REAL: Research in Education Assessment and Learning ISSN: 1947-5497 www.realjournal.org

*Research Articl*e *in Physics Education*

Unveiling ChatGPT-5's Capabilities in Physics Education: A Study on Conceptual Understanding and Reasoning

Dechao Lu[1,2], Jingbin Xue[1,2], Yuyang Qi[1,2] and Wangyi Xu[1,3,*]

[1] Chinese Education Modernization Research Institute, Hangzhou Normal University, Zhejiang China

[2] School of physics, Hangzhou Normal University, Zhejiang, China

[3.] Jing Hengyi School of Educaiton , Hangzhou Normal University, Zhejiang, China

[*] Corresponding Author: xwy@hznu.edu.cn

(Received: 03/12/2026; Accepted: 04/06/2026; Published: 04/20/2026)

DOI: https://doi.org/10.37906/real.2026.2

Abstract: This study investigates the performance of ChatGPT-5 in solving physics conceptual understanding and reasoning problems, with a focus on its implications for education. The research employed two widely used assessment tools, the Force Concept Inventory (FCI) and the Conceptual Survey of Electricity and Magnetism (CSEM), to evaluate ChatGPT-5's conceptual understanding, including both text-based and image-based questions. Additionally, eight primitive physics problems were used to assess its reasoning ability based on self-organizing representation theory. Results show that ChatGPT-5 achieved nearly perfect accuracy (100%) on text-based problems and substantial improvement (71.7%) on image-based items compared to earlier versions. In reasoning tasks, ChatGPT-5 reached an overall accuracy rate of 87.5%, outperforming both students and ChatGPT-4, particularly in abstraction, methodology, and physical representation dimensions. However, persistent weaknesses were identified in mathematical representation and in handling complex visual reasoning. These findings highlight ChatGPT-5's potential as a powerful educational tool for conceptual learning support and misconception clarification, while also revealing its current limitations in multimodal reasoning. The study underscores the necessity of integrating AI tools with teacher guidance to enhance students' comprehensive physics competencies.

Keywords: ChatGPT-5; problem solving; conceptual understanding, scientific reasoning

1. Introduction

In recent years, artificial intelligence (AI) has been increasingly integrated into the field of education, becoming a focal point of academic research. The specific applications of AI in education are mainly reflected in three aspects. First, AI can greatly enrich teaching resources (Tao et al., 2022). For example, Cai et al. demonstrated that interactive and integrated image formation experiments using augmented reality (AR) can stimulate students' interest in the subject (Cai et al., 2013). Similarly, Sidanin et al. designed a virtual physics laboratory based on nuclear experiments in higher education and investigated its impact on students' learning (Šiđanin et al., 2020). Second, AI can assist teachers or educators in evaluating learning outcomes. Kortemeyer et al. conducted studies on grading thermodynamics examinations with generative AI (Kortemeyer et al., 2024) and proposed theoretical approaches to scoring physics exams by combining psychological assessment methods (Kortemeyer & Nöhl,2 024). Nguyen et al. also explored the potential of using generative AI to evaluate and improve K-12 science education standards (Nguyen & Hayward, 2025). In addition, Li et al. demonstrated that AI enhances the accuracy, precision, recall, and F1 scores of various evaluation models, contributing to the fairness and validity of automated grading systems (Li et al., 2025). Third, AI can provide teachers and educators with more information to better understand students' learning behaviors. Krusberg et al. illustrated that AI tools in physics education, such as Physlet Physics, Andes ITS, and Microworlds, can effectively help teachers understand students' learning processes (Krusberg, 2007). Furthermore, tools such as ACTIVE Math, MATHia, Why2Atlas, Comet, and Viper have also been applied to various subjects to create assessment instruments (Chassignol et al., 2018). Although some studies have indicated that AI technologies have not significantly influenced students' academic performance in subjects such as physics (Chang et al., 2016; Lin et al., 2015), the trend of AI integration in education appears to be inevitable, with great potential to empower high-quality education in the future(Hu et al., 2022). Moreover, many studies have found that both teachers(Beege M et al., 2024; WATTANAKASIWICH P et al., 2025) and students (Sofos, 2025; Tong & Chen, 2024) hold positive attitudes toward incorporating generative AI into educational practices.

Since its release on November 30, 2022, ChatGPT reached over 100 million active users within just two months. Users employ it for writing papers, coding, storytelling, and answering questions (Zhai). Its accuracy and versatility have been widely recognized. Unlike traditional chatbots, ChatGPT is an advanced conversational AI model developed from large language models (LLMs). With a larger parameter scale and training dataset, it has stronger text generation capabilities, simulating human-like conversations and providing interactive learning experiences (Brown et al., 2020). Therefore, ChatGPT can be integrated into teaching as a learning tool to support personalized learning. However, concerns have been raised regarding its potential misuse in academic misconduct, such as students relying on it to complete homework assignments or researchers using it to write papers and reports. Academic journals, including Nature, have explicitly prohibited the use of ChatGPT or other LLMs to generate text for publications and have stated that ChatGPT cannot be listed as a co-author (Thorp, 2023). To address such widespread misuse, Coelho et al. investigated students' use of LLM platforms like ChatGPT and advocated for designing effective educational partnerships to promote the responsible use of AI tools for learning (Coelho et al., 2025).

At the same time, developing students' problem-solving ability has long been a central concern in education. In physics instruction, the ability to solve problems is considered one of the key indicators of students' core physics competencies. Consequently, several studies have investigated students' behaviors during problem solving. For instance, Bagno et al. compared the differences between novice and expert students in their methods and strategies during problem solvings (Bagno et al., 1997). With the growing maturity of AI technologies, students' problem-solving behaviors under AI support have also become a research direction. For example, Lepp et al. found that students' attitudes toward generative AI influenced their performance on programming tasks(Lepp M & Kaimre, 2025), while Lu et al. compared students' performance when completing supplementary math assignments versus when assisted by AI (Lu et al., 2024).

