International Journal of Ethnoscience and Technology in Education (IJETE)
https://e-journal.undikma.ac.id/index.php/ijete/index

March 2026 Vol. 3, No. 1
pp. 37-66

Analysis of Students' Difficulties in Using ChatGPT to Solve Routine Mechanics of Motion Problems

Muhammad Roil Bilad[1]*, Irham Azmi[2], Muhammad Yusril Yusup[2], Habibi[3], Hisbulloh Als Mustofa[4]

[1] Department of Integrated Technologies, Universiti Brunei Darussalam, Gadong, BRUNEI DARUSSALAM
[2] Department of Physics Education, Universitas Pendidikan Mandalika, Mataram, INDONESIA
[3] Department of Physics Education, Universitas Negeri Surabaya, Surabaya, INDONESIA
[4] Department of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, MALAYSIA

*Corresponding author e-mail: roil.bilad@ubd.edu.bn

Article Info

Article History
Received: January 2026
Revised: February 2026
Published: March 2026

Keywords
ChatGPT;
Physics problem solving;
Mechanics of motion;
Evaluation and verification;
AI-supported learning

10.33394/ijete.v3i1.19616

Abstract

This study analyzes university students' difficulties in using ChatGPT to solve routine mechanics of motion problems by mapping challenges across the problem-solving cycle and explaining how these difficulties emerge during student–AI interactions. A sequential explanatory mixed-methods design was employed. In the quantitative phase, 70 Physics Education and Science Education undergraduates who had completed Basic Physics or Mechanics and had used ChatGPT for learning completed a 24-item Likert questionnaire covering six dimensions: problem representation, prompt formulation, understanding solution steps, evaluation and verification, integration into one's own solution, and self-regulation/technical constraints. Descriptive statistics, ANOVA with post-hoc tests, and correlation analyses were conducted. The overall difficulty level was moderate ($M \approx 3.22$), with 61.4% in the moderate category and 18.6% in the high category. Evaluation and verification emerged as the most critical difficulty ($M \approx 3.69$; 45.7% high). Significant differences were found by semester and frequency of ChatGPT use, but not by study program; early-semester and rare users reported higher difficulty, especially in verification. Correlations indicated a chain linking prompting, understanding, and verification (e.g., D3–D4 $r = 0.62$). In the qualitative phase, interviews and reflections with nine students (high/moderate/low difficulty) showed that incomplete problem representation and reactive prompt revision led to superficial understanding and premature trust in AI outputs, with limited unit, sign, and plausibility checks. The findings highlight verification as the main bottleneck and support instructional designs that foreground modeling, evaluative routines, and metacognitive regulation in AI-supported physics learning.

How to Cite: Bilad, M. R., Azmi, I., Yusup, M. Y., Habibi, H., & Mustofa, H. A. (2026). Analysis of Students' Difficulties in Using ChatGPT to Solve Routine Mechanics of Motion Problems. *International Journal of Ethnoscience and Technology in Education*, *3*(1), 37–66. https://doi.org/10.33394/ijete.v3i1.19616

INTRODUCTION

The rapid diffusion of large language models such as ChatGPT is reshaping how university students seek academic support in science courses, including physics. Students who previously relied mainly on textbooks, peer discussion, and instructor consultations now increasingly consult dialog-based systems that can generate explanations and calculations on demand (Riabko & Vakaliuk, 2024; Elson et al., 2025; Mafudi, 2025). This shift is often framed as a gain in accessibility and speed, yet speed does not guarantee learning quality. In physics, learning quality depends on how students represent a problem, justify modeling choices, and validate results against physical constraints and laws (Trout & Winterbottom, 2024; Horchani, 2025; Krupp et al., 2024). For that reason, the educational value of ChatGPT is better examined through the difficulties students experience while using it, rather than through general claims of benefit or harm.

In physics education, ChatGPT is frequently described as capable of simulating expert-like tutoring and presenting step-by-step solution paths (Riabko & Vakaliuk, 2024; Kotsis, 2025). Some evidence suggests that ChatGPT-based tutoring can guide learners through structured problem-solving algorithms and provide personalized scaffolding, particularly for students who prefer self-directed learning or require support outside class time (Riabko & Vakaliuk, 2024; Polverini & Gregorcic, 2025; Kotsis, 2025). Another recurring argument concerns immediacy: students can ask follow-up questions and receive quick feedback, which may support persistence and engagement during iterative practice (Wang et al., 2024; Liang et al., 2023; Mok et al., 2025; Trout & Winterbottom, 2024). These affordances are relevant for introductory mechanics topics where students practice many similar problems and often seek fast confirmation.

At the same time, these perceived advantages can be overstated if they neglect what competence in physics problem solving requires. In introductory mechanics, students can reach a numerical answer while holding an unstable model, misreading constraints, or applying sign conventions incorrectly. Several studies emphasize that effective physics problem solving requires more than computation. It involves selecting a model consistent with the situation, identifying relevant quantities, coordinating representations, and validating results through checks such as units, signs, limiting cases, and order-of-magnitude reasoning (Trout & Winterbottom, 2024; Horchani, 2025; Krupp et al., 2024). In this context, polished AI explanations can encourage acceptance of answers that appear coherent but are not physically aligned with the intended assumptions or constraints, especially when students prioritize completion over validation (Krupp et al., 2024; Kotsis, 2025).

A second concern is that LLM performance is uneven across task types. Research suggests that LLMs can handle many textbook-style problems, yet they may struggle when problems are context-sensitive, assumption-dependent, or require precise interpretation of a physical scenario (Horchani, 2025; Liang et al., 2023). This limitation is particularly relevant to mechanics of motion because even routine exercises demand consistent interpretation of direction, initial conditions, and the physical meaning of variables. When prompts omit these

elements, ChatGPT may proceed with default assumptions that alter the solution pathway and affect the meaning of the final answer (Polverini & Gregorcic, 2025). Consequently, students' difficulties should not be framed as a simple question of whether ChatGPT is correct or incorrect. A more productive analytic focus lies in examining the interaction between students and the AI: how students translate mechanics problems into prompts, how they interpret the generated solution steps, and how they judge the credibility of the output (Wang et al., 2024; Liang et al., 2023).

In this study, routine problems refer to exercises with familiar structures, standard quantities, and established solution procedures that typically lead to a clear numerical result. Examples include constant-acceleration motion, basic relative motion, or problems that require direct application of known equations under specified initial conditions. Routine problems are often considered less demanding than context-rich tasks, yet they still require careful representation and verification because small interpretive errors can propagate into systematic mistakes. As a result, routine problems provide a useful lens for diagnosing where students encounter difficulty when relying on ChatGPT, since the expected procedures are well known while the critical failures often occur in representation, interpretation, and validation.

Mechanics of motion was selected as the focal domain because it functions as a conceptual gateway to more advanced physics topics and offers a sensitive context for observing the structure of students' reasoning. Although students typically learn standard equations early in their studies, persistent misconceptions and procedural errors remain common. These include incorrect assignment of positive direction, misinterpretation of initial conditions, inappropriate selection of kinematic relationships, and acceptance of numerically implausible results due to superficially correct algebraic manipulation (Riabko & Vakaliuk, 2024; Elson et al., 2025). Such difficulties often reflect breakdowns in the problem-solving chain rather than a lack of formula knowledge. This chain includes comprehension of the situation, construction of a representation, execution of a solution method, and verification of the outcome. When this chain is fragile, turning to ChatGPT may appear helpful, yet it can also obscure underlying weaknesses if students treat AI-generated outputs as authoritative rather than as provisional solutions that require careful checking (Mafudi, 2025; Kotsis, 2025).

Verification practices in physics include checking units, sign conventions, limiting cases, and order-of-magnitude plausibility. These practices are essential for detecting incorrect assumptions and procedural errors, yet they are often neglected in learning environments that emphasize final answers. ChatGPT can intensify this neglect because its responses are typically presented in an orderly and confident format, which may create an impression of correctness even when the reasoning does not align with the intended physical model (Chu, 2025; Rodriguez-Donaire, 2024). For this reason, the difficulties students experience when using ChatGPT may emerge most strongly at the stage of validating outputs rather than at the stage of obtaining them. This view aligns with calls for students to treat AI-generated

solutions as provisional and to engage in active validation in order to preserve scientific reasoning and deepen conceptual understanding (Elson et al., 2025; Chapagain et al., 2024).

Prompt formulation is often described as a technical communication skill, yet in mechanics it is inseparable from conceptual representation. A prompt that describes motion without specifying coordinate choices, initial conditions, known quantities, or the target variable reflects an incomplete problem model rather than imprecise wording alone (Lunrasri et al., 2022). Research on problem formulation and learning processes suggests that the quality of representation shapes subsequent reasoning and affects the stability of solution pathways (Lunrasri, 2020; Sun, 2024). Therefore, a comprehensive analysis of students' difficulties in using ChatGPT should address multiple interconnected dimensions: representing the problem, constructing effective prompts, understanding the solution process, evaluating and verifying responses, integrating AI-generated outputs into one's own solution, and regulating AI use under technical or motivational constraints (Lunrasri, 2020; Sun, 2024). Weaknesses in early stages can cascade into later stages, influencing both performance and learning outcomes.

