Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 20 No. 4 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijet.v20i04.57879

PAPER

Evaluating an AI-Powered Moodle Plugin for Enhancing Conceptual Understanding in Secondary Physics

Rabih Kahaleh[1,2] (✉), Victor Lopez[2], Rodrigue Imad[1,3], Elitza Maneva[2]

[1]University of Balamand, Koura, Lebanon

[2]Universitat Autònoma de Barcelona, Barcelona, Spain

[3]UMR CNRS 6285, Brest, France

rabih.kahaleh@balamand.edu.lb

ABSTRACT

This pilot study explores the feasibility of integrating large language model (LLM) assistants into physics education through a Moodle plugin designed to address conceptual understanding in Newtonian mechanics. Using OpenAI's GPT for real-time Socratic dialogue, the plugin guides students through misconception-targeted questions adapted from the Force Concept Inventory (FCI). Aligned with principles of inquiry-based learning, the intervention compares AI-guided feedback with instructor-guided materials over a one-week, three-session classroom study. Results suggest that students receiving AI-guided Socratic dialogue showed greater conceptual gains on certain targeted items (e.g., Newton's Third Law), whereas instructor guidance proved more effective for other concepts (e.g., mass and free-fall independence). Survey feedback highlights the immediacy and interactive nature of the AI while also noting a preference for the clarity provided by instructors. Qualitative analysis of open-ended question responses suggests that AI-driven dialogue promotes deeper reasoning when restating student ideas and scaffolding reflection. These preliminary findings underscore the potential of LLMs to support conceptual change in physics education when thoughtfully embedded within learning management systems, highlighting the complexity and value of personalized, interactive feedback for addressing student misconceptions.

KEYWORDS

large language models (LLMs), force concept inventory (FCI), learning management system (LMS), conceptual change, physics misconceptions, Socratic dialogue, Moodle plugin, artificial intelligence in STEM education, constructivist learning

1 INTRODUCTION

Advancements in educational technology continue to reshape the landscape of physics and engineering education, introducing innovative tools that can both scale and personalize learning. Central to ongoing challenges in this domain are persistent misconceptions about foundational concepts such as Newton's laws of motion. Despite formal instruction, students often retain intuitive but scientifically

Kahaleh, R., Lopez, V., Imad, R., Maneva, E. (2025). Evaluating an AI-Powered Moodle Plugin for Enhancing Conceptual Understanding in Secondary Physics. *International Journal of Emerging Technologies in Learning (iJET)*, 20(4), pp. 81–97. https://doi.org/10.3991/ijet.v20i04.57879

Article submitted 2025-07-25. Revision uploaded 2025-08-20. Final acceptance 2025-08-20.

inaccurate beliefs, for example, that heavier objects fall faster or that continuous force is required to sustain motion. These alternative conceptions, rooted in everyday experiences, have proven resistant to change even after repeated exposure to scientific explanations [1], [2].

Physics education research shows that students often approach physical phenomena with coherent yet flawed mental models. Conceptual change theory suggests that students learn best when they are encouraged to rethink their ideas considering new and convincing explanations [3]. Diagnostic tools such as the Force Concept Inventory (FCI) have been widely adopted to identify and analyze these misconceptions through targeted distractors [4], [5]. However, research consistently indicates that diagnosis alone is not sufficient; effective conceptual change requires intentional instructional interventions that confront and restructure students' reasoning. A substantial body of evidence supports the effectiveness of pedagogical strategies such as Socratic questioning, peer instruction, and guided inquiry in promoting conceptual change and critical thinking [6], [7], [8]. These methods encourage exploration, metacognitive reflection, and alignment of assessment and feedback with deep conceptual understanding. In recent years, technology-enhanced learning environments—particularly learning management systems (LMS) such as Moodle—have enabled the delivery of content and assessment at scale. However, traditional LMS tools, including most adaptive quizzes, are limited in dialogic depth, as most automated feedback cannot replicate the reasoning-centered, interactive support typical of expert human tutoring [9], [10]. While many LMS platforms provide immediate right/wrong feedback, they rarely support reflective dialogue that helps students uncover and address their underlying misconceptions.

Recent advancements in artificial intelligence, especially large language models (LLMs), are opening new possibilities for delivering personalized, misconception-sensitive feedback in digital environments. GPT-powered tools can now simulate reflective dialogue by engaging students in Socratic questioning that mirrors expert tutoring strategies. When integrated into platforms such as Moodle, these new-generation LLMs can simulate the dynamic, misconception-sensitive tutoring traditionally provided by experts [11], [12], [13], [14].

Early studies indicate that these integrations can improve student engagement, understanding, and self-regulation through real-time, personalized feedback [15], [16]. Given the absence of pilot testing validation for the adapted FCI items, this study can be considered as exploratory. The aim is to provide preliminary findings on the pedagogical effectiveness of integrating GPT-powered Socratic dialogue into Moodle-based physics instruction, thereby guiding future validation efforts. Despite the promise of AI, scalable methods for delivering targeted, reasoning-focused support in digital learning remain limited.

To address this gap, this pilot study investigates two primary questions:

1. How does the GPT-powered assistant plugin affect student performance on FCI-based quiz items compared to traditional instruction?
2. How do students using the GPT-powered assistant plugin perceive their own learning process compared to students following traditional instruction?

By addressing these questions, this study may open new possibilities for using AI to deliver responsive, misconception-focused feedback at scale, advancing efforts to enhance conceptual learning in STEM education.

2 MATERIALS AND METHODS

2.1 Context and participants

The study was conducted at an urban, coeducational secondary school in Lebanon, with instruction delivered in English and aligned to the national curriculum. Of the 62 students initially enrolled across Grades 10–12, 48 completed all phases and were included in the analysis. The final sample comprised students who completed the pre-test, the post-test, and the perception survey; students who missed any of these phases were excluded from analysis.

To ensure balanced group composition, students were stratified by prior achievement (low/medium/high) and gender, then randomly assigned to the experimental group (AI-assisted) or the control group (instructor-assisted). This stratified, randomized design aligns with the best practices for comparing interactive engagement and traditional instructional methods in physics education [17].