As noted above, since its release, ChatGPT has been widely adopted for various tasks (Cooper, 2023; Zhai, 2023). While its benefits to education have been widely acknowledged, its use has also raised risks and challenges. By simply entering a question into the dialogue box, students can obtain detailed answers from ChatGPT, which may encourage academic dishonesty such as cheating on exams or plagiarizing assignments. This could undermine deep thinking and compromise educational equity (Anders, 2023). Such issues are among the main concerns expressed by teachers and parents. This raises critical questions: Is ChatGPT truly as powerful as widely claimed? How well does it actually perform in solving scientific problems?

In the domain of ChatGPT's performance on physics problem solving, a number of studies have already been reported. Gregorcic and Pendrill examined ChatGPT-3.5's ability to solve basic physics problems and found that the model produced flawed reasoning processes (Gregorcic Bor & Pendrill Ann-Marie, 2023). Simó and Rezende tested ChatGPT-3.5 on various high school physics problems (forces and motion), including definitional, simple calculation, multi-step calculation, reasoning, and Fermi problems. The results showed that the model performed well on definitional and simple calculation problems, but poorly on complex and open-ended problems (López-Simó & Rezende Junior, 2024). Kortemeyer tested ChatGPT-3.5 with FCI items, real homework, clicker questions, and final exam problems in mechanics, reporting an average weighted score of 53.05%, equivalent to a grade point average of only 1.5, barely sufficient for course credit. The model's performance on calculation problems was especially problematic (Kortemeyer, 2023). In contrast, Kieser, Wulff, and Kuhn showed that ChatGPT-4 could accurately solve FCI problems (Kieser et al., 2023). Following the release of ChatGPT-4 in March 2023, West administered 23 FCI items to both ChatGPT-3.5 and ChatGPT-4, converting visual questions into text format. The study revealed that ChatGPT-3.5 achieved an average score of 65.22%, while ChatGPT-4 reached 95.65%, comparable to expert-level performance (C. G. West, 2023). To further evaluate ChatGPT's scientific reasoning abilities, Tong et al. proposed expanding the range of problem types and conducted systematic studies on ChatGPT-4, including its performance on the FCI (D. Hestenes, M. Wells, and G. Swackhamer, 1992), the Conceptual Survey of Electricity and Magnetism (CSEM), primitive physics problems(Tong et al.,2024), and China's national physics examinations at the secondary (Zeng & Tong, 2024) and high school (Liu & Tong, 2024) levels. The findings indicated that ChatGPT-4 outperformed the average level of both middle and high school students in solving physics problems.

On August 8, 2025, the release of ChatGPT-5 raises an important question: as the latest large-scale AI model, will ChatGPT-5 demonstrate superior performance in solving physics problems? Furthermore, previous studies have predominantly focused on FCI-related conceptual understanding. Due to earlier limitations of LLMs in handling multimodal inputs, most problem-solving tasks have been restricted to text-based inputs, excluding physics image-based problems. This has led to an incomplete evaluation of ChatGPT's physics problem-solving abilities. Therefore, the present study takes ChatGPT-5 as the research object, testing its performance on both conceptual understanding and reasoning problems in physics, including questions presented in both text and visual formats.

2. Methods

2.1. Research Subjects

In this study, the performance of ChatGPT-5 in solving physics problems was compared with that of ChatGPT-4 and students reported in previous research. It should be noted that since ChatGPT is under continuous development, the same version may yield different testing results at different times. The version of ChatGPT-5 used in this study was the earliest release, launched in August 2025.

2.2. Testing Materials

To comprehensively evaluate ChatGPT-5's ability to solve physics problems, we adopted and extended the materials tested by Tong et al. (2024). The physics problems were categorized into two groups: conceptual problems and reasoning problems.

The conceptual problems covered mechanics and electromagnetism. The Force Concept Inventory (FCI) (Hestenes & Halloun, 1995) and the Conceptual Survey of Electricity and Magnetism (CSEM) (Maloney et al., 2001), which are widely recognized assessment tools in physics education, were employed in this study. Since ChatGPT-5 allows multimodal input, all FCI and CSEM items, including both text-based and graphical/image-based questions, were used as prompts to evaluate its performance in conceptual understanding. The FCI consists of 30 items and the CSEM consists of 32 items, yielding a total of 62 questions.

Assessing students' physics reasoning ability is an important component of physics education. Based on the principle that "phenomena are the source of physics," Xing and Chen introduced the concept of primitive physics problems, which refer to naturally existing physics-related problems in daily life and production that have not been artificially processed (Xing, 2009). These problems are presented through text-based descriptions of phenomena, which are suitable for ChatGPT-based evaluation, as they connect physical phenomena with students' calculation and reasoning abilities. To evaluate such problems, Xing proposed the self-organizing representation theory (Xing & Chen, 2005), which assesses six dimensions: abstract representation, image representation, assignment representation, physical representation, methodological representation, and mathematical representation. In this study, we adopted the assessment instruments developed by Xing for middle and high school primitive physics problems, comprising a total of eight items (Xing et al., 2015; Xing & Chen, 2008).

2.3. Data Analysis

For the 62 FCI and CSEM items, which are multiple-choice questions, scores were assigned based on the correctness of ChatGPT-5's final answers: one point for a correct answer and zero for an incorrect one. Thus, the maximum possible score for the conceptual understanding part was 62.

In contrast, the primitive physics problems used to test physics reasoning ability required evaluation based on the self-organizing representation theory, as solving them involves abstract modeling, defining physical variables, selecting physical laws, and performing mathematical calculations. Following Xing et al., one point was awarded for each correct representation and zero points for each incorrect one. Considering ChatGPT-5's limited ability to generate figures/images, image representation was excluded from the evaluation. Therefore, abstract, assignment, physical, methodological, and mathematical representations were adopted as scoring criteria. Each primitive physics problem had a maximum of five points, giving a total possible score of 40 for all eight problems.