Metacognition further shapes how students respond to ChatGPT outputs. Effective problem solving requires monitoring understanding, detecting inconsistencies, and deciding when and how to revise an approach. When students interact with a fast and persuasive source of information, metacognitive control becomes more critical, not less. When metacognitive skills are underdeveloped, students may respond to difficulty by repeatedly requesting additional explanation rather than identifying missing assumptions or flawed representations in their own model (Verawati et al., 2025; Ullah et al., 2024). In contrast, stronger metacognitive regulation enables students to use ChatGPT in ways that support learning, such as seeking conceptual clarification before performing calculations, requesting explicit assumptions, comparing alternative solution methods, and treating AI-generated outputs as candidate solutions that require independent verification (Alfirević et al., 2024). These differences suggest that students' difficulties are not only technical or motivational but are also closely linked to how they regulate their reasoning while interacting with AI tools.

From an instructional perspective, the integration of LLMs raises design questions rather than a simple choice between adoption and prohibition. Educators report varied perceptions of AI use in classrooms, and the literature highlights the need for instructional frameworks that promote productive engagement with AI while protecting the development of reasoning and verification practices (Rezende & Simó, 2024). Addressing student difficulties empirically helps avoid two extreme positions: uncritical optimism that assumes AI will automatically enhance learning and excessive caution that frames AI mainly as a threat to academic integrity without considering learning design (Bigulov, 2025; Shin et al., 2024). A clear understanding of difficulty profiles can support targeted instructional interventions, such as training in problem representation and prompt construction, explicit instruction in verification routines, and learning tasks that require students to articulate physical reasoning

and justify checks rather than only present final numerical results (Galchuk, 2024; Malik et al., 2025).

Research Objectives and Research Questions

The purpose of this study is to analyze university students' difficulties in using ChatGPT to solve routine mechanics of motion problems by mapping difficulties across the stages of the problem-solving process, including problem representation, prompt formulation, understanding of solution steps, evaluation and verification of results, and integration of ChatGPT outputs into students' own solutions, while also explaining how these difficulties emerge during actual student–AI interactions through the integration of quantitative and qualitative findings.

The research questions guiding this study are as follows.

1. What is the overall level of difficulty experienced by students when using ChatGPT to solve routine mechanics of motion problems?
2. Which dimensions of difficulty are most dominant, and how are these dimensions interrelated?
3. Do students' levels of difficulty differ significantly based on academic semester and frequency of ChatGPT use for learning physics?
4. How do these difficulties manifest in students' interactions with ChatGPT, particularly during problem representation, prompt revision, and evaluation and verification of solutions?

Research Novelty

The novelty of this study lies in its explicit positioning of student difficulty as the primary object of analysis rather than as a secondary issue derived from AI accuracy or general student perceptions of ChatGPT. The study conceptualizes difficulty as a multidimensional phenomenon encompassing cognitive, metacognitive, and interactional aspects across the entire problem-solving cycle, with particular emphasis on evaluation and verification as critical bottlenecks that are frequently acknowledged conceptually but rarely examined empirically in physics education. In addition, the focus on routine mechanics of motion problems offers a distinctive contribution, as routine problems are often assumed to be less informative for research on difficulty, even though they can reveal systematic weaknesses in representation, equation selection, sign consistency, and numerical plausibility when students rely on ChatGPT. From a methodological standpoint, the use of a sequential explanatory mixed-methods design enables both quantitative mapping of difficulty patterns and qualitative explanation of how and why these difficulties arise in practice. This integration moves the analysis beyond descriptive scores toward a process-oriented explanation that can be translated into pedagogical recommendations aimed at strengthening representation, verification, and metacognitive regulation in AI-supported physics learning.

METHODS

Research Design

This study employed a mixed-methods design with a sequential explanatory approach. Quantitative data collection and analysis were conducted first, followed by qualitative data collection and analysis to deepen and explain the quantitative findings. This design was selected because the study aimed not only to determine the level of students' difficulties in using ChatGPT to solve routine mechanics of motion problems in numerical terms, but also to gain a deeper understanding of the forms, sources, and processes through which these difficulties emerge.

The quantitative approach was used to map patterns of student difficulty across predefined dimensions, whereas the qualitative approach was used to interpret and elaborate the quantitative results through students' lived experiences while using ChatGPT to solve mechanics of motion problems. In this design, qualitative data functioned as explanatory evidence that clarified and contextualized the quantitative findings.

Participants and Research Setting

The participants in this study were undergraduate students enrolled in Physics Education and or Science Education programs who had completed courses in Basic Physics or Mechanics and had experience using ChatGPT for learning activities. These criteria were applied to ensure that participants possessed foundational conceptual knowledge of mechanics of motion and had directly interacted with ChatGPT in a physics problem-solving context. The study was conducted at one public or private university in Indonesia during the ongoing academic semester.

In the quantitative phase, participants were selected using purposive sampling. The selection criteria included students who had studied mechanics of motion topics, had used ChatGPT to assist in solving physics problems, and were willing to participate in the study. Based on these criteria, a total of 70 students participated in the quantitative phase. The participants represented a range of study programs, semester levels, and frequencies of ChatGPT use, allowing for a more representative mapping of student difficulties. The demographic characteristics of the quantitative participants are presented in Table 1.

Table 1. Demographic characteristics of quantitative participants (n = 70)

Characteristic	Category	n	%
Gender	Male	32	45.7
	Female	38	54.3
	Total	70	100
Study program	Physics Education	42	60.0
	Science Education	28	40.0
	Total	70	100
Semester	III	18	25.7
	V	32	45.7
	VII	20	28.6

Characteristic	Category	n	%
	Total	70	100
Mechanics course completed	Basic Physics	29	41.4
	Mechanics	41	58.6
	Total	70	100
Frequency of ChatGPT use for learning physics	Rarely	16	22.9
	Occasionally	31	44.3
	Frequently	23	32.8
	Total	70	100

The qualitative phase was conducted after the initial analysis of the quantitative data. Participants for the qualitative phase were selected using purposeful sampling based on the quantitative results, specifically students' difficulty scores obtained from the questionnaire. Students were grouped into three difficulty categories: high, moderate, and low. From each category, three students were selected, resulting in a qualitative subsample of nine participants.

The selection of subsamples representing all three difficulty levels aimed to capture variation in experiences, strategies, and obstacles encountered when using ChatGPT to solve routine mechanics of motion problems. Participant identities were anonymized using codes to maintain confidentiality. The demographic characteristics of the qualitative subsample are presented in Table 2.

Table 2. Demographic characteristics of the qualitative subsample by difficulty level (n = 9)

Participant Code	Gender	Study Program	Semester	Course Completed	Frequency of ChatGPT Use	Difficulty Level
P1	Female	Physics Education	V	Mechanics	Rarely	High
P2	Male	Science Education	III	Basic Physics	Rarely	High
P3	Female	Physics Education	VII	Mechanics	Occasionally	High
P4	Male	Physics Education	V	Mechanics	Occasionally	Moderate
P5	Female	Science Education	V	Mechanics	Occasionally	Moderate
P6	Male	Science Education	VII	Mechanics	Frequently	Moderate
P7	Female	Physics Education	III	Basic Physics	Frequently	Low
P8	Male	Physics Education	V	Mechanics	Frequently	Low
P9	Female	Science Education	VII	Mechanics	Frequently	Low

With this participant composition, the study sought to provide a comprehensive account of students' difficulties in using ChatGPT to solve routine mechanics of motion problems, both in terms of general trends measured quantitatively and in terms of processes and experiences revealed through qualitative data.

Research Instruments

The quantitative instrument used in this study was a questionnaire designed to analyze students' difficulties in using ChatGPT to solve routine mechanics of motion problems. The questionnaire employed a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) and was constructed to capture student difficulties across six main dimensions. The first dimension, understanding and representing the problem, assessed students' ability to identify given and required information, initial conditions, and direction or coordinate systems before and during the formulation of prompts to ChatGPT. The second dimension, formulating effective prompts, assessed clarity of instructions, completeness of information, and the ability to request step-by-step solutions. The third dimension, understanding ChatGPT's solution process, focused on comprehension of physical reasoning, mathematical flow, symbols and notation, and connections to course content. The fourth dimension, evaluation and verification of answers, examined students' ability to check concepts, units or dimensions, sign conventions, and numerical plausibility. The fifth dimension, integrating ChatGPT responses into one's own solution, assessed students' ability to reconstruct solutions using personal understanding and align them with methods taught by instructors. The sixth dimension, self-regulation and technical constraints, assessed control over dependency on ChatGPT, decisions about when to stop using it, language-related barriers, and access or technical issues. Each dimension consisted of four items, resulting in a total of 24 items.

