly, the overall impact of the intervention was evaluated using a one-tailed Wilcoxon signed-rank test, justified by the a priori directional hypothesis that GAI mediation would not reduce performance but would instead produce either improvement or no change. The use of a one-tailed test was explicitly specified to avoid *post hoc* inferential bias [27–29].

Formally, the global null hypothesis H_{0G} stated that the median of the difference scores between conditions was less than or equal to zero, that is, that Activity 2 (A2) did not outperform Activity 1 (A1) in total score. The global alternative hypothesis H_{1G} stated that the median difference was greater than zero, indicating that A2 produced higher scores than A1. The results supported this prediction: H_{0G} was rejected in favor of H_{1G} , with a Wilcoxon signed-rank statistic of 0 and $p = 0.0312$ at $\alpha = 0.05$. The mean improvement of 8.3 points per group suggests a consistent effect of the hybrid teacher-GAI mediation, although reporting the corresponding effect size is recommended to complete standard statistical reporting [29].

Beyond the global analysis, the structure of score differences was examined at the problem level. As shown in Table II, Shapiro–Wilk tests applied to the problem-specific differences did not indicate significant departures from normality for Problem 1 (P_1), Problem 3 (P_3), and Problem 5 (P_5), whereas evidence of non-normality was

TABLE II. Shapiro–Wilk tests for problem-level score differences.

Problem	W (Shapiro–Wilk)	p value	Interpretation ($\alpha = 0.05$)
1	1.0000	1.0000	Fail to reject the null hypothesis; distribution compatible with normality
2	0.7501	0.0200	Reject the null hypothesis; evidence of deviation from normality
3	0.8269	0.1012	Fail to reject the null hypothesis
4	0.7596	0.0247	Reject the null hypothesis; significant deviation from normality
5	0.9076	0.4207	Fail to reject the null hypothesis

TABLE III. Problem-level comparison of scores between activities using the Wilcoxon signed-rank test.

Problem	N diffs $\neq 0$	Median Activity 1	Median Activity 2	Median diff. (A2–A1)	W	p value (two-tailed)
1	6	10.0	11.0	1.0	0	0.03125
2	5	9.0	11.5	2.5	0	0.06250
3	5	9.0	11.0	2.0	0	0.06250
4	6	6.5	10.0	3.5	0	0.03125
5	4	8.5	9.5	1.0	0	0.12500

observed for Problem 2 ($P2$) and Problem 4 ($P4$). Given the small sample size ($N = 6$) and the ordinal nature of the rubric scores, a nonparametric approach was retained for all problem-level contrasts in order to avoid unstable analytical decisions based on low-power normality tests [26].

Subsequently, Wilcoxon signed-rank tests with two-tailed contrasts were applied separately for each problem, as summarized in Table III. This analysis enabled paired comparisons of scores obtained in Activity 1 ($A1$) and Activity 2 ($A2$) at the problem level, reporting the medians for each condition as well as the median of the difference scores ($A2 - A1$). From a descriptive perspective, a consistent directional pattern is observed: across all five problems, the median score in $A2$ exceeds that in $A1$, with median differences ranging from 1.0 to 3.5 points. From an inferential standpoint, at $\alpha = 0.05$, statistically significant differences were identified for Problems 1 ($P1$) and 4 ($P4$), whereas for Problems 2 ($P2$), 3 ($P3$), and 5 ($P5$), the results did not reach the threshold for statistical significance, despite exhibiting the same direction of improvement.

For each problem k , hypotheses were formulated as follows for the two-tailed tests. The null hypothesis H_{0k} stated that the median of the score differences between $A2$ and $A1$ was equal to zero, indicating no systematic change between conditions. The alternative hypothesis H_{1k} stated that the median difference was different from zero, indicating a systematic change between conditions. The choice of a two-tailed contrast at the problem level is justified as a methodologically conservative approach and is aligned with standard reporting practices when multiple comparisons are conducted and when unwarranted directional testing without a strict and explicitly declared *a priori* theoretical constraint is to be avoided—a practice widely