To ensure accuracy and inter-rater reliability, three researchers with prior experience in scoring primitive physics problems independently graded the responses based on the five representation categories. In the first round of scoring, the three raters reached full agreement on 35 out of 40 representations across the eight problems (see Table I), corresponding to an agreement rate of 87.5%, which was slightly low. After reflective discussion and clarification of scoring standards, a second round of scoring was conducted. In this round, the raters achieved full agreement on 37 out of 40 representations (see Table II), corresponding to an agreement rate of 92.5%.

Table I. Results of the first-round scoring of physics reasoning problems

Representation	Abstraction	Assignment	Methodology	Physics	Mathematics
M-Q1	①②③	①②③	③	①②③	
M-Q2	①②③	①②③	①②③	①②③	③
M-Q3	①②③	①②③	①②③	①②③	①②③
M-Q4	①②③	①②③	①②③	①②③	①②③
H-Q1	①②③	①②③	①②③	①②③	
H-Q2	①②③	①②③	①②③	①②③	①②③
H-Q3	①②③	①②③	①②③	①②③	①②③
H-Q4	①②③	①②③	②③	①②	①②

M denotes the assessment tool for middle school primitive physics problems, while H denotes the assessment tool for high school primitive physics problems. Symbols ①, ②, and ③ represent Researcher 1, Researcher 2, and Researcher 3, respectively.

Table II. Results of the second-round scoring of physics reasoning problems

Representation	Abstraction	Assignment	Methodology	Physics	Mathematics
M-Q1	①②③	①②③	③	①②③	
M-Q2	①②③	①②③	①②③	①②③	③
M-Q3	①②③	①②③	①②③	①②③	①②③
M-Q4	①②③	①②③	①②③	①②③	①②③
H-Q1	①②③	①②③	①②③	①②③	
H-Q2	①②③	①②③	①②③	①②③	①②③
H-Q3	①②③	①②③	①②③	①②③	①②③
H-Q4	①②③	①②③	②③	①②③	①②③

M denotes the assessment tool for middle school primitive physics problems, while H denotes the assessment tool for high school primitive physics problems. Symbols ①, ②, and ③ represent Researcher 1, Researcher 2, and Researcher 3, respectively.

3. Results

Based on this framework, we conducted a comparative analysis of the performance of ChatGPT-5, ChatGPT-4, and students in solving physics conceptual understanding and reasoning problems.

3.1. Performance of ChatGPT-5 in Solving Physics Problems

3.1.1. Performance on Physics Conceptual Understanding Problems

The researchers used the 30 items from the FCI and the 32 items from the CSEM, including both text-based and graphical/image-based questions, as prompts in the ChatGPT-5 dialogue box. ChatGPT-5 was required to provide answers accompanied by brief explanations. The researchers then reviewed each response and determined its accuracy. The scoring results of ChatGPT-5 are presented in Table III-Table V.

It was found that ChatGPT-5 answered 16 text-based questions correctly and concisely, achieving a 100% accuracy rate, which is consistent with the performance of ChatGPT-4 reported in previous studies(Tong et al., 2024). Furthermore, the findings suggest that ChatGPT-5 demonstrated strong adaptability in directly solving graphical/image-based physics conceptual understanding problems. Regardless of whether the items came from the FCI or CSEM, the lowest accuracy rate was 70%, with an average accuracy of 71.739%. These results indicate that ChatGPT-5 has achieved significant improvement in handling multimodal input and solving physics problems, showing considerable developmental potential.

Table III. Performance of ChatGPT-5 on text-based questions

	FCI(Total of 10 questions)	CSEM(Total of 6 questions)
Number of correct questions	10	6
Percentage (of plain text)	100%	100%
Average Accuracy	100%	

The plain text questions for FCI include Q1, 2, 3, 4, 13, 25, 26, 27, 29, 30, and the plain text questions for CSEM include Q1, 2, 10, 11, 16, 21.

Table IV. Performance of ChatGPT-5 on graphical/image-based questions

	FCI(Total of 20 questions)	CSEM(Total of 26 questions)
Number of correct questions	14	19
Percentage (image question)	70%	73.077%
Average Accuracy	71.739%	

The graphic/image questions for FCI include Q5-12, 14-24, 28, while the graphic/image questions for CSEM include Q3-9, 12-15, 17-20, 22-32.

Table V. Overall performance of ChatGPT-5 on conceptual understanding problems

	FCI(Total of 30 questions)	CSEM(Total of 32 questions)
Number of correct questions	24	25
Percentage (of all questions)	80%	78.125%

In terms of plain text questions, we take Q30 from the FCI test as an example:

Despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court.

Consider the following forces:
1. A downward force of gravity.
2. A force by the "hit".
3. A force exerted by the air.

Which of the above forces is (are) acting on the tennis ball after it has left contact with the racquet and before it touches the ground?

(A)1 only.
(B)1 and 2.
(C)1 and 3.
(D)2 and 3.
(E)1, 2, and 3.

In this case, we get the response from ChatGPT-5 as follows:
Answer: (C).

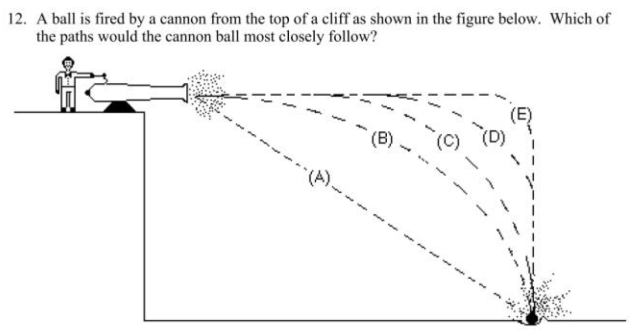
Reasoning: Once the ball loses contact with the racquet, the "hit" force (2) no longer acts - forces require continuous interaction. After that moment, only:

1. Gravity acts downward.
2. Air forces (drag, and possibly lift from spin or wind) act while the ball is in flight.

Thus, after leaving the racquet but before hitting the ground, the acting forces are (1) and (3), making (C) correct.