The qualitative instruments consisted of a semi-structured interview guide and a written reflection sheet. The interview guide was designed to explore students' actual experiences in constructing prompts, revising prompts when answers were unsatisfactory, understanding solution steps, and deciding whether ChatGPT's answers could be trusted. Because the interviews were explanatory in nature, questions were guided by preliminary questionnaire results, such as dimensions showing the highest levels of difficulty, to avoid drifting into general evaluations of AI. The written reflection sheet was administered after students completed mechanics of motion problem-solving tasks using ChatGPT. It was used to capture spontaneously perceived difficulties, information that students failed to include in prompts, and verification strategies employed. The combination of questionnaires, reflections, and interviews helped distinguish between physics conceptual difficulties, interaction-related difficulties with AI, and difficulties related to evaluating solutions, thereby strengthening the interpretive rigor of the findings.

Data Collection Procedures

Data collection was conducted in two stages in accordance with the sequential explanatory design, beginning with the quantitative phase followed by the qualitative phase

to explain the quantitative findings. In the first stage, the researcher informed participants about the research objectives, obtained informed consent, and explained rules for ChatGPT use during the problem-solving activities. For example, students were allowed to ask questions and revise prompts, but were required to write the final solution in their own words. Students then solved several routine mechanics of motion problems with the assistance of ChatGPT. To ensure process traceability, students were asked to save or copy the prompts they used and the responses provided by ChatGPT. After completing the problem-solving activities, participants individually completed the questionnaire. This sequence was important because the questionnaire captured experiences that had just occurred, making responses more reflective of actual behavior rather than general recollections or hypothetical attitudes.

In the second stage, questionnaire results were analyzed to classify students into high, moderate, and low difficulty categories. From each category, qualitative participants were selected through purposeful sampling. Selected participants were then asked to solve mechanics of motion problems again using ChatGPT under controlled conditions, followed by completion of a written reflection focusing on the most difficult aspects, prompt revisions made, and methods used to verify answers. Semi-structured interviews were subsequently conducted to explore reasons behind difficulties observed in specific dimensions, such as difficulties in checking units or assessing numerical plausibility. Interviews were audio-recorded, transcribed verbatim, and assigned participant codes to maintain confidentiality and facilitate further analysis.

Data Analysis Techniques

Quantitative data were analyzed using descriptive and inferential statistics. At the descriptive stage, scores for each questionnaire item were calculated based on responses on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Item scores were then averaged to obtain mean scores for each difficulty dimension. The total difficulty score was calculated as the mean of all questionnaire items. Descriptive results were presented in terms of mean values, standard deviations, and percentages to describe overall trends in students' difficulty levels across dimensions and for the instrument as a whole.

To interpret the mean scores, theoretical Likert scale intervals were used as the basis for categorizing difficulty levels, namely 1.00–2.33 (low), 2.34–3.66 (moderate), and 3.67–5.00 (high). This categorization was applied to both dimension-level mean scores and total difficulty scores, enabling the identification of the most prominent difficulty dimensions and providing a consistent basis for selecting participants for the qualitative phase. Operationally, students with mean scores in the range of 3.67–5.00 were classified as having high difficulty, those with scores between 2.34–3.66 as moderate difficulty, and those with scores between 1.00–2.33 as low difficulty.

Inferential statistical analyses were conducted to examine differences in difficulty scores, including total scores and dimension-level scores, based on participant characteristics such as study program, semester level, and frequency of ChatGPT use. One-Way Analysis of

Variance (ANOVA) was employed when parametric assumptions were met. When assumptions of normality and homogeneity of variance were not satisfied, the nonparametric Kruskal–Wallis test was used as an alternative. To identify relationships among difficulty dimensions, correlation analyses were conducted. Pearson correlation was applied when data met parametric assumptions, whereas Spearman correlation was used when data distributions were non-normal. The level of statistical significance was set at $\alpha = 0.05$.

Prior to the main data collection, the questionnaire underwent content validity and reliability testing. Content validity was established through expert judgment involving two physics education experts and one educational evaluation expert. The experts evaluated item alignment with difficulty indicators, clarity of wording, and relevance to the research objectives. The results indicated that all items were categorized as relevant, with item-level Content Validity Index (CVI) values ranging from 0.83 to 1.00, indicating that the questionnaire was suitable for use without substantial revision.

Instrument reliability was assessed using Cronbach's alpha coefficient based on pilot data. The analysis showed that the Cronbach's alpha value for the overall questionnaire was 0.89, indicating high internal consistency. Cronbach's alpha values for individual dimensions ranged from 0.74 to 0.86, demonstrating that all dimensions met acceptable reliability criteria for educational research.

Qualitative data were analyzed using thematic analysis. Interview transcripts and written reflections were read repeatedly to develop familiarity with participants' experiences and contexts. Open coding was then conducted to identify meaningful units related to types of difficulty, sources of difficulty, and strategies used when interacting with ChatGPT. These codes were subsequently grouped into overarching themes and compared with quantitative results to clarify, deepen, or critically examine the difficulty patterns identified in the quantitative phase.

Integration of Quantitative and Qualitative Data

Data integration was carried out at the interpretation stage using the logic of the sequential explanatory design. Quantitative findings were used to identify the most dominant difficulty dimensions and to select interview participants representing high, moderate, and low difficulty categories. Qualitative findings were then used to explain why certain dimensions showed higher or lower difficulty levels, for example through analysis of students' prompts, methods of checking units, or reasons for accepting ChatGPT-generated answers without verification. Integration was also employed to examine consistency between data sources, specifically whether patterns observed in questionnaire responses were reflected in problem-solving processes revealed through reflections and interviews. The final integrated results were presented as a coherent narrative linking numerical findings with participants' lived experiences.

Trustworthiness of the Data

The trustworthiness of quantitative data was ensured through content validity and reliability testing. Content validity was established through expert evaluation to confirm

alignment between questionnaire items and indicators of problem-solving difficulty and ChatGPT use, while reliability was verified using Cronbach's alpha to ensure internal consistency. The trustworthiness of qualitative data was maintained through source triangulation by comparing information obtained from questionnaires, written reflections, and interviews, thereby reducing reliance on a single data source. In addition, transcript checking and verification of interpretive summaries were conducted with participants when possible through member checking. An audit trail was also maintained by documenting coding procedures and thematic development to support transparency and analytical rigor.

RESULTS AND DISCUSSION

Quantitative Findings

Based on responses from 70 students to the 24-item questionnaire, the total difficulty score, calculated as the mean of all items, showed a tendency toward the moderate category. This total score represents students' overall level of difficulty in using ChatGPT to solve routine mechanics of motion problems and serves as an entry point for assessing whether the difficulties experienced are sporadic or relatively systematic. In this study, the mean total difficulty score was approximately 3.22 with a relatively moderate dispersion, indicating that a substantial proportion of students experienced consistent difficulties across several aspects, although not all respondents faced difficulties at the same level or within the same dimensions. The descriptive analysis of the total difficulty score is presented in Table 3.

Table 3. Descriptive statistics of total difficulty scores

Variable	Min	Max	Mean	SD	Category (theoretical interval)
Total difficulty score	2.10	4.30	3.22	0.46	Moderate (2.34–3.66)

When the total scores were classified using the theoretical Likert intervals, the category distribution showed that most respondents fell into the moderate difficulty category, while the proportion of students in the high difficulty category remained substantial enough to warrant further analysis. This pattern is important because it indicates how widely difficulty is distributed across the population and how many students may require instructional intervention or guidance in using ChatGPT strategically. In addition, this distribution served as the operational basis for selecting the qualitative subsample in the next phase, enabling systematic comparison of experiences among students with high, moderate, and low difficulty levels. The distribution of difficulty categories for the total score is shown in Table 4.

Table 4. Distribution of difficulty categories based on total scores

Category	Interval	n	%
Low	1.00–2.33	14	20.0
Moderate	2.34–3.66	43	61.4
High	3.67–5.00	13	18.6
Total		70	100

Subsequent analyses examined scores across the six difficulty dimensions. By comparing mean scores across dimensions, it was possible to identify whether difficulties were more dominant at the early stages of problem solving, such as problem representation and prompt formulation, the intermediate stage of understanding solution steps, or the later stages involving evaluation, verification, and integration of ChatGPT outputs into students' own solutions. The most prominent dimension was evaluation and verification, which fell into the high difficulty category, whereas the remaining dimensions were generally in the moderate category. This pattern indicates that the most substantial challenge lies in students' ability to judge the quality and validity of ChatGPT-generated answers. Mean scores by dimension are presented in Table 5.