Instructional materials and activities were co-developed and supervised by a team of six experienced physics teachers (three with master's degrees in science and three with bachelor's degrees and 8–20 years of teaching experience), supporting fidelity of implementation. Participants represented a range of socioeconomic backgrounds, and all had access to school-provided laptops and Internet during study sessions. This approach is consistent with recent recommendations for implementing adaptive learning research in diverse educational settings using LMS platforms [18]. All students had met the national English proficiency standards for their grade level. Prior experience with AI or educational technology was not systematically documented; however, all students had completed school-based computer literacy modules. We acknowledge that differences in prior exposure to AI tools may have influenced student engagement—a potential limitation of the study. Attrition resulted from students failing to complete either the pre-test, post-test, or perception survey. There were no substantial differences in grade level or gender between those who completed the study and those who did not. However, detailed analysis of other characteristics among non-completers was not conducted. Strict ethical protocols were maintained, including informed consent, data anonymization, and privacy safeguards; only anonymized session IDs were logged by the plugin. Table 1 provides a summary of the study context, participant characteristics, group assignment, and ethical procedures.

Table 1. Study context and group assignment

Category	Details	Category	Details
School Context	Urban, coeducational, Lebanon; English-medium, national curriculum	**Grade Levels**	Grades 10–12
Initial Enrollment	62 students	**Final Sample**	48 students (completed all phases)
Grade Distribution	Grade 10: 14 students Grade 11: 17 students Grade 12: 17 students	**Group Size/ Assignment**	24 per group Stratified by prior achievement (low/med/high) & gender, then randomized
Control Group	Instructor-guided feedback (PDFs, follow-up Q&A)	**Experimental Group**	AI-assisted feedback (GPT-3.5 Assistant)
Teacher Involvement	6 teachers (8–20 yrs exp., 3 with master's in science, 3 with Bachelor of Science); co-developed materials and supervised sessions		
Ethics & Privacy	informed consent: data anonymized; no personal identifiers stored; plugin logs session IDs only;	**Instrumental Development**	Team-developed quizzes and surveys; peer review to minimize bias

2.2 Research design

The study employed a three-day classroom implementation, structured in four phases: instrument development, AI plugin development, controlled classroom deployment, and evaluation. Students were randomly assigned to either the experimental (AI-guided) or control (instructor-guided) group. Both groups completed a pre-test, participated in group-specific intervention sessions, and then completed a post-test and perception survey. The specific sequence of classroom activities, timing, and feedback approaches for each group are summarized in Table 2.

Table 2. Sequenced classroom sessions and group-specific feedback paths

Day	Time	Experimental Group ($n = 24$)	Control Group ($n = 24$)
Day 1 Pre-Test Quiz (50 minutes)	20 minutes	• Explain the study role to students • Highlight the importance of their participation	
	20 minutes	• Login to Moodle and Answer Pre-test 6 multiple-choice conceptual questions	
	10 minutes	• Access the pre-test quiz review through the Moodle Plugin chat panel interface • Engage in 3–6 Socratic dialogue turns per item to reflect on and revise answers	• Instructor-led discussion and concept explanation for each question • Watch a 5-minute instructional video • Participate in a 5-minute live Q&A with the instructor
Day 2 Intervention (50 minutes)	50 minutes	• Log in to Moodle • Plugin detects group assignments and redirects students to the AI chat panel (experimental) or to the instructor-prepared resource page (control)	
		• Review pre-test quiz answers through the AI chat panel and participate in 3–6 Socratic dialogue turns per item	• The instructor revisits pre-test quiz concepts on the resource page, leads discussion with students, and provides further explanation through instructional video and live Q&A session
Day 3 Post-test Quiz (50 minutes)	10 minutes	• Log in to Moodle • Plugin detects group assignments and redirects all students to the post-test quiz	
		• Access the pre-test quiz review through the AI chat panel interface • Engage in 1–2 Socratic dialogue turns per item to reflect on and revise answers	• Watch a 5-minute instructional video on the resource page • Join a 5-minute live Q&A session with the instructor to clarify key concepts
	20 minutes	• Login to Moodle and answer Post-test 6 multiple-choice conceptual questions	
	20 minutes	• Complete the perception survey to evaluate clarity, engagement, confidence, and effectiveness of the AI plugin-guided Socratic discussions	• Complete the perception survey to evaluate clarity, engagement, confidence, and effectiveness of the instructor-led discussions

This structured design ensured both groups engaged with the same core physics concepts, differing only in feedback approach: the experimental group received real-time, personalized Socratic prompts from the AI plugin, while the control group participated in instructor-led discussions. The post-test was conceptually aligned with the pre-test but modified to reduce recall, and all sessions were conducted under standardized, supervised conditions to maintain instructional fidelity.

Phase 1 – design of pre/post-tests and surveys. To develop valid and pedagogically grounded assessment tools, two core instruments were created: a pair of conceptual diagnostic quizzes and a student perception survey. The pre- and

post-tests were adapted from six items in the FCI—specifically items 5, 7, 11, 13, 17, and 21—selected for their alignment with key Newtonian mechanics concepts and common student misconceptions [17]. Six experienced high school physics teachers (four male, two females; 8–20 years' experience; two with master's degrees) collaboratively adapted the items across three two-hour workshops. Adaptation focused on systematically changing surface features and contexts to enhance relevance and reduce recall, while preserving the underlying conceptual structure. For example, the context of a "rocket in space" was replaced by a "sled on ice," while the misconception targeted and the correct response remained unchanged. Distractors were deliberately crafted to reflect misconceptions documented in the literature (e.g., the belief that continuous force is needed to keep an object moving) [4]. A conceptual mapping of pre- and post-test items, their everyday contexts, and targeted misconceptions is provided in Table 3. Although every effort was made to preserve the diagnostic function of the original items, surface feature changes may have inadvertently altered the conceptual demands of the questions, a recognized risk when adapting established instruments [4], [18].