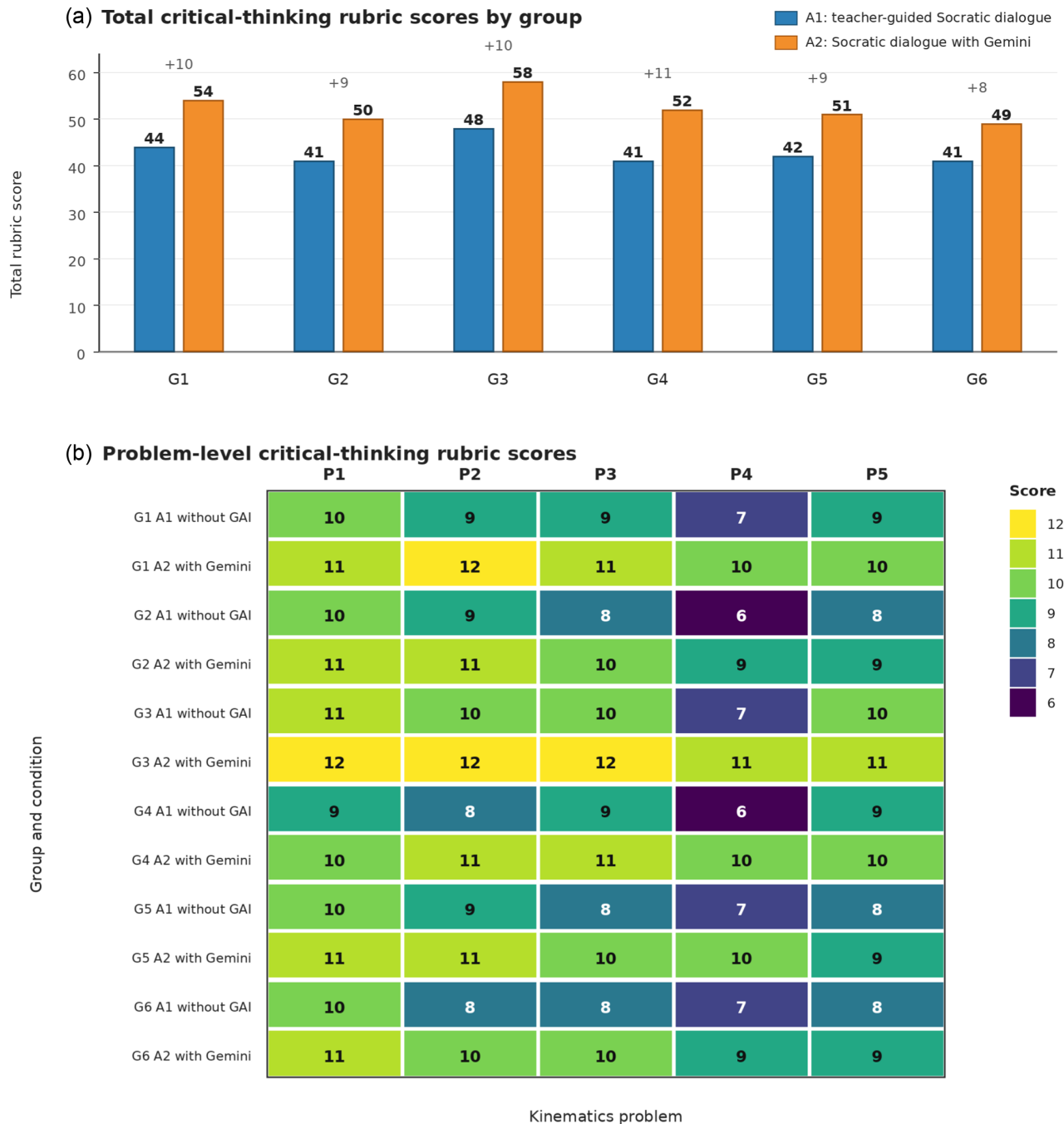
cautioned against in the methodological literature due to its potential to bias inference [28,29].

However, when the theoretical framework specifies *a priori* a clear directional expectation—namely, that mediation supported by generative artificial intelligence should not result in performance detriment and that any observable change, if present, would be in the direction of improvement—the use of a one-tailed contrast can be defended, provided that this directionality is declared before data inspection and reported transparently. Under this condition, for each problem k , the null hypothesis H_{0k} is defined as the median of the difference scores being less than or equal to zero, whereas the alternative hypothesis H_{1k} is defined as the median being greater than zero. This formulation focuses inference on the theoretically relevant direction of change while reducing the risk of opportunistic reinterpretation of results [27–29].

Figure 2 presents a heatmap visualization of the redistribution of critical-thinking scores for six groups across five kinematics problems, comparing Activity 1 ($A1$), implemented through traditional teacher-guided Socratic dialogue, with Activity 2 ($A2$), implemented through the same Socratic approach but mediated by generative artificial intelligence, following the same problem order. This matrix-based visualization facilitates the identification of patterns and contrasts between instructional conditions by representing numerical magnitudes through color gradients and enabling rapid comparative reading across rows and columns [30].

The use of a cold-to-warm color scale to represent lower to higher scores makes both the direction and intensity of changes visually salient, leveraging the perceptual correspondence between magnitude ordering and visual ordering. This approach enhances the discrimination of score increases and decreases between conditions and supports

Critical-thinking rubric scores across groups and kinematics problems



Descriptive visualization; scores should be interpreted in light of the fixed-order preexperimental design.

FIG. 2. Critical-thinking rubric scores across groups and kinematics problems. (a) Total critical-thinking rubric scores for the six groups in Activity 1 (A1; teacher-guided Socratic dialogue) and Activity 2 (A2; teacher-guided Socratic dialogue mediated by generative artificial intelligence); values above each pair of bars indicate the change in total score from A1 to A2. (b) Problem-level critical-thinking rubric scores for the same six groups across five kinematics problems (P1–P5) under A1 and A2. Color intensity represents score magnitude from lower (cool colors) to higher values (warm colors), enabling visual comparison across groups, problems, and instructional conditions. The visualization is descriptive, and the scores should be interpreted in light of the fixed-order preexperimental design.

TABLE IV. Triangulation of quantitative and qualitative results using a convergence matrix.