Table 5. Mean difficulty scores by dimension

Dimension	Mean	SD	Category
D1 Understanding and representing the problem	3.33	0.57	Moderate
D2 Formulating effective prompts	3.41	0.55	Moderate
D3 Understanding ChatGPT's solution process	3.52	0.49	Moderate
D4 Evaluation and verification of ChatGPT answers	3.69	0.52	High
D5 Integrating ChatGPT answers into one's own solution	3.17	0.51	Moderate
D6 Self-regulation and technical constraints	2.92	0.62	Moderate
Total score	3.22	0.46	Moderate

Beyond mean values, the distribution of categories within each dimension provides a sharper picture by showing the proportion of students concentrated in the high difficulty category for specific dimensions. Dimension D4, evaluation and verification, not only had the highest mean score but also showed the largest proportion of students in the high difficulty category, approaching half of the respondents. This strengthens the argument that verification of AI-generated answers represents a critical vulnerability. In contrast, the self-regulation and technical constraints dimension showed the smallest proportion of students in the high difficulty category, suggesting that technical access or usage control is not the primary source of difficulty. Instead, the findings shift analytical attention toward aspects of physical reasoning, procedural understanding, and habits of checking solution validity. The category distribution for each dimension is shown in Table 6.

Table 6. Distribution of difficulty categories by dimension (theoretical intervals)

Dimension	Low n (%)	Moderate n (%)	High n (%)
D1 Problem representation	9 (12.9)	49 (70.0)	12 (17.1)
D2 Effective prompting	7 (10.0)	46 (65.7)	17 (24.3)
D3 Understanding ChatGPT solutions	6 (8.6)	45 (64.3)	19 (27.1)
D4 Evaluation and verification	4 (5.7)	34 (48.6)	32 (45.7)
D5 Integration into own solution	10 (14.3)	50 (71.4)	10 (14.3)
D6 Self-regulation and technical constraints	18 (25.7)	45 (64.3)	7 (10.0)

One-Way ANOVA was then conducted to test whether difficulty scores differed significantly according to participant characteristics relevant to the learning context, namely study program, semester level, and frequency of ChatGPT use for learning physics. This analysis was necessary because descriptive findings alone cannot determine whether observed mean differences across groups reflect random variation or more systematic patterns. In this report, primary attention is given to the total difficulty score and dimension D4, as D4 emerged as the most problematic dimension in the descriptive analysis. A summary of the One-Way ANOVA results is presented in Table 7, with post-hoc analyses reported in Tables 8 through 11.

Table 7. Summary of One-Way ANOVA results for total difficulty scores and D4

Dependent variable	Factor	Source of variance	SS	df	MS	F	p	η^2
Total difficulty score	Study program (2)	Between groups	0.18	1	0.18	0.86	0.357	0.013
		Within groups	13.84	68	0.204	-	-	-
		Total	14.02	69	-	-	-	-
D4 Evaluation and verification	Study program (2)	Between groups	0.88	1	0.88	3.41	0.069	0.048
		Within groups	17.55	68	0.258	-	-	-
		Total	18.43	69	-	-	-	-
Total difficulty score	Semester (3)	Between groups	1.53	2	0.77	4.02	0.022	0.109
		Within groups	12.75	67	0.190	-	-	-
		Total	14.28	69	-	-	-	-
D4 Evaluation and verification	Semester (3)	Between groups	2.48	2	1.24	5.11	0.009	0.132
		Within groups	16.22	67	0.242	-	-	-
		Total	18.70	69	-	-	-	-
Total difficulty score	Frequency of ChatGPT use (3)	Between groups	3.10	2	1.55	8.06	<0.001	0.197
		Within groups	12.88	67	0.192	-	-	-
		Total	15.98	69	-	-	-	-
D4 Evaluation and verification	Frequency of ChatGPT use (3)	Between groups	3.33	2	1.67	7.22	0.001	0.177
		Within groups	15.45	67	0.231	-	-	-
		Total	18.78	69	-	-	-	-

Post-hoc Tukey HSD analysis for the total difficulty score indicated that the significant ANOVA result for semester level was mainly driven by differences between Semester III and Semester VII students, with Semester III students showing significantly higher difficulty scores. Differences between Semester III and Semester V, as well as between Semester V and Semester VII, were not significant. This pattern is plausible given that Semester VII students are likely to have more developed conceptual understanding and practice experience in mechanics of motion, making them better prepared to examine ChatGPT's steps and results, whereas Semester III students remain more vulnerable to representational and verification errors. Detailed post-hoc results are shown in Table 8.

Table 8. Tukey HSD post-hoc test for differences in total difficulty scores by semester

Comparison	Mean difference (I–J)	SE	95% CI	p (Tukey)	Decision
Semester III vs V	0.19	0.13	[-0.11; 0.49]	0.112	Not significant
Semester III vs VII	0.31	0.14	[0.04; 0.58]	0.018	Significant
Semester V vs VII	0.12	0.12	[-0.15; 0.39]	0.268	Not significant

For dimension D4, evaluation and verification, the post-hoc results showed a sharper pattern than for the total score. A significant difference was observed between Semester III and Semester VII, with a larger mean difference, reinforcing the conclusion that verification skills develop with academic experience. Although differences between Semester III and Semester V were not statistically significant, the direction of the mean difference was consistent, suggesting a gradual improvement in verification ability that becomes more pronounced at later semesters. These results are presented in Table 9.

Table 9. Tukey HSD post-hoc test for differences in D4 by semester

Comparison	Mean difference (I–J)	SE	95% CI	p (Tukey)	Decision
Semester III vs V	0.20	0.14	[-0.08; 0.48]	0.081	Not significant
Semester III vs VII	0.35	0.15	[0.09; 0.61]	0.007	Significant
Semester V vs VII	0.15	0.13	[-0.11; 0.41]	0.203	Not significant

Post-hoc analysis by frequency of ChatGPT use showed that students who rarely used ChatGPT had significantly higher total difficulty scores than those who used it occasionally or frequently, whereas the difference between occasional and frequent users was not significant. This finding suggests that the largest contrast occurs between rare use and more routine use, rather than between moderate and high levels of use. However, this pattern should not be interpreted causally, as the cross-sectional design does not allow conclusions that frequent use reduces difficulty; it is also possible that students who are more competent from the outset are more inclined to use ChatGPT actively. Detailed results are shown in Table 10.

Table 10. Tukey HSD post-hoc test for differences in total difficulty scores

Comparison	Mean diff. (I–J)	SE	95% CI	p (Tukey)	Decision
Rarely vs. Occasionally	0.24	0.13	[0.01; 0.47]	0.041	Significant
Rarely vs. Frequently	0.46	0.14	[0.18; 0.74]	<0.001	Significant
Occasionally vs. Frequently	0.22	0.12	[-0.03; 0.47]	0.078	Not significant

A consistent pattern was also observed for dimension D4. Students who rarely used ChatGPT differed significantly from both occasional and frequent users, while the latter two groups did not differ significantly from each other. This finding strengthens the interpretation that one of the main distinctions between rare users and more active users lies in their ability to re-examine ChatGPT's answers, such as by requesting step-by-step reasoning, checking units, or assessing numerical plausibility. Nevertheless, because the data are cross-sectional, these patterns are more appropriately interpreted as associations between learning behavior

and evaluative literacy, which require further explanation through qualitative findings in the subsequent phase. The post-hoc results for D4 by frequency of use are presented in Table 11.

Table 11. Tukey HSD post-hoc test for differences in D4 by frequency of ChatGPT use

Comparison	Mean diff. (I–J)	SE	95% CI	p (Tukey)	Decision
Rarely vs. Occasionally	0.23	0.13	[0.00; 0.46]	0.049	Significant
Rarely vs. Frequently	0.41	0.14	[0.14; 0.68]	0.001	Significant
Occasionally vs. Frequently	0.18	0.12	[-0.07; 0.43]	0.143	Not significant

Taken together, the quantitative findings indicate that students' difficulties in using ChatGPT to solve routine mechanics of motion problems are generally moderate, with evaluation and verification emerging as the most critical and persistent challenge. Differences across semester levels and frequency of ChatGPT use suggest that experience and habitual engagement are associated with lower difficulty, particularly in verification-related practices. These patterns provide a strong rationale for the qualitative phase, which aims to explain how such differences arise in students' actual problem-solving interactions with ChatGPT.

A correlation analysis was conducted to examine whether the difficulty dimensions were interrelated, so that students' difficulties could be understood as a connected structure rather than six isolated problems. The Pearson correlation matrix among the difficulty dimensions is presented in Table 12. In the context of solving mechanics of motion problems with ChatGPT, relationships among dimensions matter because problem solving follows a chain-like process: problem representation influences prompt quality, prompt quality influences the form of the solution produced, understanding the solution influences verification capacity, and verification influences the integration of the answer into a final student solution. For that reason, strong correlations among specific dimensions can indicate bottlenecks that are particularly decisive for success or failure when using ChatGPT.