Table 3. Conceptual mapping of pre- and post-test items

Concept	Pre-Test Context	Post-Test Context	Targeted Misconception
QC1: Newton's 1st Law (Inertia)	The rocket changes direction in space after perpendicular thrust	Sled receives sideways push on frozen lake	Misconception: Objects need a force to keep moving Scientific Idea: Objects continue in combined motion if no further force acts
QC2: Newton's 2nd Law (F = ma)	The rocket speeds up during thrust (b to c)	The shopping cart speeds up while being pushed sideways	Misconception: Force is required to keep an object moving Scientific Idea: Force causes acceleration in the direction of the push
QC3: Newton's 3rd Law (Action–Reaction)	The car pushes the truck as both speed up	The horse pulls the cart while accelerating	Misconception: The bigger object exerts more force in a collision Scientific Idea: Forces are equal and opposite, regardless of size or role
QC4: Superposition Principles (Net Force)	Elevator moving at constant speed	A wagon pulled forward by a force exactly balanced by friction, moving at a constant speed	Misconception: If an object is moving, there must be a net force acting on it Scientific Idea: Net force is zero when balanced
QC5: Friction and Force Persistence	Golf ball under gravity and air resistance	Hockey puck sliding on rough ice	Misconception: Friction only exists while pushing Scientific Idea: Friction continues acting after the initial force is gone
QC6: Mass and Free-Fall Independence	Two metal balls of different mass dropped together	Steel balls of different mass roll off table	Misconception: Heavier objects fall faster than lighter ones Scientific Idea: Mass does not affect free-fall motion

No separate pilot testing or formal validation of the adapted instruments was performed prior to deployment. While three additional physics teachers independently reviewed the instruments for conceptual equivalence and age-appropriateness, no empirical psychometric analyses (such as reliability estimation, item discrimination, or factor analysis) were conducted. This lack of formal validation represents a critical limitation and may affect the internal validity of the study findings. Although the original FCI items are well-validated [4], the adapted versions used here should be considered preliminary. The absence of pilot testing and psychometric analysis means that measurement error, changes in item difficulty, or shifts in targeted misconceptions cannot be ruled out [16]. As such, results should be interpreted with appropriate caution. In parallel, a student perception survey was developed to capture learners' reflections on clarity, engagement, and conceptual understanding. Both groups received parallel versions of the survey, with phrasing adapted to match the instructional condition (AI-guided or instructor-guided). The survey included eight Likert-scale items (QS1–QS8) and two open-ended questions (QS9–QS10), mapped in Table 4. As with the quizzes, the survey instruments were not pilot tested or formally validated, which is a further limitation.

Table 4. Mapping of parallel student perception items across experimental and control groups

Item of Perception	Experimental Group	Control Group
QS1: Understanding	The AI conversation helped me understand the concepts better	The guided material and videos helped me understand the concepts better
QS2: Mistake Correction	The AI helped me notice and correct mistakes in my thinking	The guided material and videos helped me notice and correct mistakes in my thinking
QS3: Clarity	The AI's questions and explanations were clear to follow	The guided material and its inline questions and explanations were clear to follow
QS4: Conceptual Challenge	The AI helped me challenge my previous ideas	The guided material and videos helped me challenge my previous ideas
QS5: Confidence	I felt more confident after discussing with the AI	I felt more confident after learning from the guided material and videos
QS6: Interest in Learning	Talking to the AI made me more interested in learning	Interacting with guided material and videos made me more interested in learning
QS7: Engagement	The AI conversation kept me focused and engaged	The guided material and follow-up questions kept me focused and engaged
QS8: Support & Guidance	I felt guided and supported while talking to the AI	I felt guided and supported while following guided material and videos
QS9: Open-Ended Reflection	What part of this learning experience helped you most, and why?	What part of this learning experience helped you most, and why?
QS10: Recommendation	Would you recommend this learning method for future lessons? Why or why not?	Would you recommend this learning method for future lessons? Why or why not?

Phase 2 – technical development and plugin deployment. To deliver real-time, misconception-sensitive feedback in secondary physics, a custom AI assistant and Moodle plugin were developed and deployed as part of this intervention. The system was designed to foster Socratic dialogue and formative assessment, guiding students to reflect on and refine their reasoning around core Newtonian mechanics concepts.

AI assistant development. The AI assistant was developed using OpenAI's Assistant API and configured to provide Socratic dialogue-based tutoring [19], [20]. Its pedagogical behavior was guided by a structured system prompt informed by conceptual change literature [3] and expert teacher input. The assistant was also connected to a curated vector store containing Newtonian mechanics scenarios and common misconception labels. The misconception-tagging schema and dialogue rules are available in the GitHub repository: *MoodleAIPlugin*. Figure 1 illustrates the AI assistant's configuration interface, including system prompt management and vector store integration.