Finding	Quant. evidence	Qual. themes	Conv.	Integrated synthesis
Global improvement with GAI	Wilcoxon (global): $W = 0.0000$, $p = 0.03120$	Conceptual clarity/usefulness (44%); reasoning structure (17%)	Agree	Performance gains align with reports of greater clarity and better-organized reasoning.
Improvement in P1 and P4	P1, P4: $p = 0.03125$; P4 median: 6.5 $\rightarrow$ 10.0	Critical use/verification (33%)	Partial	Largest gains coincide with stronger verification and critical control.
Positive trends (P2, P3) without significance	P2, P3: $p = 0.06250$	No problem-specific theme	Silence	Moderate gains appear without a distinct qualitative signature by problem.
Perceived usefulness of GAI	PU-GAI: 100% agreement	Conceptual clarity/usefulness (44%)	Agree	Perceived usefulness matches reports of conceptual support and clarity.
Nuances in trust and integration	TE-AI, RI: high agreement; neutrality + isolated disagreement	Prompt dependence (22%); math inaccuracies (28%)	Agree	Reservations match cautions about limits and the need for verification.
Structuring/metacognition	A1: CT dispersion/neutrality; A2: high convergent profile	A1: CT + metacognition (39%); A2: reasoning structure	Partial	Metacognitive/structural reports track the overall pattern without isolating dimensions.

efficient visual inference of performance shifts across problems and groups [31,32].

Taken together, the findings support a coherent and convergent empirical narrative across levels of analysis. At the global level, the alternative hypothesis H_{1G} received statistical support, indicating a significant improvement in performance when Socratic instruction incorporated mediation by GAI. At the problem level, the alternative hypothesis H_{1k} was supported for two of the five problems, while the remaining three exhibited a homogeneous directional pattern favoring Activity 2 (A2). In studies with small sample sizes, this pattern is methodologically informative, as consistent directionality across multiple comparisons can be compatible with a stable underlying effect even when statistical significance is not reached in all individual contrasts, particularly under limited power and conservative testing strategies at the problem level [26,29].

In addition, no negative changes were identified for any group or problem, suggesting that the incorporation of GAI did not produce observable performance detriment within the examined framework. However, this absence of decreases should be interpreted cautiously: it does not constitute inferential evidence of “no harm,” but rather indicates that no systematic declines were observed under the present sample and conditions.

Taken together, the evidence integrated across Tables III and IV and Fig. 2 supports the claim that hybrid human-AI mediation can strengthen processes associated with critical thinking in problem solving, particularly by promoting more precise explanations, explicit verification of assumptions, and metacognitive self-regulation during reasoning, provided that teacher supervision and explicit criteria for response validation are maintained [33,34].

B. Quantitative analysis of shared perception dimensions across instructional conditions

To characterize the learning experience under both Socratic instructional modalities—teacher-guided and teacher-GAI-mediated—perception surveys were administered to assess dimensions common to Activities 1 (A1) and 2 (A2). The dimensions evaluated were KSE, CL, and IM. Comparison across these shared dimensions indicates that the transition toward hybrid teacher-GAI mediation did not merely represent a change of tool, but rather a modification in how prospective physics teachers experienced the task.

For KSE, a clear increase was observed. Overall agreement rose from 84.7% to 94.4%, responses in the “Disagree” category disappeared, and the proportion of “Strongly agree” responses increased from 26.4% to 37.5%. This pattern suggests enhanced confidence in understanding and solving kinematics problems under the GAI-mediated condition. The response pattern for CL was more heterogeneous. Overall agreement increased modestly from 59.7% to 65.3%, yet some items exhibited polarization, with responses clustering at opposite positions. This indicates that, for certain items, part of the group reported higher mental demand, while others reported no increase or even a reduction in perceived effort. For instance, agreement on item CL_1 increased from 77.8% to 88.9%, indicating that several participants perceived greater attentional effort. At the same time, others described a redistribution of effort, associated with verification activities and quality control of responses generated by the GAI. Finally, IM exhibited the most pronounced effect: agreement reached 100% in A2, compared to an initial 87.5%, eliminating all neutral and disagreement responses.