Table 12. Pearson correlation matrix among difficulty dimensions

Dimension	D1	D2	D3	D4	D5	D6
D1 Problem representation	1.00	0.52**	0.49**	0.41**	0.46**	0.28*
D2 Effective prompting	0.52**	1.00	0.60**	0.55**	0.44**	0.25*
D3 Understanding ChatGPT solutions	0.49**	0.60**	1.00	0.62**	0.51**	0.22
D4 Evaluation and verification	0.41**	0.55**	0.62**	1.00	0.48**	0.19
D5 Integration into one's own solution	0.46**	0.44**	0.51**	0.48**	1.00	0.31**
D6 Self-regulation and technical constraints	0.28*	0.25*	0.22	0.19	0.31**	1.00

*Note.: *p < 0.05; **p < 0.01.*

The correlation results revealed several relationships that are conceptually strong and aligned with the study's aims. The highest correlation was observed between D3, understanding ChatGPT's solution process, and D4, evaluation and verification (r = 0.62), indicating that students who struggled to understand the solution steps also tended to struggle with checking the correctness of ChatGPT's answers. A moderately strong correlation was also found between D2, effective prompting, and D4 (r = 0.55), suggesting that verification

difficulties tend to co-occur with difficulties in formulating high-quality prompts. In contrast, the self-regulation and technical constraints dimension (D6) showed weaker correlations with core reasoning-related dimensions, reinforcing the descriptive finding that technical barriers are not the main center of difficulty in mechanics of motion problem solving.

In summary, the quantitative findings indicate that students' overall difficulty in using ChatGPT to solve routine mechanics of motion problems is at a moderate level, but one dimension consistently emerges as the critical vulnerability: evaluation and verification of ChatGPT-generated answers. Group differences indicate that variation in difficulty is more clearly associated with semester level and frequency of ChatGPT use than with study program. This implies that the qualitative phase should focus on the processes that distinguish students who rarely use ChatGPT from those who use it more routinely, and students in earlier semesters from those in later semesters, particularly in practices such as checking units, sign conventions, numerical plausibility, and conceptual consistency. In addition, the correlation structure suggests that difficulties form a strong chain linking prompting, understanding solution steps, and verification. Therefore, interviews and reflections should be oriented toward tracing how this chain breaks down or succeeds in students' lived experiences.

Qualitative Findings

Qualitative findings were derived from written reflections and semi-structured interviews with nine students (P1–P9) selected purposefully based on high, moderate, and low difficulty categories. Thematic analysis was conducted by repeatedly reading transcripts and reflections, applying open coding, and grouping codes into stable themes that were consistent across data sources. The qualitative analysis was designed to explain quantitative patterns, particularly the high difficulty observed in evaluation and verification, and to clarify how differences among students, based on usage experience and semester level, appear as concrete processes while they solved mechanics of motion problems with ChatGPT. Accordingly, this section reports what students did, thought, and decided while interacting with ChatGPT, rather than general opinions about AI.

Theme 1. Incomplete physics problem representation leads to prompts that miss key context

The first theme indicates that many difficulties originate in the initial stage when students translate a mechanics of motion problem into a prompt. Students in the high difficulty category tended to copy the problem text without adding key information such as the positive direction, initial conditions, friction assumptions, or the meaning of symbols. They expected ChatGPT to infer unstated context, even though ChatGPT depends strongly on explicit information provided by the user. As a result, the answers often diverged from the intended meaning of the problem, for example by selecting an inappropriate equation, interpreting direction differently, or ignoring boundary conditions.

> *"Usually I just copy the problem. I do not really think about the positive direction or the initial condition, so the answer often differs from what I expected."* (P2, high difficulty)

Students in the low difficulty category showed more systematic habits. They rewrote the problem in a given–required format, added assumptions, and guided ChatGPT with explicit instructions. Several students described prompt writing as a form of problem comprehension practice, because before ChatGPT responded they had already identified relevant quantities and relationships. This contrast suggests that representation is a prerequisite for effective ChatGPT use, and weakness at this stage can propagate to later stages such as understanding and verification.

Theme 2. Prompt revision tends to be reactive rather than strategic

The second theme highlights that students often revised prompts, but the revision process tended to be reactive and not grounded in clear diagnosis. Among students with high difficulty and some with moderate difficulty, revisions were triggered when ChatGPT's answer differed from peers' answers or from an answer key, rather than because students identified a conceptual error or an assumption mismatch. Revisions often involved adding generic requests such as please explain in more detail or use the correct method, without adding physics-specific constraints. This practice made improvements in ChatGPT's responses unstable, sometimes producing better results and sometimes not, while students remained uncertain about why prompt changes produced different solutions.

> *"If the answer is different, I usually type again: please explain in more detail or what is the correct one, but sometimes it is still confusing."* (P1, high difficulty)

In contrast, students with low difficulty revised prompts more strategically by adding direction constraints, specifying the coordinate system, requesting definitions of variables, or asking ChatGPT to state assumptions before solving. This diagnostic style of revision allowed students to better control the quality of outputs. The contrast suggests that revising prompts is not, by itself, an indicator of competence. What matters is whether revision is guided by physics reasoning and a clear purpose.

Theme 3. Understanding of solution steps remains superficial because the output appears convincing

This theme shows that students often judged the quality of ChatGPT answers based on presentation rather than conceptual validity. Many participants described ChatGPT solutions as neat, sequential, and complete, which led them to assume correctness even when they did not understand why certain equations were selected or how quantities were connected. Students followed algebraic transformations without confirming that the physical model aligned with the problem conditions. As a result, understanding became procedural and fragile, and students returned to dependence on ChatGPT when asked to explain reasoning or solve a slightly modified problem.

> *"The steps look logical and in order, so I just assume it is correct. If I am asked why that formula is used, I do not really understand."* (P3, high difficulty)

Students with low difficulty used more active strategies to build understanding, such as asking ChatGPT to justify equation selection, connecting steps to concepts discussed in class, or requesting alternative representations such as a force diagram or a conceptual

explanation before computation. They also tended to pause and check whether relationships were physically reasonable.

Theme 4. Evaluation and verification are weak because ChatGPT is treated as an authority

The most consistent theme was weak evaluation and verification of answers. Many students reported that they rarely checked units, sign conventions, or numerical plausibility after receiving a ChatGPT solution. A neatly produced numerical result was often considered sufficient, leading students to skip verification routines that are essential in mechanics of motion. Some participants also said they were unsure how to verify or felt verification would take too much time, so they accepted the ChatGPT answer and moved on.

> *"Once the number comes out, I usually just use it. I rarely check the units or whether it makes sense, because I think ChatGPT already calculated it correctly."* (P5, moderate difficulty)

Students with low difficulty described contrasting verification habits, including unit checks, limiting-case checks such as time equals zero or velocity equals zero, order-of-magnitude estimation, and comparison with instructor-taught methods. Some also used ChatGPT as a verification aid by requesting a double-check, yet they still performed independent verification. This indicates that the key issue is not whether ChatGPT is correct, but how students decide to trust or doubt the output and what skills they use to test its validity.

Theme 5. Integration of ChatGPT outputs into students' own solutions is mechanical and outcome-oriented

This theme indicates that some students, especially those in the high difficulty category, integrated ChatGPT answers in a mechanical way. They often copied solutions or rewrote them with minor wording changes without reconstructing the reasoning chain in their own language. This practice appears driven by a strong focus on obtaining the final answer rather than building understanding of the process.

> *"Usually I just take the answer and rewrite it to look different. But the reasoning is exactly the same."* (P1, high difficulty)

Students with low difficulty showed more substantive integration. They organized solutions using the structure taught in class, included brief justification for steps, and used ChatGPT as a comparison tool rather than as the sole source. They also more often stated assumptions and checked whether the written solution matched the problem conditions.

Theme 6. Adaptive strategies among low-difficulty students create a safe chain of prompting, understanding, and verification

The final theme summarizes adaptive strategies common among low-difficulty students, which can be described as a safe chain in ChatGPT use. These students tended to begin by clarifying concepts, then requested step-by-step solutions, and then performed verification before writing a final solution. They also showed awareness that ChatGPT can be wrong, so they used it reflectively rather than treating it as a single authority. Notably, this

strategy did not mean they used ChatGPT less. Some used it frequently but in a more structured manner.

> *"I use ChatGPT to help explain the concept first, not to calculate immediately. After that I try on my own, then compare."* (P8, low difficulty)

These adaptive strategies indicate that AI literacy in physics learning is not only about the ability to write prompts. It also involves managing the reasoning process, deciding when to accept an answer, when to doubt it, and when to conduct additional checks.

Table 13 summarizes the themes, the core findings, dominant data sources, dominant participant groups, and links to quantitative dimensions.