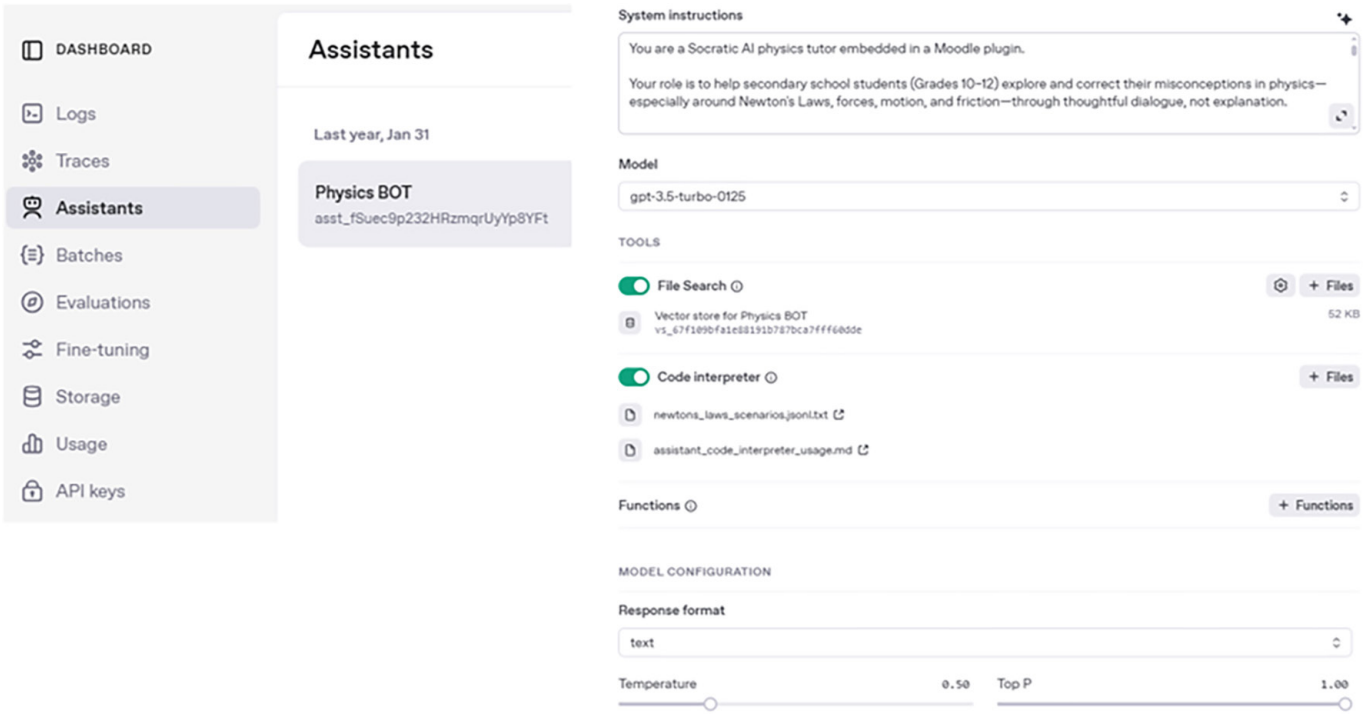

Fig. 1. OpenAI assistant configuration interface

Plugin Functionality, Structure, and Feedback Modalities: The AI assistant was integrated into a custom Moodle plugin, developed as a local module and embedded into the quiz review workflow. Each quiz item was manually tagged with its corresponding misconception category (e.g., inertia, net force, action–reaction) by consensus among physics teachers and literature review. These tags triggered targeted Socratic dialogue or instructor support upon student answer submission. The plugin automatically detected each student's group assignment and routed them accordingly.

Experimental Group: AI-Guided Feedback: Students in the experimental group accessed an embedded AI chat panel after submitting each answer

(see Figure 2). They engaged in three to six structured dialogue turns per quiz item, guided by the assistant's system prompt. Teachers supervised AI–student interactions in real time during classroom sessions and reviewed logs internally to ensure instructional appropriateness and alignment with pedagogical strategy. No formal external evaluation or systematic coding of AI responses was conducted during deployment. Dialogue memory was restricted to the current session for privacy and fairness. Session logs and transaction data were anonymized after quiz completion; no personally identifying information was stored.