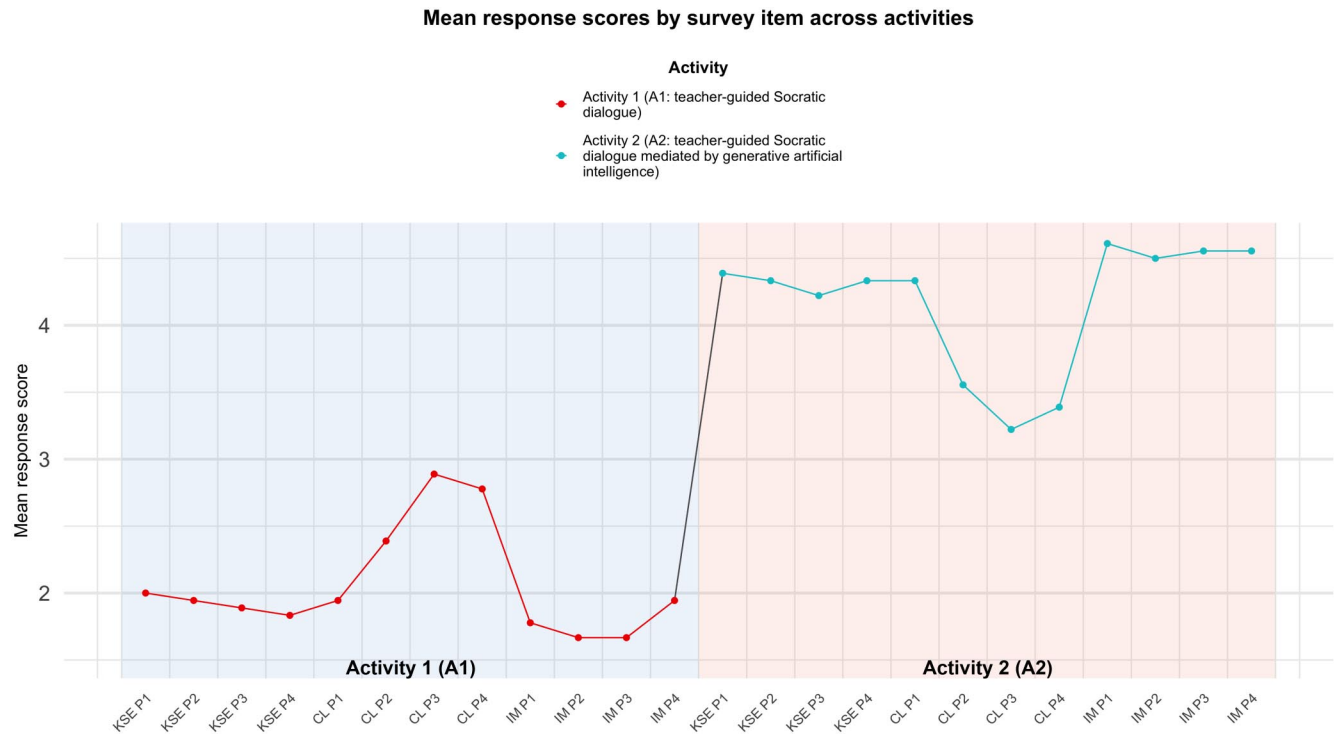


FIG. 3. Mean response scores by survey item for Activity 1 (A1; teacher-guided Socratic dialogue) and Activity 2 (A2; teacher-guided Socratic dialogue mediated by generative artificial intelligence). Items correspond to shared perception dimensions—KSE, CL, and IM. The figure illustrates the overall increase in KSE and IM under the GAI-mediated condition, while cognitive load exhibits greater variability across items.

In summary, hybrid teacher-GAI mediation strengthened self-efficacy and intrinsic motivation, whereas cognitive load displayed a differentiated response, reflecting the dual nature of the tool as both a facilitator and a source of demands for critical supervision. Figure 3 illustrates the evolution of mean scores by item, allowing for a cautious conclusion: although age-based subgroups are small, KSE and IM converge toward high levels in A2, while CL remains the most variable dimension. From a pedagogical standpoint, the perceptual benefits of GAI appear to be broadly distributed in terms of competence and motivation; however, the experience of mental effort seems to be modulated by individual differences such as prior experience, study habits, ability to formulate effective queries, tolerance for uncertainty, and the need for validation—a pattern also reported in recent studies on factors conditioning the educational use of ChatGPT and related tools [35].

C. Quantitative analysis of activity-specific perception dimension

In the perceptual analysis of dimensions not shared between the two activities, a highly positive evaluation pattern emerges, with nuanced differences associated with cognitive complexity. In Activity 1 (A1), SI consistently reached high levels of agreement across all indicators,

ranging from 88.9% to 100%, with minimal neutrality (0%–5.6%) and no disagreement. In contrast, CT was also predominantly evaluated favorably but exhibited greater dispersion, with agreement levels ranging from 77.8% to 94.4%, increased neutrality (5.6%–22.2%), and isolated instances of mild disagreement (5.6%).

In Activity 2 (A2), PU-GAI achieved unanimous approval across all indicators (100% agreement), corroborating the role of GAI as a learning support. More nuanced patterns emerged for TE-AI and RI. Although overall agreement remained high (88.9%–100%), neutrality reached up to 11.1% for TE-AI and up to 16.7% for RI, along with isolated instances of mild disagreement for RI (5.6%). Rather than weakening the findings, these patterns ground them: students confirm the usefulness of GAI while simultaneously identifying conditions for its effective use, such as the need for verification, sensitivity to prompt formulation, and awareness of potential errors—features consistent with a more reflective and self-regulated mode of engagement.

In summary, the two instructional activities define a convergent profile characterized by high acceptance of both Socratic interaction and GAI mediation. Within this profile, more immediate benefits—such as perceived usefulness, clarity, and guidance—are nearly unanimous, whereas more complex constructs, such as critical thinking, and dimensions related to criteria and reliability of

use—such as trust in educational AI and resource integration—exhibit higher levels of neutrality and occasional disagreement. These response patterns are consistent with processes of deep learning and the progressive development of reasoning.