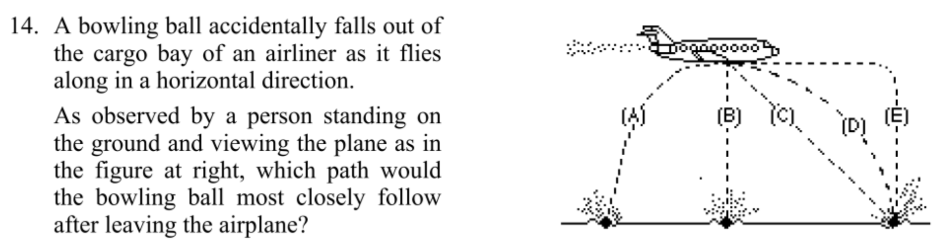
We found that ChatGPT-5's responses are very concise; it not only understands the concept of force but also pays attention to the concept of 'action force' that ordinary students often misunderstand in the questions. This indicates that ChatGPT-5 has a good understanding of the misconceptions related to physics concepts, especially compared to general students who might think that a force is needed to maintain the motion of an object.In terms of graphical/image problems, we found that ChatGPT-5 often analyzes correctly but chooses the wrong answer, especially when the answer needs to be selected from an image. We take FCI's Q12 as an example:

Figure I. FCI's Q12



In this case, the response from ChatGPT-5 is as follows:
C–Projectile motion: launched horizontally, then curves downward (parabola).
We found that although ChatGPT-5 recognized that the trajectory of a cannonball is a parabola, it still incorrectly chose C, failing to notice that the motion trajectory at the beginning of curve C is a horizontal straight line. Similarly, in FCI Q14:

Figure II. FCI's Q14



The response from ChatGPT-5 is as follows:
C–The ball keeps the plane's horizontal speed while falling: it stays below the plane (curved path).
We found that ChatGPT-5 also recognized that the horizontal initial velocity of the projectile is the same as that of the airplane; however, it did not explicitly emphasize that the vertical motion should follow free fall. Although it noted in parentheses that the trajectory of the projectile is a curve, it still chose the incorrect option C.

From this, we can qualitatively conclude that ChatGPT-5 is generally able to identify the contents of an image, understand them, and integrate basic physics knowledge into its analysis. Nevertheless, its reasoning sometimes shows omissions. In other words, if ChatGPT-5 were asked to revisit the image for further processing or to re-select the correct answer, it might still make mistakes-much like students who often reflect after an exam, "I was thinking along the right lines, but I still got it wrong."

3.1.2. Performance of ChatGPT-5 on Physics Reasoning Problems

In order to evaluate ChatGPT-5's performance on physics reasoning tasks, a total of eight standard primitive physics problems were administered as prompts. The answers were reviewed according to the established scoring rubric, assessing whether each response sufficiently included the five representational dimensions and whether the final answer was correct.

The analysis showed that ChatGPT-5 achieved a score of 17 out of 20 points (85%) on middle-school-level problems and 18 out of 20 points (90%) on high-school-level problems. Overall, its mean accuracy rate on reasoning problems was 87.5%, representing a clear improvement compared to the previous benchmark of 80%.

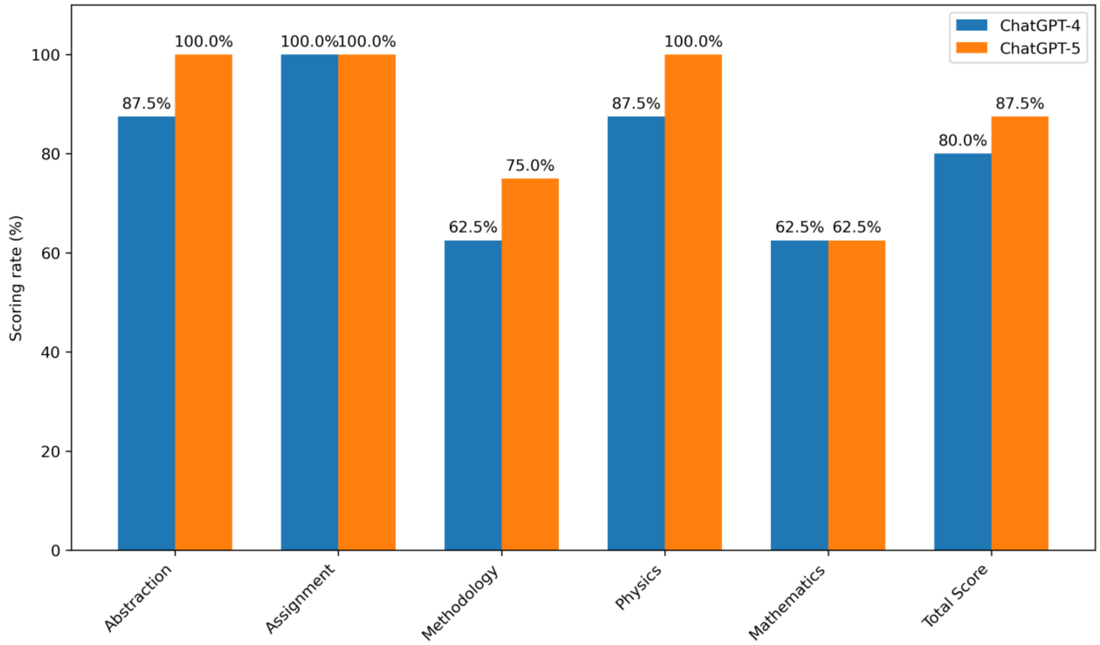
Despite this progress, several systematic weaknesses were observed. ChatGPT-5 frequently exhibited basic errors in methodological and mathematical representations. For example, in the problem M-Q1 "Carrying a box upstairs", ChatGPT-5 correctly recognized that the task involved the principle of lever equilibrium. However, it failed to analyze the lines of action of the forces accurately, assuming equal distances for the front and rear individuals without considering the actual geometry. Interestingly, it also provided an alternative solution from a work–energy perspective, yet similarly overlooked differences in height.