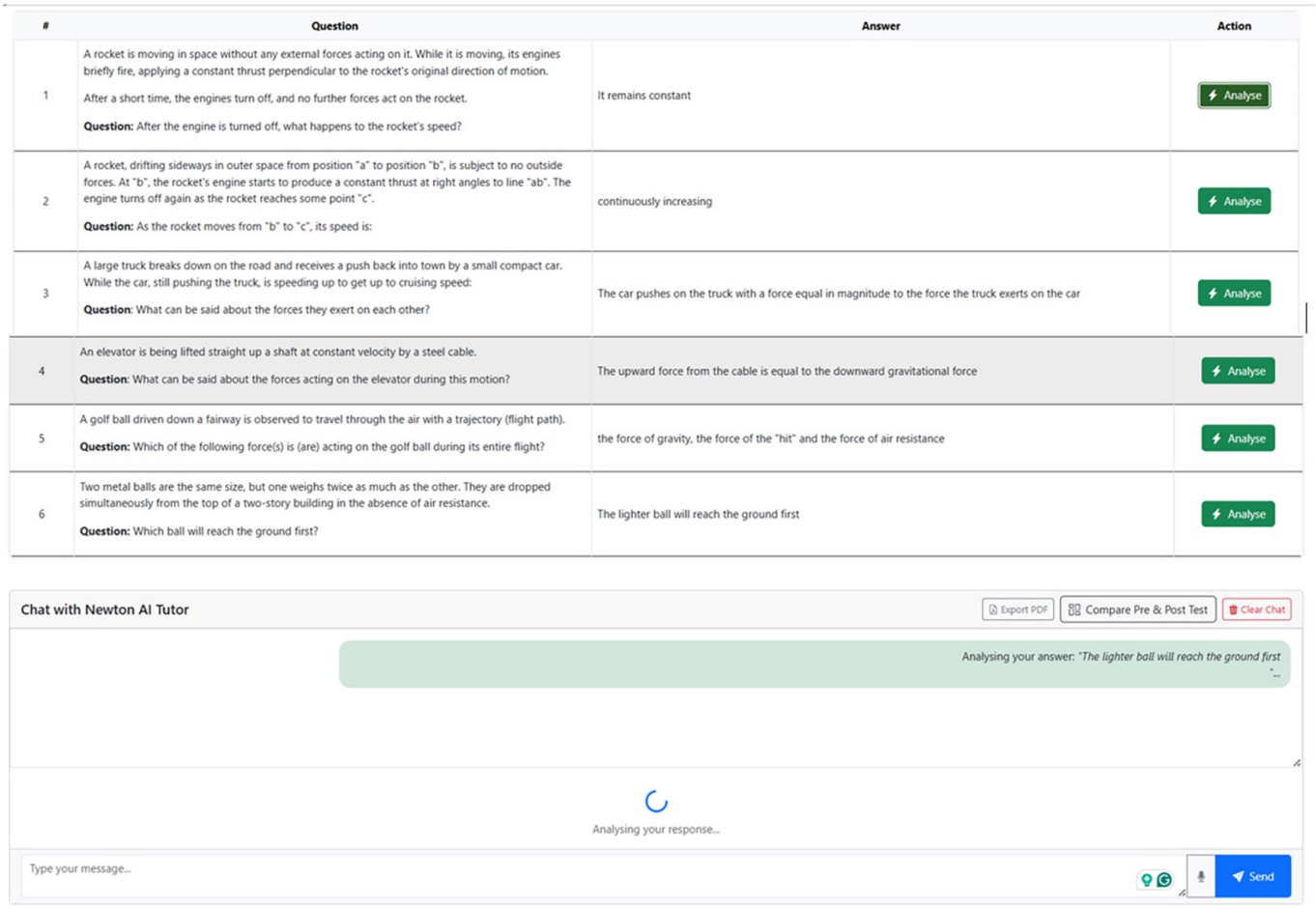

Fig. 2. Moodle plugin interface – experimental group

Control Group: Instructor-Guided Feedback: Students in the control group were redirected to a Moodle resource page after each quiz item. This page included a short instructional video, a concept summary PDF, and a reflection question aligned with the same core physics concept as the AI intervention (see Figure 3). After reviewing these materials, students participated in teacher-facilitated discussions, where instructors used structured questioning to help clarify reasoning, address misconceptions, and reinforce conceptual understanding. The intent and instructional structure of these discussions closely mirrored those of the AI-guided approach. As with the experimental group, all session logs and transaction data were anonymized after quiz completion, and no personally identifying information was stored.

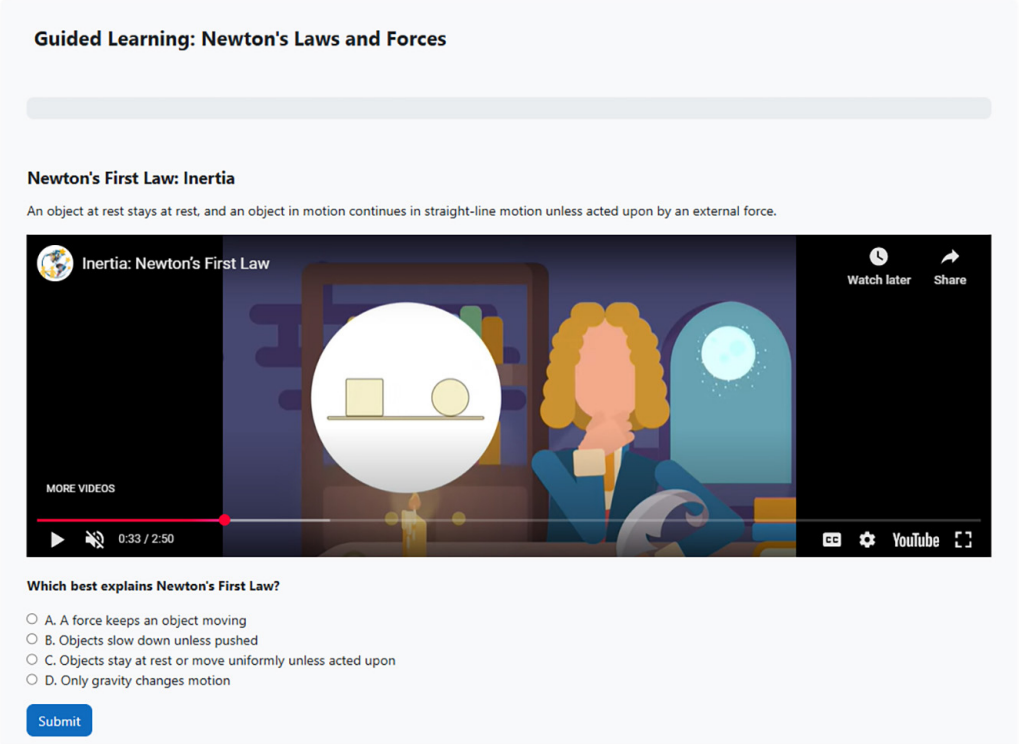

Fig. 3. Moodle instructor page – control group

Deployment and testing. Prior to classroom implementation, the plugin and AI assistant underwent two supervised dry runs with participating teachers and IT staff. These trials confirmed technical reliability, correct integration with Moodle, and alignment with instructional goals. During live classroom sessions, teachers monitored AI–student interactions to ensure appropriate tone and feedback, intervening if necessary. Misconception tagging was developed collaboratively by the research team and informed by existing literature; no formal external validation protocol was implemented for tagging accuracy at this stage.

Phase 3 – classroom implementation. Classroom implementation took place over one instructional week, following the school's standard Monday–Wednesday–Friday schedule. Activities were distributed across three 50-minute physics sessions: Day 1 included the pre-test and initial intervention, Day 2 continued the intervention, and Day 3 concluded with the post-test and a student perception survey. Each group followed a distinct instructional path: the experimental group received AI-guided feedback via the Moodle plugin, while the control group engaged with instructor-guided feedback using video tutorials followed by a live Q&A session. After completing the pre-test on Day 1, students were directed to separate classroom settings according to their group assignment. The control group joined a live session with the instructor, while the experimental group remained in the computer lab to begin interacting with the AI plugin through the Moodle interface. This setup was maintained consistently throughout the three sessions. The computer lab used for testing and intervention was equipped with 48 laptops prepared for student use and two additional laptops reserved for technical emergencies. The lab had a stable internet connection with a bandwidth of 10 Mbps, ensuring smooth access to the Moodle platform and AI plugin throughout the study (see Figure 4).

Fig. 4. Classroom deployment of the intervention. Left: Instructor-guided group viewing structured video and Q&A material. Right: AI-guided group engaging with the plugin via laptops

Phase 4 – data analysis plan. To evaluate the impact of the intervention, both quantitative and qualitative data were collected and analyzed. The core assessment consisted of six multiple-choice questions aligned with Newtonian mechanics concepts and well-documented student misconceptions. Each item was scored as either 0 (incorrect) or 1 (correct), and the proportion of students answering each item correctly was calculated within each group. These values were treated as normalized scores ranging from 0 to 1. A score of 0.333 for a given pre-test item, for example, indicates that 33.3% of students in that group selected the correct answer. Normalized scores were computed separately for the pre-test and post-test phases and averaged across students within each group. Learning gains were calculated as the raw difference between post-test and pre-test scores for each item: *Gain* = *Post-test Score* – *Pre-test Score*. These item-level gains were then averaged to obtain an overall gain per group. This approach allowed for detailed comparison of learning improvement across specific conceptual areas. To compare group performance and gains statistically, the Mann–Whitney U test was used. This non-parametric test is appropriate for modest sample sizes and non-normally distributed data.