D. Qualitative analysis of student reflections across instructional conditions

A thematic analysis suitable for small samples, following the framework proposed by Braun and Clarke [36], was conducted to examine the underlying experiences reflected in the open-ended responses. In Activity 1 (A1), students were asked, “What did you learn from the teacher’s Socratic questions?” Analysis of the 18 responses revealed three dominant themes. The most frequently reported theme was conceptual clarification and deepening (44%), highlighting the role of guided questioning in refining understanding of kinematics concepts. This was followed by critical thinking and metacognitive awareness (39%), where students emphasized reflection on assumptions, reasoning processes, and error detection. A third theme concerned the structuring of reasoning (28%), with respondents noting improved organization and coherence in their problem-solving explanations.

In Activity 2 (A2), students responded to the question, “In which aspects was generative AI useful or confusing?” The thematic distribution shifted while preserving continuity with A1. Conceptual usefulness and clarity again emerged as a dominant theme (44%), indicating that GAI was perceived as reinforcing understanding. A substantial proportion of responses emphasized critical use and verification (33%), reflecting students’ awareness of the need to evaluate, contrast, and validate AI-generated outputs. Additional themes included structuring of reasoning (17%) and instrumental efficiency (17%), the latter associated with faster exploration of solution pathways and alternative explanations. Alongside these benefits, students explicitly identified limitations: prompt dependence (22%) and mathematical inaccuracies (28%), which required continuous verification. Rather than undermining the instructional value of GAI, these limitations appear to have fostered more deliberate, reflective, and self-regulated engagement with the problem-solving process.

E. Convergent mixed-methods interpretation of learning outcomes

To integrate evidence from performance outcomes (critical-thinking rubric), perceptions (survey data), and reported experiences (open-ended responses), a qualitative-quantitative convergence matrix was employed, as presented in Table IV. This integrative approach enables systematic comparison of findings and the identification of coherent patterns across data sources without redundancy in result reporting [37,38].

Within the quantitative component, the global comparison between activities showed higher scores under the GAI-mediated condition (A2), with a Wilcoxon statistic of $W = 0.0000$ and a p value of 0.03120. At the problem level, statistically significant differences were identified for Problems 1 ($P1$) and 4 ($P4$), both with $p = 0.03125$. In particular, $P4$ exhibited a marked median increase, from 6.5 in A1 to 10.0 in A2. For $P2$ and $P3$, a favorable trend toward A2 was observed ($p = 0.06250$), although the conventional significance threshold of $\alpha = 0.05$ was not reached. Finally, $P5$ showed a more moderate change, with $p = 0.12500$. Regarding perceptual evidence, a highly favorable evaluation was identified. In A2, PU-GAI reached 100% agreement. Dimensions associated with reliability and criteria of use, such as TE-AI and RI, maintained high levels of agreement, although with neutrality levels of up to 11.1% and 16.7%, respectively, and an isolated instance of disagreement (5.6%) for RI. Taken together, these results indicate high acceptance accompanied by cautious appraisal.

The qualitative analysis revealed thematic cores that complement the observed quantitative patterns. In A2, conceptual usefulness and clarity emerged in 44% of responses, critical use and verification in 33%, and structuring of reasoning in 17%. Reported limitations included mathematical inaccuracies (28%) and prompt dependence (22%).

Overall, the triangulated evidence primarily indicates convergence between performance gains and high perceived usefulness, reinforced by narratives emphasizing clarity and instructional scaffolding. In addition, partial agreement is observed for dimensions related to trust and resource integration, where quantitative acceptance coexists with qualitative cautions highlighting the need for verification and sensitivity to prompt formulation.

IV. DISCUSSION

This study compared two forms of Socratic mediation in the resolution of conceptual and numerical problems in translational kinematics within initial physics teacher training: one guided exclusively by the instructor and another supported by GAI. Across the full set of findings, a consistent pattern emerges. Performance assessed through a critical-thinking rubric and several learning-related perceptions tended to improve when Socratic dialogue was complemented by GAI, whereas cognitive load did not follow a uniform trajectory and appeared to be modulated by the additional effort required for verification and quality control of AI-generated responses.

From a performance perspective, the observed global improvement is consistent with recent reports positioning AI-based tutors as supports capable of enhancing learning in authentic educational contexts, including physics, when they are integrated with clear pedagogical intent rather than deployed as substitutes for students’ intellectual

work [39]. Along similar lines, multiple studies have shown that AI-supported Socratic dialogues can foster reasoning and critical-thinking processes by sustaining probing questions, demanding justification, and encouraging argumentative revision [40,41]. Evidence from mathematics education further supports this mechanism: when dialogical exchanges are structurally constrained, AI tools can promote deeper thinking and autonomous discovery rather than mechanical response generation [42].

The contribution of the present work becomes more meaningful when situated against a long-standing difficulty in the teaching of kinematics: the tendency to prioritize algorithmic procedures and data substitution over modeling, variable interpretation, and metacognitive reflection [1,2]. Research on active methodologies in science education suggests that critical thinking is strengthened when shared argumentation, conceptual analysis, and guided inquiry are emphasized rather than mere repetition of procedures [3,4,43]. In translational kinematics specifically, learning requires the construction of mental models, the interpretation of relationships between quantities, and abstract reasoning—competencies that are central to the professional formation of future physics teachers. Within this framework, the Socratic method provides a plausible pathway for sustaining explanations grounded in reasons rather than results, and its articulation with GAI may expand opportunities for feedback and error reconstruction, provided that interaction remains reflective and pedagogically supervised [6].