Another notable case occurred in M-Q4 "A boat with stones in a swimming pool". ChatGPT-5 classified the situation into three cases depending on the relative densities of the stones and water ($\rho_{stones} > \rho_{water}, \rho_{stones} = \rho_{water}, \rho_{stones} < \rho_{water}$). Since the correct solution was contained within one of these cases, the response was judged as acceptable. This multi-scenario reasoning strategy illustrates ChatGPT-5's capacity for comprehensive thinking. Nevertheless, it raises an important pedagogical concern: whether such "overthinking" might guide students toward redundant or even divergent solution paths in real learning contexts. This issue warrants further investigation.

Table VI. The performance of ChatGPT-5 in solving physics reasoning problems

Representation	Abstraction	Assignment	Methodology	Physics	Mathematics	score
M-Q1	1	1	0	1	0	60%
M-Q2	1	1	1	1	0	80%
M-Q3	1	1	1	1	1	100%
M-Q4	1	1	1	1	1	100%
H-Q1	1	1	1	1	0	80%
H-Q2	1	1	1	1	1	100%
H-Q3	1	1	1	1	1	100%
H-Q4	1	1	0	1	1	80%
Total Score	8	8	6	8	5	35(87.5%)

Figure III. Comparison of performance between ChatGPT-4 and ChatGPT-5 in solving physics reasoning problems



In summary, we found that ChatGPT-5 exhibits a high accuracy in solving pure text problems and shows strong potential in addressing graphical/image issues. At the same time, in terms of solving physics reasoning problems, ChatGPT-5 has made certain improvements in method representation and physical representation dimensions compared to ChatGPT-4.

3.2. Students' Performance in Solving Physics Problems

3.2.1. Performance in Conceptual Understanding of Physics Problems

The Force Concept Inventory (FCI) is an assessment tool that covers fundamental concepts of introductory mechanics. It is widely used to evaluate students' understanding of mechanics concepts and to reflect their mastery of basic mechanics knowledge. Caballero et al. (2012) administered an FCI pre-test to 5,000 university students, with an average score of 48%. After completing a traditional mechanics course and a Matter and Interactions course, students' average FCI post-test scores were 71.3% and 59.3%, respectively. Similarly, West tested 415 university students with the FCI and reported an average score of 56.30% (West, 2023).

The Conceptual Survey of Electricity and Magnetism (CSEM) is another widely used assessment tool that includes fundamental concepts such as charge, electric field, electric potential, and magnetic field. It is commonly used to examine students' conceptual understanding of electromagnetism. For example, Maloney et al. (Maloney et al, 2001) conducted pre- and post-tests with more than 5,000 physics students from 30 different institutions. Their results showed that students enrolled in calculus-based courses achieved an average pre-test score of 31%, while those in algebra-based courses scored 25%. After instruction, their scores improved to 47% and 44%, respectively.

These studies collectively indicate that students' ability to solve physics problems involving conceptual understanding is relatively weak, and that they often encounter significant difficulties in grasping the underlying concepts.

3.2.2. Performance in Physics Reasoning Problems

Based on existing research on students' performance in solving primitive physics problems, it has been found that students generally perform poorly when dealing with such problems, as shown in the following table.

Table VII. Score distribution of students in solving primitive physics problems

Score	0	1	2	3	4	5	6	7	8	9	10	11	12
High	1	7	9	4	10	9	10	7	3	2	2	9	3
Middle	125	51	40	29	14	21	30	15	19	9	6	4	7
Total	126	58	49	33	24	30	40	22	22	11	8	13	10
Score	13	14	15	16	17	18	19	20	21	22	23	24	
High	0	1	1	0	1	0	0	0	0	0	1	0	
Middle	5	3	1	3	1	2	0	0	0	0	0	0	
Total	5	4	2	3	2	2	0	0	0	0	1	0	

Referring to the six levels of the self-organized representation theory, we cite the findings of Xing H. J. et al., as shown in Table VIII.

Table VIII. Scores of secondary school students in solving primitive physics problems

Representation	Abstraction	Assignment	Image	Methodology	Physics	Mathematics
Full	4	4	4	4	4	4
Mean	0.9330	0.5696	0.8557	0.3557	0.4871	0.3170
Rate	23.32%	14.24%	21.39%	8.89%	12.18%	7.93%

We found that students scored relatively low in method representation and mathematical representation when solving primitive physics problems. This phenomenon is closely related to the prevailing educational philosophy in physics teaching. For a long time, physics education has primarily focused on knowledge transmission, emphasizing the memorization and reproduction of concepts and laws, while the explicit training of scientific inquiry methodologies has been relatively weak. The formation of method representation requires students to actively analyze problem situations, select appropriate methods, and engage in logical reasoning during the problem-solving process. However, this process is often conveyed implicitly rather than explicitly in classroom teaching. As a result, students lack a deep understanding of the principles and conditions underlying method selection. Consequently, when facing open-ended or novel problems, they often struggle to flexibly transfer and apply knowledge, leading to low scores in method representation.