The U statistics were calculated as: $U = \frac{n_1}{n_2} + \frac{2n_1(n_1+1)}{2} - R_1$ where n_1 and n_2 represent the sample sizes of the two groups, and R_1 is the sum of ranks for group 1. In addition to test performance, all students completed a post-intervention perception survey, which included eight Likert-scale items and two open-ended questions. The Likert items measured perceptions of understanding, clarity, mistake correction, engagement, and confidence. For consistency with the performance data presentation and to facilitate unified interpretation across all quantitative measures, the Likert scale responses were normalized to a 0–1 scale using the formula: $Normalized\ Score = \frac{Original\ Score - 1}{4}$. This transformation maintains the relative

differences between groups while providing a consistent metric across all quantitative analyses. Open-ended reflections were reviewed informally to identify recurring observations or points of feedback. A summary of the data sources and analysis methods is provided in Table 5.

Table 5. Summary of data sources, scoring, and analysis methods

Component	Description
Assessment Items	12 multiple-choice questions (6 pre-test, 6 post-test) aligned with Newtonian misconceptions
Scoring	Each item scored 0 or 1; normalized scores (0–1 scale); group averages calculated per item
Quantitative Data	Pre-test and post-test scores; item-level and group-level raw gains
Survey Data	Post-intervention perception survey with 8 Likert-scale items and 2 open-ended responses
Statistical Method	Mann–Whitney U test for between-group comparisons
Qualitative Review	Informal review of open-ended responses and AI–student dialogue interactions
Triangulation	Combined interpretation of test scores and student survey responses

Additional methodological limitations include the risk of inflated Type I error due to multiple item-level comparisons without correction, absence of effect size reporting, use of non-standard normalized scores, lack of formal survey validation, and informal qualitative analysis. These factors may affect interpretability and reproducibility; results should be interpreted accordingly. See Discussion for implications and future recommendations.

3 RESULTS

This section presents the outcomes of the intervention by integrating quantitative test performance, student survey responses, and qualitative observations from AI–student interactions. The findings provide insight into how AI-guided Socratic feedback, delivered through the LMS plugin, influenced students' conceptual understanding of Newtonian mechanics compared to instructor-guided feedback.

3.1 Quantitative analysis: learning gains and group comparison

Analysis of student performance. Table 6 presents a detailed breakdown of student performance by displaying the normalized proportion of correct responses for each quiz question, separated by group (control vs. experimental) and by test phase (pre-test and post-test). This approach allows for a clear comparison of learning gains between groups across specific conceptual items. Learning gains were calculated as the raw difference between each group's post-test and pre-test performance on every item, providing insight into the progress made through each instructional approach. To assess whether these gains differed significantly between groups, the Mann–Whitney U test was used as a nonparametric alternative appropriate for small sample sizes and ordinal data [21].

Table 6. Normalized averages with overall mean by group (0–1 scale)

Question	Pre-Test Control Group	Post-Test Control Group	Control Gain (Post-Test – Pre-Test)	Pre-Test Experimental Group	Post-Test Experimental Group	Experimental Gain (Post-Test – Pre-Test)	Mann–Whitney U p-Value
Q1	0.333	0.125	–0.208	0.5	0.292	–0.208	0.92
Q2	0.292	0.391	0.099	0.25	0.625	0.375	0.17
Q3	0.583	0.696	0.113	0.167	0.708	0.541	**0.019**
Q4	0.458	0.167	–0.291	0.542	0.292	–0.25	0.93
Q5	0.292	0.5	0.208	0.208	0.708	0.5	0.091
Q6	0.208	0.739	0.531	0.417	0.5	0.083	**0.014**
Average	0.361	0.436	0.075	0.347	0.521	0.174	0.35

Both groups showed overall improvement in conceptual understanding. In this exploratory study, the experimental group demonstrated a higher average gain (0.174 vs. 0.075), though not statistically significant overall. Item-level differences provide early evidence that instructional effectiveness may depend on the specific misconception targeted.

3.2 Analysis of student perceptions

After the instructional intervention, all participants completed a post-study survey consisting of ten items. The first eight items (QS1–QS8) measured student perceptions using a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) across dimensions such as understanding, mistake correction, clarity, challenge, confidence, interest, engagement, and support. The final two items (QS9, QS10) were open-ended, allowing participants to provide additional comments or suggestions. Given the small sample size and the limited number of open-ended questions, responses to these items were reviewed to identify general patterns and representative remarks, rather than undergoing detailed qualitative or thematic analysis. Table 7 presents the quantitative results for the eight Likert-scale items (QS1–QS8), and the open-ended responses (QS9–QS10) are analyzed qualitatively in the following paragraphs.