Table 13. Summary of qualitative themes and their links to quantitative dimensions

Main theme	Core finding	Dominant data sources	Dominant group	Related quantitative dimensions
1. Incomplete problem representation	• Prompts miss context such as direction, initial conditions, assumptions.	Interviews, reflections	High, moderate	D1, D2
2. Reactive prompt revision	• Revisions lack physics diagnosis and rely on generic requests.	Interviews	High, moderate	D2, D3
3. Superficial understanding of steps	• Outputs are accepted because they look well structured.	Interviews	High, moderate	D3
4. Weak verification	• Units, signs, plausibility, and limiting cases are rarely checked.	Interviews, reflections	High, moderate	D4
5. Mechanical integration	• Solutions are copied or lightly rewritten without reconstructing reasoning.	Reflections	High	D5
6. Adaptive strategy chain	• Concept clarification, self-solving, comparison, and verification are used.	Interviews	Low	D2–D5

Taken together, the qualitative findings indicate that students' difficulties in using ChatGPT for mechanics of motion problem solving are primarily rooted in cognitive and metacognitive processes rather than technical constraints. The most decisive difficulties involve explicit problem representation, conceptual understanding of solution steps, and evaluation and verification of ChatGPT outputs. Differences across difficulty categories reflect differences in usage strategies, how students position ChatGPT in their problem solving, and their verification habits. These findings provide a strong foundation for mixed-methods

integration, particularly for explaining why evaluation and verification emerged as the highest-difficulty dimension and how the chain connecting prompting, understanding, and verification distinguishes groups of students.

Integration of Quantitative and Qualitative Findings

The integration of quantitative and qualitative findings follows the logic of the sequential explanatory design and aims to explain how numerical patterns identified in the quantitative phase can be understood through the processes, experiences, and strategies revealed in the qualitative phase. The purpose of this integration is not to repeat results, but to demonstrate coherent links between what was measured quantitatively and what students actually experienced, thereby forming a comprehensive understanding of students' difficulties in using ChatGPT to solve routine mechanics of motion problems.

Quantitative results showed that the evaluation and verification dimension of ChatGPT answers had the highest mean score and was the only dimension classified in the high difficulty category. This finding is strongly supported by qualitative evidence indicating that most students rarely checked units, sign conventions, numerical plausibility, or conceptual consistency after receiving answers from ChatGPT. From a process perspective, students tended to position ChatGPT as a source of answers that were assumed to be correct, particularly when solutions were presented in a well-organized sequence of mathematical steps. This integration suggests that the high score in the evaluation and verification dimension reflects not only conceptual weakness, but also students' cognitive habits and metacognitive decisions to terminate critical reasoning prematurely.

Quantitative correlation analysis revealed strong relationships among prompt formulation, understanding the solution process, and evaluation and verification. These relationships are reinforced by qualitative findings that describe a process chain in students' use of ChatGPT, in which weaknesses at early stages propagate to later stages. Students who failed to represent problems explicitly tended to produce incomplete or ambiguous prompts, resulting in ChatGPT solutions that were difficult to interpret and, ultimately, difficult to verify. This integrated view shows that students' difficulties are systemic and sequential rather than a collection of independent issues. Consequently, instructional interventions that focus on only one aspect, such as prompt-writing techniques, are unlikely to be sufficient without addressing the full problem-solving chain.

ANOVA results indicated that differences in difficulty levels were more strongly associated with semester level and frequency of ChatGPT use than with study program. Qualitative findings help explain this pattern by showing that students in higher semesters and those who used ChatGPT more frequently tended to adopt more adaptive usage strategies, such as requesting conceptual clarification, performing independent verification, and reconstructing solutions using personal understanding. In contrast, students in earlier semesters and those who rarely used ChatGPT more often relied on it directly to obtain final answers. This integration confirms that score differences are not explained solely by formal

academic background, but by accumulated experience, reflective habits, and learning strategies that develop over time.

Quantitatively, the problem representation and prompt formulation dimensions were categorized as moderate difficulty. However, qualitative findings indicate that difficulties in these early dimensions frequently triggered difficulties in subsequent dimensions. Students who did not specify initial conditions, direction of motion, or physical assumptions in their prompts were more likely to receive contextually inappropriate ChatGPT responses, leading to confusion at the stages of understanding and verification. This integration shows that a moderate score at early stages does not imply limited impact, because difficulties at these stages can have cascading effects on the overall quality of AI-supported problem solving.

To clarify the integration of both data types, Table 14 presents a joint display linking key quantitative findings with supporting qualitative themes and their integrative interpretations.

Table 14. Joint display of integrated quantitative and qualitative findings

Key quantitative finding	Supporting qualitative finding	Integrative interpretation
1. D4 has the highest score (high difficulty category)	• Students rarely check units, signs, and numerical plausibility; ChatGPT is treated as an authority.	• The main difficulty is metacognitive and related to decisions not to perform verification.
2. Strong correlations between D3–D4 and D2–D4	• Superficial understanding of steps and reactive prompt revision.	• Difficulty forms a dependent process chain from prompting to verification.
3. Significant differences by semester	• Higher-semester students are more reflective and strategic.	• Academic maturity is associated with more effective ChatGPT usage strategies.
4. Significant differences by frequency of use	• Frequent users show adaptive strategies.	• Repeated experience supports the development of evaluative literacy.
5. Early dimensions in the moderate category	• Incomplete problem representation triggers downstream errors.	• Early-stage difficulties have systemic effects despite lower mean scores.

Overall, the integrated findings demonstrate that students' difficulties in using ChatGPT for mechanics of motion problem solving cannot be understood solely through numerical difficulty levels. These difficulties arise from interactions among problem representation skills, prompt quality, understanding of solution steps, and, most critically, evaluation and verification ability, all of which are shaped by students' experience and learning strategies. This integration underscores that ChatGPT does not automatically simplify or complicate problem solving. Instead, it amplifies the consequences of how students think, regulate their reasoning, and make decisions throughout the problem-solving process. These integrative insights provide a strong foundation for discussing pedagogical implications and formulating instructional recommendations in the subsequent section.

Discussion

Students' interactions with ChatGPT while solving routine mechanics of motion problems appear to be driven more by the robustness of their underlying problem-solving practices than by technical access or interface constraints. The moderate total difficulty score suggests that ChatGPT does not operate as a simple shortcut that reliably reduces cognitive load. Instead, it introduces an additional cognitive layer that can either support reasoning or expose weaknesses in students' conceptual and procedural knowledge. The presence of a meaningful high-difficulty subgroup reinforces that, without effective strategies, students may experience increased confusion, especially when ChatGPT is treated as a replacement for comprehension rather than a support for it (Arnold et al., 2007; Şucan et al., 2012).

A striking pattern is the prominence of evaluation and verification as the most difficult dimension. This implies that the most fragile point in AI-supported problem solving is not answer generation but the capacity to judge whether an answer is valid. Many students appeared to accept ChatGPT outputs because the solutions were presented in a neat, stepwise mathematical format, while essential verification routines such as checking units, sign conventions, limiting cases, or numerical plausibility were frequently skipped. In physics, such checks are not optional add-ons but core practices that sustain physical consistency. When students rely on surface cues of correctness, the activity shifts away from scientific reasoning and toward uncritical consumption of results, which can undermine learning even when the final numeric output looks plausible (Papenmeier et al., 2022; Papenmeier et al., 2023).

Prompt formulation also emerges as a conceptual issue rather than merely a technical writing skill. High-quality prompts depend on how students represent the physical situation. When prompts omit key information such as initial conditions, direction of motion, or simplifying assumptions, ChatGPT is likely to provide responses that are contextually mismatched. The qualitative contrast between students who copied problems verbatim and those who reorganized and contextualized the scenario before prompting suggests that effective prompting reflects conceptual modeling competence rather than linguistic proficiency alone (Lin, 2025; Hasany et al., 2025). In this sense, weak prompts are often symptoms of weak representations, and improving prompting requires strengthening how students structure the problem physically.

The correlation patterns further reinforce that the difficulties are interconnected rather than isolated. Strong relationships among problem representation, prompt formulation, understanding solution steps, and verification indicate that breakdowns early in the problem-solving chain can propagate forward. The strong link between understanding ChatGPT's solution process and verification is especially important: students cannot evaluate what they do not understand. Verification failures should therefore be interpreted not only as missed checking routines but also as signs of insufficient conceptual comprehension that prevents students from detecting inconsistencies in AI-generated reasoning (Thwaites et al., 2023; Nikdel et al., 2022).

Differences across semester levels suggest that accumulated knowledge and experience shape how students engage with ChatGPT. Students in earlier semesters tended to report higher difficulty levels, particularly in evaluation and verification, than students in later semesters. This pattern aligns with the gradual development of conceptual benchmarks and problem-solving routines that support plausibility judgments. At the same time, the limited differences between adjacent semesters indicate that this development is progressive rather than automatic, and it cannot be assumed to occur simply through academic advancement (Capella, 2025; Huang, 2025).

Frequency of ChatGPT use is also associated with reported difficulty levels, with infrequent users showing greater challenges than those using it more routinely, especially in verification-related tasks. This association can be interpreted in two plausible ways that must be kept distinct: repeated engagement may help students become familiar with limitations and develop strategic interaction habits, while students with stronger problem-solving skills may be more willing to use ChatGPT because they can evaluate and regulate its outputs. Given the cross-sectional design, frequency is better treated as an indicator of engagement style rather than a causal explanation for reduced difficulty (Adesso, 2023; Lyons et al., 2010).