With regard to mathematical representation, solving primitive physics problems requires not only basic algebraic computation skills but also the ability to transform physical situations into mathematical models and to derive quantitative results through logical reasoning. However, school instruction tends to focus more on the direct application of formulas and repetitive practice of familiar problem types. Students' mathematical modeling and reasoning abilities therefore rely heavily on their individual understanding and accumulation. The lack of systematic interdisciplinary training makes it difficult for students to handle problems that involve multi-step reasoning, simultaneous variables, or approximate estimation. Thus, their low performance in mathematical representation is not merely due to calculation errors but reflects deeper deficiencies in integrating mathematics with physics and in applying methodological thinking.

3.3. Comparison of ChatGPT-5, ChatGPT-4, and Students in Solving Physics Problems

First, we compare the performance of ChatGPT-5, ChatGPT-4, and students in solving physics concept understanding problems. According to existing research, results vary among studies due to differences in student populations, even when the same concept assessment tools are used. Therefore, based on the surveys conducted by West and Maloney et al., we estimate that the average student score rate is approximately 60% on the FCI and 45% on the CSEM, yielding an overall average of about 50% for physics concept understanding problems. However, previous studies have shown that ChatGPT-4 achieved nearly 100% accuracy on these tasks. As an iterative successor, ChatGPT-5 unsurprisingly also reached an average score rate close to 100% on physics concept understanding problems, specifically for the text-based items in the FCI and CSEM.

In contrast, relatively few studies have investigated performance on graphical/diagrammatic items. Based on our testing results, we found that ChatGPT-5 achieved approximately 71.739% accuracy on such problems. This indicates that, although ChatGPT-5 has reached near-perfect performance on text-based concept understanding problems, its performance still lags on problems involving graphical/diagrammatic information. This discrepancy may stem from two main factors: (1) graphical problems require both information extraction and physical modeling, and ChatGPT-5 may still miss details or misinterpret spatial relations; (2) some graphical problems contain key information not explicitly presented but embedded in features such as proportions, orientations, and positions, which impose higher demands on multimodal integration. Nevertheless, compared with earlier versions that performed much worse on graphical problems, ChatGPT-5's accuracy of 71.739% demonstrates substantial progress, indicating great potential in multimodal physics problem solving. Looking forward, further enhancement of the model's visual perception and spatial reasoning capabilities could enable its performance on graphical problems to approach or even match that on text-based ones. However, misuse of ChatGPT-5 could threaten the validity of these standardized assessment tools.

From an educational perspective, these findings provide dual implications. On one hand, ChatGPT-5's high accuracy on text-based problems suggests that it can serve as an effective classroom tool for conceptual explanation and misconception clarification, especially for personalized tutoring, after-class review, and online learning support. On the other hand, its limitations on graphical/diagrammatic problems highlight the current imperfections of AI in multimodal information processing. Students still need systematic training to develop skills in extracting information from diagrams, engaging in spatial reasoning, and constructing models. Therefore, in teaching practice, ChatGPT-5 can be integrated as an auxiliary learning resource alongside human instruction-leveraging its strengths in concept understanding while preventing overreliance on its still-developing visual reasoning capacity-to foster students' comprehensive physics literacy.

Next, we compare the performance of ChatGPT-5, ChatGPT-4, and students in solving physics reasoning problems. To better capture the differences, we adopt the five representational dimensions of the self-organizing representation theory (excluding image representation, given the current limitations of ChatGPT's image output). Based on existing literature and the findings of this study, we obtained the data summarized in Table IX.

Table IX. The Performance Differences among Secondary School Students, ChatGPT-4, and ChatGPT-5 in Solving Primitive Physics Problems

Representation	Abstraction	Assignment	Methodology	Physics	Mathematics	Total
Middle school students	23.32%	14.24%	8.89%	12.18%	7.93%	23.32%
ChatGPT-4	87.50%	100%	62.50%	87.50%	62.50%	80.00%
ChatGPT-5	100%	100%	75.00%	100%	62.50%	87.50%

The table provides a detailed comparison of the performance of secondary school students, ChatGPT-4, and ChatGPT-5 in solving primitive physics problems. Due to current technical limitations, ChatGPT still shows deficiencies in generating image-based feedback for physics reasoning tasks. However, in terms of abstraction, assignment, methodological, physical, and mathematical representations, ChatGPT-5 not only significantly outperforms secondary school students but also demonstrates further improvements over ChatGPT-4 in abstraction, methodological, and physical representations.

At the same time, we observed that ChatGPT-5's progress in mathematical representation is not particularly pronounced. This suggests that, when it comes to mathematical derivation and quantitative analysis, ChatGPT-5 still faces certain limitations. Mathematical representation requires not only accurate computational ability but also the appropriate translation of physical situations into mathematical models, while maintaining precision and consistency throughout multi-step logical reasoning. Although ChatGPT-5 exhibits stability in language understanding and formula generation, it is still prone to errors when handling long-chain reasoning, approximate calculations, or complex inter-variable relationships.

Interestingly, this finding parallels the generally low performance of students in mathematical representation, suggesting that both human learners and AI systems require more targeted training in mathematical modeling and reasoning. From an educational perspective, it is therefore necessary to combine the strengths of AI tools with teacher guidance: on the one hand, leveraging ChatGPT-5's advantages in abstraction, methodology, and physical representation to support students in modeling and strategic decision-making; on the other hand, strengthening mathematical reasoning and validation skills through classroom instruction and practice. Such an integrated approach may foster more comprehensive competencies in solving physics problems.

5. Discussion and Conclusion

Compared with the performance of ChatGPT-4 and secondary school students, ChatGPT-5 demonstrates significant advantages in solving both physics concept understanding problems and physics reasoning problems. This result not only provides new empirical evidence for physics education but also implies that future reforms in physics teaching must systematically consider the application pathways of artificial intelligence, particularly AIGC (AI-generated content) technologies. In terms of text-based conceptual problems, ChatGPT-5 has achieved nearly perfect accuracy, indicating its near-expert capability in concept explanation and misconception clarification. However, in graphical or image-based problems, although its average accuracy has reached 71.739%-a marked improvement over its predecessor-it still shows deficiencies in information detail extraction and spatial reasoning. This suggests that artificial intelligence has not yet reached a fully reliable level in multimodal information integration.