Within physics education, emerging studies already report uses of ChatGPT aimed at strengthening teachers' dialogical skills and exploring customizable Socratic tutors, with the potential to foster active reflection and adaptive scaffolding [44,45]. Complementary work has shown that GAI can support conceptual understanding and learning efficiency by offering immediate feedback, alternative solution pathways, and opportunities for self-correction [46]. It is also relevant that comparisons between student-generated solutions and AI-generated responses allow the identification of both strengths and blind spots in each mode of reasoning, opening space for instructional practices centered on contrast, justification, and critical evaluation [47].

From the participants' perspective, the observed increase in self-efficacy and intrinsic motivation aligns with studies reporting positive effects of GAI on engagement and learning perceptions, although such effects are contingent on task characteristics and the quality of pedagogical guidance [48]. At the same time, the literature cautions that unguided use of chatbots may lead to poorly structured interactions, superficial reflection, and even reduced performance, particularly when argumentation and metacognitive control are not explicitly required [39]. In this sense, the practical contribution of the SOCRATIC-IA approach does not lie in “using AI” *per se*

but in framing its participation as a cognitive scaffold: prompting explanations, contrasting assumptions, requesting evidence, checking dimensional and conceptual consistency, and enforcing reasoning traceability before accepting a response.

Nevertheless, the methodological limitations identified call for a cautious interpretation of the findings. The preexperimental design, fixed sequence of conditions, and repeated use of the same problems across phases introduce strong threats to internal validity, including practice effects, familiarity, and potential carryover. Additionally, the small sample size and the mixture of analytical units—group-level performance versus individual-level perceptions—constrain the scope of inference. For a high editorial standard, it also remains necessary to report the reliability of the scoring process more explicitly, for instance, through interrater consistency, and to complement statistical significance with effect sizes and confidence intervals so that the magnitude of change is as transparent as its statistical support.

Taken together, the study is best interpreted as situated and preliminary evidence. It suggests that Socratic mediation can be strengthened through GAI when interaction is deliberately designed to demand higher-order thinking and when the teacher retains the role of epistemic regulator of the dialogue. Future research should incorporate counterbalancing or crossover designs, truly isomorphic tasks, larger samples, and assessments of retention and transfer. In addition, deeper qualitative analyses of the dialogues themselves—such as question quality, justification patterns, and mechanisms of self-correction—would help disentangle more precisely which aspects of change can be attributed to the Socratic strategy, which to GAI mediation, and which to accumulated practice.

V. CONCLUSIONS

This study compared the traditional Socratic method with a Socratic approach mediated by GAI to enhance critical thinking in prospective physics teachers during the resolution of conceptual and numerical problems in translational kinematics. The preexperimental mixed-methods design contributes to physics education research and related STEAM disciplines by providing empirical evidence of the role of GAI as a catalyst for deep learning and the development of higher-order cognitive skills.

The findings indicate that higher critical-thinking scores were observed during the second, GAI-mediated Socratic dialogue condition. Although this pattern was consistent across the participating groups, the fixed sequential design does not allow the observed differences to be attributed independently to GAI mediation, because practice, accumulated learning, familiarity, and carryover effects cannot be excluded. Specifically, statistically significant within-sample differences were observed in the quality of responses for both conceptual and numerical questions

in Problems 1 and 4, along with positive trends in the remaining problems analyzed. These results suggest that GAI can enrich Socratic dialogue by supporting conceptual accuracy, inferential elaboration, and metacognitive revision. The absence of negative performance impacts, together with the overall improvement observed, supports the pedagogical feasibility and potential benefits of integrating GAI tools within this disciplinary context.

Analysis of the secondary research questions further revealed that GAI mediation positively influenced students' perceptions of learning. Significant increases were observed in kinematics self-efficacy and intrinsic motivation among prospective teachers, who evaluated the role of GAI in their learning process favorably. These subjective gains were consistent with the objective improvements in performance. Cognitive load, however, exhibited a more complex pattern. While GAI facilitated access to information and conceptual clarification, it also required students to engage in monitoring and quality control of generated responses. This finding underscores the importance of pedagogical designs that integrate GAI as a teacher-supervised cognitive scaffold rather than as an answer-providing tool.

Within the conditions examined in this study, the SOCRATIC-IA model provides a preliminary instructional framework for integrating teacher-guided Socratic dialogue and generative AI in kinematics problem solving. The findings support further investigation of its feasibility and pedagogical value, but they do not establish its effectiveness beyond the present sample and instructional context. Further research using larger samples, controlled or counterbalanced designs, equivalent nonidentical tasks, and measures of retention and transfer is needed to determine whether the observed patterns can be replicated and attributed more specifically to GAI-mediated Socratic support.

By integrating Socratic dialogue with the adaptive and feedback capabilities of GAI, this approach not only enhances conceptual and numerical problem solving in translational kinematics but also fosters self-efficacy and intrinsic motivation—key factors for autonomous and lifelong learning. More broadly, this research opens avenues for exploring how pedagogically grounded

applications of generative AI can transform STEAM learning environments by cultivating deeper, more transferable forms of reasoning.