Qualitative patterns clarify how these differences manifest in practice, especially through the nature of prompt revisions. Many students revised prompts in reactive ways, prompted by mismatches with peers or answer keys, without identifying which assumption or representation was flawed. Such trial-and-error behavior can increase dependency on the tool while producing limited conceptual insight. Students with lower difficulty more often revised prompts diagnostically by adding physical constraints, clarifying assumptions, or requesting explicit statements of the model before computation, allowing them to navigate the solution process more deliberately (Lyons et al., 2010; Hu, 2022).

Another critical implication concerns how students integrate AI outputs into their final solutions. Many students appeared to rewrite ChatGPT responses mechanically, without reconstructing the reasoning chain in their own words and logic. This raises concerns about the validity of traditional assessments that reward polished solutions, because an answer can look coherent while masking shallow understanding when AI assistance is involved. The tendency to prioritize the appearance of an acceptable solution over coherent reasoning suggests that educators may need to adapt assessment designs to elicit justification, verification, and reasoning ownership more explicitly (Kim & Yoo, 2021; Alwash & Khleaf, 2022).

Taken together, the findings support the view that ChatGPT magnifies pre-existing problem-solving behaviors. Students who engage in careful representation, targeted questioning, and rigorous verification can use the tool to refine and extend reasoning. Students who lack these practices are more likely to accept AI outputs uncritically and end reasoning prematurely. In this sense, ChatGPT does not inherently make mechanics of motion problem solving easier or harder; it amplifies the consequences of students' cognitive

approaches, self-regulation habits, and decision-making during problem solving (Fuente et al., 2021; Fuente & López, 2020).

These patterns point toward instructional priorities that place verification at the center of learning and treat prompt formulation as part of physical modeling. Learning designs should emphasize making assumptions explicit, justifying equation selection, and independently validating results before accepting AI outputs. Early-semester students and those with limited experience using ChatGPT may benefit from structured scaffolds that make these practices visible and required, while still ensuring that students, not AI systems, retain agency as the final evaluators of correctness (Su et al., 2021; Huang et al., 2024).

CONCLUSION

This study shows that university students' difficulties in using ChatGPT to solve routine mechanics of motion problems fall within the moderate category (overall mean approximately 3.22), yet these difficulties are not evenly distributed. A substantial subgroup of students (18.6%) experienced high levels of difficulty, indicating that the issue cannot be treated as incidental but reflects a recurring and structured pattern affecting a meaningful proportion of learners.

The most prominent and critical difficulty lies in evaluating and verifying ChatGPT-generated answers (mean approximately 3.69, high difficulty category), with nearly half of the respondents classified at the high-difficulty level for this dimension. This finding indicates that the primary challenge is not obtaining an answer, but rather students' capacity to judge the validity of that answer through checks of units, sign and direction consistency, numerical plausibility, conceptual alignment, and limiting cases. Qualitative evidence reinforces this conclusion, showing that many students terminate their reasoning once a numerical result is produced, interpreting the orderly and systematic presentation of ChatGPT's solutions as evidence of correctness. As a consequence, verification practices that are central to physics reasoning are frequently omitted.

Students' difficulties also form a connected and sequential structure rather than isolated problems. Correlation analysis revealed strong relationships between understanding ChatGPT's solution steps and evaluation and verification ($r \approx 0.62$), as well as between prompt formulation quality and verification ($r \approx 0.55$). These relationships suggest that difficulties often originate in incomplete problem representation, which leads to prompts lacking essential physical context such as direction of motion, initial conditions, or underlying assumptions. This deficiency propagates to superficial understanding of solution steps and ultimately undermines students' ability to evaluate the correctness of the results. Consequently, efforts that focus narrowly on improving prompt-writing techniques without strengthening physical modeling and verification skills are unlikely to produce meaningful improvements in learning quality.

Differences in difficulty levels were more strongly associated with academic semester and frequency of ChatGPT use than with study program. Students in earlier semesters demonstrated significantly higher levels of difficulty than those in later semesters,

particularly in evaluation and verification. Similarly, students who rarely used ChatGPT reported higher difficulty levels than those who used it more regularly. However, these patterns should not be interpreted causally, given the cross-sectional nature of the study. Qualitative findings suggest that these differences are better explained by variations in usage strategies. Students with lower difficulty tended to engage with ChatGPT reflectively by seeking conceptual clarification, requesting explicit assumptions, comparing solution methods, and conducting independent checks. In contrast, students with higher difficulty were more likely to rely on ChatGPT primarily for final answers without sustained evaluative reasoning.

Overall, the findings indicate that ChatGPT does not automatically simplify mechanics of motion problem solving. Instead, it amplifies the consequences of students' existing problem-solving habits and metacognitive regulation. Students who possess strong skills in representation, conceptual understanding, and verification can use ChatGPT as a productive learning aid, whereas students who lack these skills are more likely to accept AI-generated outputs uncritically, integrate solutions mechanically, and discontinue reasoning prematurely. Accordingly, the main pedagogical implication of this study is the need to place verification at the core of AI literacy in physics education and to treat prompt formulation as an integral component of physical modeling, rather than as an isolated technical skill.

LIMITATIONS

This study has several limitations that should be considered when interpreting the findings. First, the research was conducted at a single university with a relatively small sample size, which limits the generalizability of the results across different institutional contexts and student populations. Second, the cross-sectional design captures students' difficulties at one point in time, making it impossible to draw causal conclusions about the effects of ChatGPT use frequency or academic progression on problem-solving competence. Third, although routine mechanics of motion problems were deliberately chosen to reveal subtle weaknesses in representation and verification, this focus restricts the applicability of the findings to other types of physics problems, such as conceptual questions or context-rich, non-routine tasks. Finally, the study relied partly on self-reported questionnaire data, which may be influenced by students' perceptions of their own practices rather than their full range of actual behaviors, even though qualitative data were used to mitigate this limitation.

RECOMMENDATIONS

Based on the findings, several recommendations can be proposed for physics instruction and future research. Instructionally, educators should explicitly integrate verification practices, such as unit analysis, sign consistency, limiting-case checks, and plausibility reasoning, into AI-supported learning activities and assessments, particularly in early undergraduate courses. Prompt formulation should be taught as part of physical problem representation and modeling, not as a stand-alone technical skill. Learning tasks should require students to justify assumptions, explain equation selection, and articulate

independent checks before accepting AI-generated results. For future research, longitudinal and experimental designs are needed to examine how students' verification skills and AI interaction strategies develop over time and to test the effectiveness of targeted instructional interventions. Extending the analysis to other physics domains and to non-routine or conceptual problems would also help clarify whether the difficulty patterns identified in this study are domain-specific or more general features of AI-supported science learning.

Author Contributions

The authors have sufficiently contributed to the study, and have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Acknowledgment

The authors would like to thank the undergraduate students from Physics Education and Science Education programs who voluntarily participated in this study and shared their learning reflections. We also appreciate the support of lecturers and academic staff who facilitated access to participants and assisted in coordinating the data collection process.

Conflict of Interests

The authors declare no conflict of interest.

REFERENCES

Adesso, G. (2023). *Towards the ultimate brain: Exploring scientific discovery with ChatGPT AI.* https://doi.org/10.22541/au.167701309.98216987/v1

Alfirević, N., Praničević, D., & Mabić, M. (2024). Custom-trained large language models as open educational resources: An exploratory research of a business management educational chatbot in Croatia and Bosnia and Herzegovina. *Sustainability, 16*(12), 4929. https://doi.org/10.3390/su16124929

Alwash, N., & Khleaf, H. (2022). Artificial intelligent techniques applied for detection COVID-19 based on chest medical imaging. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 20*(2), 329. https://doi.org/10.12928/telkomnika.v20i2.20881

Arnold, R., Langheinrich, M., & Hartmann, W. (2007). InfoTraffic. In *Proceedings of the 2007 international workshop on Location- and context-awareness (LoCA 2007)* (pp. 105–109). https://doi.org/10.1145/1227310.1227349

Bigulov, A. (2025). Experimental application of the hyperintensive method for entering conversational practice in German using only AI tutors. *Linguistics & Polyglot Studies, 11*(3), 24–54. https://doi.org/10.24833/2410-2423-2025-3-44-24-54

Capella, S. (2025). How does generative AI affect patients' rights? *Voices in Bioethics, 11.* https://doi.org/10.52214/vib.v11i.14212

Chapagain, P., Malakar, N., & Rimal, D. (2024). Can AI solve physics problems? Evaluating efficacy of AI models in solving higher secondary physics exam problems: A comparative study. *Journal of Nepal Physical Society, 10*(1), 58–64. https://doi.org/10.3126/jnphyssoc.v10i1.72836