In reasoning problems, ChatGPT-5 scores significantly higher than secondary school students and outperforms ChatGPT-4 in abstraction, methodological, and physical representations, reflecting a more mature ability to analyze contexts, select strategies, and apply physical laws. Nevertheless, its progress in mathematical representation is less pronounced, as it remains prone to errors in long-chain reasoning and quantitative modeling. This limitation interestingly mirrors the generally low performance of students in mathematical representation, highlighting that both human learners and AI systems require specialized training to strengthen mathematical reasoning and modeling.

From the perspective of educational practice, the high performance of ChatGPT-5 represents both an opportunity and a challenge. On the one hand, it can serve as an important support tool for physics learning, particularly by providing instant feedback and diverse perspectives in personalized tutoring, after-class review, and open-ended inquiry tasks. On the other hand, overreliance on its generated final answers may undermine students' independent problem-solving and reflective abilities, fostering "copy-based learning" rather than genuine deep understanding. To mitigate this risk, teachers can guide students to focus not only on the results but also on the reasoning processes of ChatGPT-5, and encourage collective examination of AI's solution pathways in group learning activities, thereby reducing individual path dependence and cognitive blind spots.

Furthermore, our findings show that ChatGPT-5, when faced with open-ended physical phenomena, often provides multiple possible situational analyses. While such "overly comprehensive" reasoning may inspire students to adopt multi-perspective thinking, without proper teacher guidance it may also make students' solution approaches redundant or even ambiguous. Thus, future classroom designs should explore optimized modes of human–AI collaboration, combining the divergent thinking of AI with the focusing guidance of teachers to promote inquiry that is both broad and precise.

Overall, ChatGPT-5 holds tremendous potential as a tool for physics learning and teaching, with its true value depending on how educators integrate it scientifically and ethically into classroom and extracurricular contexts. Future education systems should prioritize the following aspects: first, incorporating AI literacy and ethics into curriculum design to help students rationally evaluate and critically use AI; second, developing targeted training tasks that address AI's current limitations in multimodal reasoning and mathematical modeling, thereby enhancing students' ability to collaborate with AI in solving complex problems; and third, promoting teacher professional development so that teachers can reconstruct instructional strategies and assessment methods in AI-assisted learning environments.

Looking ahead, as artificial intelligence continues to evolve, the knowledge production modes of the human scientific community will undergo profound transformations. Only when technological literacy and ethical literacy are unified in education can AI truly become a driving force for high-quality development in physics education, rather than a crutch that replaces students' independent thinking.

Funding: This research received no external funding

Acknowledgments: The research is supported in part by the Zhejiang Philosophy and Social Science Planning Project 23YJKT03YB. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References:

Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? Patterns, 4(3), 100694. https://doi.org/10.1016/j.patter.2023.100694

Bagno, E., & Eylon, B. S. (1997). From problem solving to a knowledge structure: An example from the domain of electromagnetism. American Journal of Physics, 65(8), 726–736. https://doi.org/10.1119/1.18642

Beege, M., Hug, C., & Nerb, J. (2024). AI in STEM education: The relationship between teacher perceptions and ChatGPT use. Computers in Human Behavior Reports, 16, 100494. https://doi.org/10.1016/j.chbr.2024.100494

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Caballero, M. D., Greco, E. F., Murray, E. R., Bujak, K. R., Marr, M. J., Catrambone, R., Kohlmyer, M. A., & Schatz, M. F. (2012). Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. American Journal of Physics, 80(7), 638–644. https://doi.org/10.1119/1.3703517

Cai, S., Chiang, F.-K., & Wang, X. (2013). Using the augmented reality 3D technique for a convex imaging experiment in a physics course. International Journal of Engineering Education, 29(4), 856–865.

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: A narrative overview. Procedia Computer Science, 136, 16–24. https://doi.org/10.1016/j.procs.2018.08.233

Chang, R.-C., Chung, L.-Y., & Huang, Y.-M. (2016). Developing an interactive augmented reality system as a complement to plant education and comparing its effectiveness with video learning. Interactive Learning Environments, 24(6), 1245–1264. https://doi.org/10.1080/10494820.2014.982131

Coelho, R., Bjune, A. E., Ellingsen, S., Muñoz Solheim, B., Thormodsæter, R., Wasson, B., & Cotner, S. (2025). A call for clarity: Biology students advocate for guidelines for the use of generative AI in higher education. Journal of Science Education and Technology, 34(4), 853–865. https://doi.org/10.1007/s10956-025-10216-1

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. Physics Education, 58(3), 035021. https://doi.org/10.1088/1361-6552/acc299Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30(3), 141–158. https://doi.org/10.1119/1.2343497

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. The Physics Teacher, 33(8), 502–506. https://doi.org/10.1119/1.2344278

Hu, X., Sun, S., Yang, W., & Ding, G. (2022). 人工智能赋能教育高质量发展：需求、愿景与路径 [Artificial intelligence empowering the high-quality development of education: Demands, vision, and paths]. *现代教育技术* [*Xiandai Jiaoyu Jishu*], 32(1), 5–15. https://doi.org/10.3969/j.issn.1009-8097.2022.01.001

Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. Physical Review Physics Education Research, 19, 020150. https://doi.org/10.1103/PhysRevPhysEducRes.19.020150

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

Kortemeyer, G., Nöhl, J., & Onishchuk, D. (2024). Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. Physical Review Physics Education Research, 20, 020144. https://doi.org/10.1103/PhysRevPhysEducRes.20.020144