Table 7. Survey responses by group – normalized scale (0–1, where 0 = Strongly Disagree, 1 = Strongly Agree)

Survey Item	Control Group Mean	Experimental Group Mean	Mann–Whitney U p-Value
QS1: Understanding	0.76	0.55	**0.002**
QS2: Mistake Correction	0.61	0.59	0.888
QS3: Clarity	0.59	0.58	0.932
QS4: Conceptual Challenge	0.70	0.60	0.065
QS5: Confidence	0.65	0.62	0.634
QS6: Interest in Learning	0.78	0.64	**0.040**
QS7: Engagement	0.72	0.68	0.858
QS8: Support & Guidance	0.71	0.62	0.247

To complement the performance data, students reflected on their learning experience through the survey. Results revealed important contrasts in how each group experienced the intervention. Students in the control group, who received instructor-guided video instruction, consistently reported higher agreement across several items. For example, on QS1 ("This method helped me understand the concepts better"), the control group averaged 0.76 compared to 0.55 in the experimental group ($p = 0.002$). This perception was also evident in their open-ended responses:

- *"The teacher explanation in the video made things clear."*
- *"Watching the teacher-video while solving helped me stay focused."*

Similarly, for QS6 ("*This method made me more interested in learning*"), control group students rated their experience more positively (0.78 vs. 0.64, $p = 0.040$), suggesting that the familiar, structured format supported motivation and comfort. In contrast, students in the experimental group, who engaged in Socratic dialogue with the AI assistant, described their experience as more cognitively demanding but also more reflective. Though their mean ratings were generally lower, their open-ended responses emphasized the benefits of thinking through their reasoning:

- *"The AI made me explain my thinking. It helped me notice my mistake."*
- *"Asking questions made me understand why I chose that answer."*

Students reported that being challenged to reconsider their initial ideas and engaging in a dialogic structure—despite requiring more effort—was valuable, often prompting deeper thinking and moments of realization. While no significant differences emerged between groups on items such as Clarity, Confidence, or Support, overall patterns suggested two distinct instructional experiences: the control group benefited from clarity, reassurance, and structured explanation, whereas the experimental group experienced a more productive struggle that sometimes led to self-correction and insight. These findings indicate that the instructional methods differed not only in perceived satisfaction but also in the nature of cognitive engagement they promoted.

4 DISCUSSION AND LIMITATIONS

This study explored the potential of a GPT-powered Socratic assistant, integrated into Moodle quizzes, to support conceptual learning in Newtonian mechanics. Both the experimental (AI-assisted) and control (instructor-guided) groups demonstrated learning gains, with evidence suggesting that the effectiveness of each instructional approach may depend on the misconceptions being addressed. These findings align with the broader literature on conceptual change, which emphasizes the importance of targeted feedback and cognitive engagement in shifting students' underlying ideas. Importantly, students' perceptions of the two instructional approaches revealed notable contrasts. The control group reported greater clarity and satisfaction, appreciating the structured and familiar format of instructor-guided video instruction. In contrast, students in the experimental group described the AI-based Socratic dialogue as cognitively demanding but also valuable for prompting reflection and deeper engagement with their reasoning. This suggests a pedagogical trade-off between immediate comfort and the productive struggle that often characterizes conceptual change in learning. Several limitations should be considered when

interpreting these results. First, the modest sample size limited the statistical power of between-group comparisons and may have increased the risk of both Type I and Type II errors. Additionally, while efforts were made to preserve conceptual alignment between pre- and post-test items, subtle changes in context or wording may have influenced students' interpretations and responses, potentially affecting construct validity. Another important consideration is the consistency and fidelity of the interventions across groups. Although the AI assistant operated using structured Socratic dialogue rules, no formal analysis was conducted on the consistency or instructional quality of the AI-generated prompts. Teacher feedback indicated that some AI responses were occasionally generic or repetitive, which may have reduced their instructional impact. Similarly, the control group benefitted from a blend of video instruction and live Q&A, which offered a different mix of interaction compared to the AI-only format. Such differences complicate direct comparisons and highlight the need for more tightly controlled studies in the future. It is also important to acknowledge potential confounding variables related to the physical environment. The two groups were assigned to different settings (computer lab versus classroom), which may have influenced comfort, access to technology, or exposure to distractions. Future studies should consider rotating or standardizing physical settings to minimize these effects. Finally, the scope and generalizability of the findings are constrained by the study's context. The intervention was conducted in a single private school with English-medium instruction over a short period, and no delayed post-test was included. As such, the results may not extend to other educational contexts or provide insight into long-term conceptual change. Further research in more diverse settings, with longer follow-up, is needed to better understand the potential and limitations of AI-assisted Socratic dialogue in physics education.

5 FUTURE DIRECTIONS AND CONCLUSION

This study provides encouraging early indications that AI-assisted Socratic dialogue, delivered through Moodle, may promote conceptual learning in physics. Both the AI-guided and instructor-led groups demonstrated learning gains, but the AI group achieved higher average improvement (0.174 vs. 0.075), with item-level results showing that instructional effectiveness depends on the specific misconception being addressed. Notably, the AI group outperformed on Newton's Third Law questions, while the control group excelled on items related to mass and free fall. These nuanced findings highlight the importance of matching instructional approaches to conceptual challenges. The survey revealed a further layer of complexity. Students in the control group expressed greater satisfaction and perceived understanding, possibly reflecting comfort with traditional, structured instruction. In contrast, the AI group, while reporting lower satisfaction, frequently described experiences of reflection and self-correction. This suggests that the AI assistant fostered "productive struggle"—a process identified in the literature as critical for conceptual change—where students actively grappled with their existing ideas, sometimes experiencing temporary discomfort but ultimately achieving deeper learning. The successful integration of GPT-powered dialogue into a widely used LMS demonstrates that such pedagogical innovations are technically feasible and potentially scalable. However, the mixed student responses emphasize the need for flexible systems that support different learning preferences and promote digital equity. This variability in student experience underscores the importance of designing AI interventions that can be adapted to meet diverse needs, especially as educational

contexts become more heterogeneous. Looking forward, several research and development directions emerge. Larger, multi-site studies across varied school types and languages are essential to test the scalability and generalizability of these findings, while longer-term follow-up will be important to determine if conceptual gains persist over time. Further technical enhancements—such as adaptive prompting based on student profiles and the integration of visuals or interactive simulations—could make AI tutoring even more responsive and accessible. At the same time, exploring how teachers can complement AI feedback, for instance by reviewing student-AI interactions and providing targeted guidance, may bridge the gap between automated support and personalized instruction. Blending AI-mediated dialogue with teachers or peer-led discussion could offer the best of both worlds, support engagement and conceptual development while also meeting students' needs for human interaction. Moreover, the core features of this plugin—misconception tagging, real-time feedback, and Socratic questioning—show promise for adaptation across STEM disciplines, provided subject-specific adjustments are made. Advanced assessment techniques, including open-ended response analysis and adaptive questioning, could deepen our understanding of student thinking and further enhance feedback quality. Ultimately, the path forward requires careful attention to issues of equity and access. As technological instructional systems are deployed more widely, it will be essential to consider technological access, learning differences, language diversity, and cost-effectiveness to ensure that these innovations benefit all learners, not just a privileged few. While much work remains, this pilot study suggests that thoughtfully implemented AI dialogue may play a valuable role in supporting conceptual change, especially when combined with human expertise and adaptive instructional design. As advances in AI and learning analytics continue to converge with educational theory, there is real potential to create more personalized and effective STEM learning environments that empower both students and teachers.