APPENDIX: SUPPLEMENTARY MATERIALS

The supplementary materials for this article are available as open online resources in Figshare and document the instructional, assessment, and AI-mediated components of the intervention. Supplementary Material S1: Kinematics Problem Set presents the conceptual and quantitative kinematics problems used during the learning activities, including tasks on average speed, average velocity, acceleration, graphical interpretation, and relative motion (<https://doi.org/10.6084/m9.figshare.32733156>). Supplementary Material S2: Socratic Questions provides the teacher-designed Socratic prompts aligned with the analytic rubric, showing how students' reasoning was guided across conceptual understanding, argumentation, examples, error correction, and clarity of response (<https://doi.org/10.6084/m9.figshare.32733849>). Supplementary Material S3: Analytic Rubric contains the scoring instrument used to evaluate discipline-specific critical thinking in kinematics through five criteria and four performance levels (<https://doi.org/10.6084/m9.figshare.32735553>). Supplementary Material S4: Prompt Structure describes the role, context, instructions, and output format used during the generative AI-mediated phase, as well as the boundaries of teacher supervision and conceptual validation (<https://doi.org/10.6084/m9.figshare.32735838.v1>). Finally, Solution to Problem 1 provides a worked example of a student-group response on average speed and average velocity in circular motion, illustrating the type of reasoning, mathematical justification, and AI-supported Socratic reflection analyzed in the study (<https://doi.org/10.6084/m9.figshare.32751666.v1>).

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

-
- [1] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020106 (2015).
 - [2] I. Sánchez Soto, M. A. Moreira, and C. Caballero Sahelices, Implementation of a proposal for meaningful learning of kinematics through problem solving, *Ingeniare* **17**, 27 (2009).
 - [3] B. Mitrevski, Teaching critical thinking and problem solving in physics, *AIP Conf. Proc.* **2075**, 180001 (2019).
 - [4] M. Ubaidillah, H. Hartono, P. Marwoto, W. Wiyanto, and B. Subali, How to improve critical thinking in physics learning? A systematic literature review, *J. Educ. Cultural Psychol. Stud.* **161** (2023).