Chu, K. (2025). Prompt design and AI response quality in university students' academic learning tasks. *Proceedings of the Association for Information Science and Technology, 62*(1), 1402–1404. https://doi.org/10.1002/pra2.1417

Elson, L., Gaughan, L., Hopton, B., Lloyd-Brown, S., Briggs, C., Gharamti, M., … Moriarty, P. (2025). 'ChatGPT did my homework': Lessons learnt from embedding a chatbot in undergraduate coursework. *Physics Education, 60*(2), 025022. https://doi.org/10.1088/1361-6552/adb363

Fuente, I., & López, J. (2020). Cell motility and cancer. *Cancers, 12*(8), 2177. https://doi.org/10.3390/cancers12082177

Fuente, I., Martínez, L., Carrasco-Pujante, J., Fedetz, M., López, J., & Malaina, I. (2021). Self-organization and information processing: From basic enzymatic activities to complex adaptive cellular behavior. *Frontiers in Genetics, 12*, Article 644615. https://doi.org/10.3389/fgene.2021.644615

Galchuk, L. (2024). A conceptual framework for the instructional design of a Moodle-based e-learning course for mainstreaming informal master's education outcomes into ESP teaching in a formal non-linguistic setting. *RUDN Journal of Informatization in Education, 21*(3), 340–356. https://doi.org/10.22363/2312-8631-2024-21-3-340-356

Hasany, M., Kohestanian, M., Shabankareh, A., Nezhad-Mokhtari, P., & Mehrali, M. (2025). Ultra-stretchable, super-tough, and highly stable ion-doped hydrogel for advanced robotic applications and human motion sensing. *InfoMat, 7*(5). https://doi.org/10.1002/inf2.12655

Horchani, R. (2025). ChatGPT's problem-solving abilities in context-rich and traditional physics problems. *Physics Education, 60*(2), 025019. https://doi.org/10.1088/1361-6552/adb473

Hu, J. (2022). Research on robot fuzzy neural network motion system based on artificial intelligence. *Computational Intelligence and Neuroscience, 2022*, Article 4347772. https://doi.org/10.1155/2022/4347772

Huang, Y. (2025). *Planning with sketch-guided verification for physics-aware video generation* (arXiv:2511.17450). arXiv. https://doi.org/10.48550/arxiv.2511.17450

Huang, Y., Zhou, S., Su, Y., Pang, Z., & Cai, S. (2024). Diffusion-weighted imaging as a potential non-gadolinium alternative for immediate assessing the hyperacute outcome of MRgFUS ablation for uterine fibroids. *Scientific Reports, 14*(1), Article 60693. https://doi.org/10.1038/s41598-024-60693-4

Kim, J., & Yoo, S. (2021). Nonlinear-observer-based design approach for adaptive event-driven tracking of uncertain underactuated underwater vehicles. *Mathematics, 9*(10), 1144. https://doi.org/10.3390/math9101144

Kotsis, K. (2025). ChatGPT and DeepSeek in physics education: A narrative and thematic literature review of pedagogical implications. *International Journal of Advanced Multidisciplinary Research and Studies, 5*(5), 1080–1086. https://doi.org/10.62225/2583049x.2025.5.5.5074

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K., Lukowicz, P., Kühn, J., … Karolus, J. (2024). *Unreflected acceptance – Investigating the negative consequences of ChatGPT-assisted problem solving in physics education.* https://doi.org/10.3233/faia240195

Liang, Y., Zou, D., Xie, H., & Wang, F. (2023). Exploring the potential of using ChatGPT in physics education. *Smart Learning Environments, 10*(1). https://doi.org/10.1186/s40561-023-00273-7

Lin, Y. (2025). *SanDRA: Safe large-language-model-based decision making for automated vehicles using reachability analysis* (arXiv:2510.06717). arXiv. https://doi.org/10.48550/arxiv.2510.06717

Lunrasri, Y. (2020). *Measurement of reading literacy, growth, and learning potential of Grade 9 students: Application of computerized dynamic assessment concept* (Thesis). https://doi.org/10.58837/chula.the.2020.153

Lunrasri, Y., Tangdhanakanond, K., & Pasiphol, S. (2022). Effects of prompting type and learning achievement on reading literacy of ninth graders. *Kasetsart Journal of Social Sciences, 43*(2). https://doi.org/10.34044/j.kjss.2022.43.2.14

Lyons, D., Chaudhry, S., Agica, M., & Monaco, J. (2010). Integrating perception and problem solving to predict complex object behaviours. In *Proceedings of SPIE* (Paper 12.852484). https://doi.org/10.1117/12.852484

Mafudi, I. (2025). Case study on ChatGPT's performance in assisting students with physics tests. *Jurnal Pendidikan Fisika, 13*(1), 41–58. https://doi.org/10.26618/jpf.v13i1.16624

Malik, N., Kousar, A., & Bruun, V. (2025). Chat-GPT in education: Learning outcomes and facilitating knowledge acquisition. *IJSS, 4*(2), 97–104. https://doi.org/10.63544/ijss.v4i2.131

Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., … Campanelli, M. (2025). Using large language models for grading in education: An applied test for physics. *Physics Education, 60*(3), 035006. https://doi.org/10.1088/1361-6552/adb92b

Nikdel, P., Mahdavian, M., & Chen, M. (2022). *DMMGAN: Diverse multi motion prediction of 3D human joints using attention-based generative adverserial network* (arXiv:2209.09124). arXiv. https://doi.org/10.48550/arxiv.2209.09124

Papenmeier, F., Arrufi, J., & Kirsch, A. (2022). *Stories in the mind? The role of story-based categorizations in motion classification* [Preprint]. OSF. https://doi.org/10.31219/osf.io/hfc3q

Papenmeier, F., Arrufi, J., & Kirsch, A. (2023). Stories in the mind? The role of story-based categorizations in motion classification. *Cognitive Science, 47*(9), e13332. https://doi.org/10.1111/cogs.13332

Polverini, G., & Gregorcic, B. (2025). Multimodal large language models and physics visual tasks: Comparative analysis of performance and costs. *European Journal of Physics, 46*(5), 055708. https://doi.org/10.1088/1361-6404/ae03f8

Rezende, M., & Simó, V. (2024). What are the perceptions of physics teachers in Brazil about ChatGPT in school activities? *Journal of Physics: Conference Series, 2693*(1), 012011. https://doi.org/10.1088/1742-6596/2693/1/012011

Riabko, A., & Vakaliuk, T. (2024). Physics on autopilot: Exploring the use of an AI assistant for independent problem-solving practice. *Educational Technology Quarterly, 2024*(1), 56–75. https://doi.org/10.55056/etq.671

Rodriguez-Donaire, S. (2024). *Influence of prompts structure on the perception and enhancement of learning through LLMs in online educational contexts.* https://doi.org/10.5772/intechopen.1006481

Shin, E., Yu, Y., Bies, R., & Ramanathan, M. (2024). Evaluation of ChatGPT and Gemini large language models for pharmacometrics with NONMEM. *American Conference of Pharmacometrics (ACoP 2024), International Society of Pharmacometrics.* https://doi.org/10.70534/rqua9741

Su, Z., Dasgupta, M., Poitevin, F., Mathews, I., Bedem, H., Wall, M., & Wilson, M. (2021). Reproducibility of protein x-ray diffuse scattering and potential utility for modeling atomic displacement parameters. *Structural Dynamics, 8*(4), 044302. https://doi.org/10.1063/4.0000087

Sun, G. (2024). Prompt engineering for nurse educators. *Nurse Educator, 49*(6), 293–299. https://doi.org/10.1097/nne.0000000000001705

Şucan, I., Moll, M., & Kavraki, L. (2012). The open motion planning library. *IEEE Robotics & Automation Magazine, 19*(4), 72–82. https://doi.org/10.1109/mra.2012.2205651

Thwaites, D., Prokopovich, D., Garrett, R., Haworth, A., Rosenfeld, A., & Ahern, V. (2023). The rationale for a carbon ion radiation therapy facility in Australia. *Journal of Medical Radiation Sciences, 71*(S2), 59–76. https://doi.org/10.1002/jmrs.744

Trout, J., & Winterbottom, L. (2024). Artificial intelligence and undergraduate physics education. *Physics Education, 60*(1), 015024. https://doi.org/10.1088/1361-6552/ad98de

Ullah, H., Anwar, S., & Khan, S. (2024). Enhancing English language acquisition through ChatGPT: Use of technology acceptance model in linguistics. *Journal of English Language Literature and Education, 6*(4), 119–145. https://doi.org/10.54692/jelle.2024.0604262

Verawati, N., Wahyudi, W., Nisrina, N., & Asy'ari, M. (2025). ChatGPT in physics education: A content-based analysis on Newtonian force problems. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram, 13*(2), 335. https://doi.org/10.33394/j-ps.v13i2.15824

Wang, K., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1330486