6 DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author(s) used OpenAI's GPT-4.5 to refine and rephrase text for readability. After using this tool, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the publication. The prompts used and the generated paragraphs are available to the editors.

7 REFERENCES

[1] R. Alexander, *Towards Dialogic Teaching: Rethinking Classroom Talk*. Cambridge, UK: Dialogos, 2008.

[2] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998. https://doi.org/10.1080/0969595980050102

[3] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, "Accommodation of a scientific conception: Toward a theory of conceptual change," *Science Education*, vol. 66, no. 2, pp. 211–227, 1982. https://doi.org/10.1002/sce.3730660207

[4] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *The Physics Teacher*, vol. 30, no. 3, pp. 141–158, 1992. https://doi.org/10.1119/1.2343497

[5] I. A. Halloun and D. Hestenes, "The initial knowledge state of college physics students," *American Journal of Physics*, vol. 53, no. 11, pp. 1043–1055, 1985. https://doi.org/10.1119/1.14030

[6] C. Chin, "Teacher questioning in science classrooms: Approaches that stimulate productive thinking," *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, vol. 44, no. 6, pp. 815–843, 2007. https://doi.org/10.1002/tea.20171

[7] M. T. H. Chi, "Active-constructive-interactive: A conceptual framework for differentiating learning activities," *Topics in Cognitive Science*, vol. 1, no. 1, pp. 73–105, 2009. https://doi.org/10.1111/j.1756-8765.2008.01005.x

[8] B. Y. White and J. R. Frederiksen, "Inquiry, modeling, and metacognition: Making science accessible to all students," *Cognition and Instruction*, vol. 16, no. 1, pp. 3–118, 1998. https://doi.org/10.1207/s1532690xci1601_2

[9] S. Freeman *et al.*, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 23, pp. 8410–8415, 2014. https://doi.org/10.1073/pnas.1319030111

[10] H. Khosravi, K. Kitto, and S. Buckingham Shum, "Teaching analytics: Towards actionable learning analytics," *British Journal of Educational Technology*, vol. 53, no. 3, pp. 514–535, 2022. https://doi.org/10.1111/bjet.13107

[11] J. G. Meyer *et al.*, "ChatGPT and large language models in academia: Opportunities and challenges," *BioData Mining*, vol. 16, no. 1, p. 20, 2023. https://doi.org/10.1186/s13040-023-00339-9

[12] E. Kasneci, K. Sessler, and H. Küchenhoff, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[13] B. Gregorcic, G. Polverini, and A. Sarlah, "ChatGPT as a tool for honing teachers' Socratic dialogue skills," *Physics Education*, vol. 59, no. 4, p. 045005, 2024. https://doi.org/10.1088/1361-6552/ad3d21

[14] H. Xie, H.-C. Chu, G.-J. Hwang, and C.-C. Wang, "Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017," *Computers & Education*, vol. 140, p. 103599, 2019. https://doi.org/10.1016/j.compedu.2019.103599

[15] D. Kaleci, "Integration and application of artificial intelligence tools in the Moodle platform: A theoretical exploration," *Journal of Educational Technology and Online Learning*, vol. 8, no. 1, pp. 100–111, 2025. https://doi.org/10.31681/jetol.1595079

[16] S. Zawacki-Richter, M. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1–27, 2019. https://doi.org/10.1186/s41239-019-0171-0

[17] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, no. 1, pp. 64–74, 1998. https://doi.org/10.1119/1.18809

[18] R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir, "Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education," *Information*, vol. 15, no. 10, p. 596, 2024. https://doi.org/10.3390/info15100596

[19] OpenAI, "GPT-4 technical report," 2023. [Online]. Available: https://openai.com/research/gpt-4

[20] OpenAI, "OpenAI API documentation," 2024. [Online]. Available: https://platform.openai.com/docs

[21] P. E. McKnight and J. Najab, "Mann-Whitney U test," *The Corsini Encyclopedia of Psychology*, 2010. https://doi.org/10.1002/9780470479216.corpsy0524

8 APPENDIX

8.1 AI Plugin overview: Configuration, and socratic rules

A complete repository for this study is available at: https://github.com/rabihkahaleh/MoodleAIPlugin/. It includes the plugin's system architecture, module configuration, privacy settings, and Socratic dialogue rules. The AI assistant operates using thread-based memory, structured prompts, and rule-based scaffolding to support misconception-sensitive feedback. Core dialogue strategies such as affirmation, open-ended questioning, and reasoning reflection are documented in full. All relevant data including CSV files containing student answers, quiz questions, mapped misconceptions, and dialogue states are provided to support replication and further research.

9 AUTHORS

Rabih Kahaleh is with the Department of Computer Engineering, University of Balamand, Koura, Lebanon; Àrea de Didàctica de les Ciències Experimentals, Universitat Autònoma de Barcelona, Barcelona, Spain (E-mail: rabih.kahaleh@balamand.edu.lb).

Victor Lopez is with the Àrea de Didàctica de les Ciències Experimentals, Universitat Autònoma de Barcelona, Barcelona, Spain.

Rodrigue Imad is with the Department of Computer Engineering, University of Balamand, Koura, Lebanon; Lab-STICC, UMR CNRS 6285, Brest, France.

Elitza Maneva is with the Àrea de Ciències de la Computació i Intel·ligència Artificial, Computer Engineering, Universitat Autònoma de Barcelona, Barcelona, Spain.