- [5] L. Møller Haase, J. Egelund Holgaard, and A. Kolmos, The complexity of engineering education in a mission driven PBL university, *Transforming Engineering Education*, OA ed. (Aalborg Universitetsforlag, 2023), pp. 172–177.
- [6] Y.-R. Ho, B.-Y. Chen, and C.-M. Li, Thinking more wisely: Using the socratic method to develop critical thinking skills amongst healthcare students, *BMC Med. Educ.* **23**, 173 (2023).
- [7] S. Küchemann, S. Steinert, N. Revenga, M. Schweinberger, Y. Dinc, K. E. Avila, and J. Kuhn, Can ChatGPT support prospective teachers in physics task development?, *Phys. Rev. Phys. Educ. Res.* **19**, 020128 (2023).
- [8] N. H. Pulungan, M. Sirait, and E. M. Ginting, Analysis of linear motion kinematics on student's critical thinking skillsbased on scientific inquiry, *J. Phys.: Conf. Ser.* **1811**, 012006 (2021).
- [9] J. D. Quiceno Osorio, La inteligencia artificial y el riesgo de una analogía invertida, *Sophia Colección de Filosofía de la Educación* **39**, 315 (2025).
- [10] A. Recio Sastre, La pregunta por la enseñanza de la ética como límite de la inteligencia artificial, *Sophia, Colección de Filosofía de la Educación* **39**, 53 (2025).
- [11] H. Barrios Tao, Configuración de subjetividades mediante sistemas de inteligencia artificial, *Sophia, Colección de Filosofía de la Educación* **39**, 83 (2025).
- [12] R. A. Reza Flores, R. M. Vicari, C. Galafassi, and M. Usart Rodríguez, Percepción del docente sobre la otredad en la era de la inteligencia artificial, *Sophia, Colección de Filosofía de la Educación* **39**, 281 (2025).
- [13] N. A. Zambrano-González and J. Vallverdú Segura, Retos del aprendizaje en la era de las heurísticas artificiales, *Sophia, Colección de Filosofía de la Educación* **39**, 167 (2025).
- [14] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, 3rd ed. (SAGE Publications, Thousand Oaks, CA, USA, 2017).
- [15] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. (SAGE Publications, Thousand Oaks, CA, USA, 2017).
- [16] R. H. Ennis, Ideal critical thinkers are disposed to, *Inquiry* **26**, 4 (2011).
- [17] D. F. Halpern, *Thought and Knowledge: An Introduction to Critical Thinking*, 5th ed. (Psychology Press, Taylor & Francis Group, New York London, 2014).
- [18] R. Paul and L. Elder, *The Miniature Guide to Critical Thinking Concepts and Tools* (Bloomsbury Publishing PLC, London, UK, 2019).
- [19] P. A. Facione, *Critical Thinking: What it Is and Why it Counts*, (Insight Assessment, 2023), <https://insightassessment.com/iaresource/critical-thinking-what-it-is-and-why-it-counts/>.
- [20] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. (SAGE Publications Ltd, London, UK, 2018), Vol. 1.
- [21] D. C. Howell, *Statistical Methods for Psychology* (Cengage Learning, Belmont, CA, USA, 2009).
- [22] F. J. Gravetter and L. B. Wallnau, *Statistics for the behavioral sciences*, 10th ed. (Cengage Learning, Boston, MA 02210 USA, 2017). Vol. 1.
- [23] World Medical Association, World medical association declaration of Helsinki: ethical principles for medical research involving human subjects, *Jama, J Am Med Assoc* **310**, 2191 (2013).
- [24] E. A. Buchanan and M. Zimmer, Internet research ethics, *The Stanford Encyclopedia of Philosophy Fall*, 2025 Ed. (2012).
- [25] N. Mohd Razali and Y. Bee Wah, Power comparisons of shapiro-wilk, Kolmogorov-smirnov, lilliefors and Anderson-darling tests (**2**, 21 (2011)).
- [26] B. W. Yap and C. H. Sim, Comparisons of various types of normality tests, *J. Stat. Comput. Simul.* **81**, 2141 (2011).
- [27] H. D. Kimmel, Three criteria for the use of one-tailed tests, *Psychol. Bull.* **54**, 351 (1957).
- [28] C. M. Lombardi and S. H. Hurlbert, Misprescription and misuse of one-tailed tests, *Austral Ecol.* **34**, 447 (2009).
- [29] G. D. Ruxton and M. Neuhäuser, When should we use one-tailed hypothesis testing?, *Methods Ecol. Evol.* **1**, 114 (2010).
- [30] L. Wilkinson and M. Friendly, The History of the Cluster Heat Map, *Am. Stat.* **63**, 179 (2009).
- [31] C. Ware, *Information Visualization*, 3rd ed. (Data Visualization Research Lab, University of New Hampshire, Durham, USA, 2012), Vol. 1.
- [32] B. Rogowitz and L. Treinish, Data visualization: The end of the rainbow, *IEEE Spectrum* **35**, 52 (1998).
- [33] E. Kasneci *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Differ.* **103**, 102274 (2023).
- [34] W. Holmes and I. Tuomi, State of the art and practice in AI in education, *Eur. J. Educ.* **57**, 542 (2022).
- [35] A. M. Al-Abdullatif and M. A. Alsubaie, ChatGPT in learning: Assessing students' use intentions through the lens of perceived value and the influence of AI literacy, *Behav. Sci.* **14**, 845 (2024).
- [36] V. Braun and V. Clarke, Conceptual and design thinking for thematic analysis, *Qualitative Psychol.* **9**, 3 (2022).
- [37] A. Killett, P. E. Langdon, H. Ryan, C. Shiggins, R. Heywood, O. F. Jimoh, M. Redley, and K. Bunning, Inclusion of adults with conditions that have the potential to affect capacity and or communication in research: triangulation from a mixed-methods study of current practice and values across multiple stakeholders, *BMJ Open* **13**, e068366 (2023).
- [38] S. Tonkin-Crine, S. Anthierens, K. Hood, L. Yardley, J. W. L. Cals, N. A. Francis, S. Coenen, A. W. van der Velden, M. Godycki-Cwirko, C. Llor, C. C. Butler, T. J. M. Verheij, H. Goossens, and P. Little and on behalf of the GRACE IN TRO/CHAMP consortium, Discrepancies between qualitative and quantitative evaluation of randomised controlled trial results: Achieving clarity through mixed methods triangulation, *Implementation Sci.* **11**, 66 (2016).
- [39] 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).
- [40] H. Fakour and M. Imani, Socratic wisdom in the age of AI: A comparative study of ChatGPT and human tutors in enhancing critical thinking skills, *Front. Educ.* **10**, 1528603 (2025).

- [41] J. Jabbour, K. Kleinbard, O. Miller, R. Haussman, and V. J. Reddi, SocratiQ: A generative AI-powered learning companion for personalized education and broader accessibility, in *Proceedings of the Workshop on Computer Architecture Education* (ACM, Tokyo Japan, 2025), pp. 1–10.
- [42] Y. Ding, H. Hu, J. Zhou, Q. Chen, B. Jiang, and L. He, Boosting large language models with Socratic method for conversational mathematics teaching, in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24* (Association for Computing Machinery, New York, NY, USA, 2024), pp. 3730–3735.
- [43] Y. P. Chavesta Incio and M. d. l. Sánchez Trujillo, El pensamiento crítico en los estudiantes de una Escuela Nacional Superior de Arte, *Revista de Estudios y Experiencias en Educación* **22**, 65 (2023).
- [44] B. Gregorcic, G. Polverini, and A. Sarlah, ChatGPT as a tool for honing teachers' Socratic dialogue skills, *Phys. Educ.* **59**, 045005 (2024).
- [45] E. Tufino and B. Gregorcic, Creating a customisable Socratic AI physics tutor, *Phys. Educ.* **60**, 065037 (2025).
- [46] E. M. Torralba, Generative AI scaffolding in physics education: A phenomenological analysis of its role and implications in STEM learning, *Schrödinger* **6**, 175 (2025).
- [47] A. Sirnoorkar, D. Zollman, J. T. Lavery, 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.* **7**, 100318 (2024).
- [48] N. A. F. Shukor and K. Osman, The Role of artificial intelligence in developing critical thinking among science students: A systematic literature review, *Int. J. Acad. Res. Prog. Educ. Dev.* **14**, 164 (2025)..