Kortemeyer, G., & Nöhl, J. (2025). Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study. Physical Review Physics Education Research, 21, 010136. https://doi.org/10.1103/PhysRevPhysEducRes.21.010136

Krusberg, Z. A. C. (2007). Emerging technologies in physics education. Journal of Science Education and Technology, 16(5), 401–411. https://doi.org/10.1007/s10956-007-9068-0

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. Science, 208(4450), 1335–1342. https://doi.org/10.1126/science.208.4450.1335

Lepp, M., & Kaimre, J. (2025). Does generative AI help in learning programming: Students' perceptions, reported use and relation to performance. Computers in Human Behavior Reports, 18, 100642. https://doi.org/10.1016/j.chbr.2025.100642

Li, T., Haudek, K., & Krajcik, J. (2025). Utilizing deep learning AI to analyze scientific models: Overcoming challenges. Journal of Science Education and Technology, 34(4), 866–887. https://doi.org/10.1007/s10956-025-10217-0

Lin, H. C. K., Chen, M. C., & Chang, C. K. (2015). Assessing the effectiveness of learning solid geometry by using an augmented reality–assisted learning system. Interactive Learning Environments, 23(6), 799–810. https://doi.org/10.1080/10494820.2013.817435

Liu, D., & Tong, D. (2024). ChatGPT4 解决科学问题能力的探究--以"北上广"三市中考物理试题为例 [A study on ChatGPT-4's ability to solve scientific problems: A case study of the Chinese secondary school entrance examination from Beijing, Shanghai, and Guangzhou]. *中学物理* [*Zhongxue Wuli*], 42(08), 15–19.

Lu, Y., Supriya, K., Shaked, S., Simmons, E. H., & Kusenko, A. (2024). Incentivizing supplemental math assignments and using AI-generated hints improve exam performance, especially for racially minoritized students. arXiv. https://doi.org/10.48550/arXiv.2412.19961

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. American Journal of Physics, 69(S1), S12–S23. https://doi.org/10.1119/1.1371296

Nguyen, H., & Hayward, J. (2025). Applying generative artificial intelligence to critiquing science assessments. Journal of Science Education and Technology, 34(1), 199–214. https://doi.org/10.1007/s10956-024-10177-x

López-Simó, V., & Rezende Junior, M. F. (2024). Challenging ChatGPT with different types of physics education questions. The Physics Teacher, 62(4), 290–294. https://doi.org/10.1119/5.0160160

Šiđanin, P., Plavšić, J., Arsenić, I., & Krmar, M. (2020). Virtual reality (VR) simulation of a nuclear physics laboratory exercise. European Journal of Physics, 41(6), 065802. https://doi.org/10.1088/1361-6404/ab9c90

Kostas, A., Paraschou, V., Spanos, D., Tzortzoglou, F., & Sofos, A. (2025). AI and ChatGPT in higher education: Greek students' perceived practices, benefits, and challenges. Education Sciences, 15(5), 605. https://doi.org/10.3390/educsci15050605

Tao, Y., Wang, Y., & Xing, H. (2022). Zhongxue wuli kecheng shuziziyuan jianjie [Introduction to digital resources for secondary school physics curriculum]. Wuli Jiaoshi (Physics Teacher), 43(11), 75.

Thorp, H. H. (2023). ChatGPT is fun, but not an author. Science, 379(6630), 313. https://doi.org/10.1126/science.adg7879

Wan, T., & Chen, Z. (2024). Exploring generative AI–assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Physical Review Physics Education Research, 20, 010152. https://doi.org/10.1103/PhysRevPhysEducRes.20.010152

Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q. (2024). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. Asia Pacific Education Review, 25(5), 1379–1389. https://doi.org/10.1007/s12564-023-09913-6

Tong, W., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Ithaca: doi:https://doi.org/10.1103/PhysRevPhysEducRes.20.010152

Wattanakasiwich, P., Kaewkhong, K., & Katwibun, D. (2025). Physics instructors' acceptance and implementation of generative AI. Physical Review Physics Education Research, 21, 010155. https://doi.org/10.1103/PhysRevPhysEducRes.21.010155

West, C. G. (2023). AI and the FCI: Can ChatGPT project an understanding of introductory physics? arXiv. https://doi.org/10.48550/arXiv.2303.01067

West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.17012

Xing, H. J. (2009). 自组织表征理论：一种物理问题解决的新理论 [Self-organization representation theory: A new theory of physics problem solving]. 课程·教材·教法 [Curriculum, Teaching Material and Method], 29(4), 60–64. https://doi.org/10.19877/j.cnki.kcjcjf.2009.04.013

Xing, H. J., & Chen, Q. M. (2005). 论原始物理问题的教育价值及其启示 [On the educational value of primitive physics problems and its implications]. 课程·教材·教法 [Curriculum, Teaching Material and Method], (1), 56–61.https://doi.org/10.19877/j.cnki.kcjcjf.2005.01.017

Xing, H. J., & Chen, Q. M. (2008). 原始物理问题测量工具：编制与研究 [Measurement tools for primitive physics problems: Development and research]. 课程·教材·教法 [Curriculum, Teaching Material and Method], (11), 59–63.

Xing, H. J., Shi, Y., Hu, Y. Y., & Tian, H. X. (2015). 初中原始物理问题测量工具：编制与研究 [Measurement tools for junior high school primitive physics problems: Development and research]. 课程·教材·教

法[Curriculum, Teaching Material and Method], 35(2), 69–73. https://doi.org/10.19877/j.cnki.kcjcjf.2015.02.010

Zeng, Y., & Tong, D. (2024). ChatGPT4 解决科学问题能力的研究--以高考全国物理卷为例 [A study on ChatGPT-4's ability to solve scientific problems: A case study of the national college entrance examination physics papers]. *中学物理* [*Zhongxue Wuli*], 42(5), 22–27.

Zhai, X. (2023). ChatGPT user experience: Implications for education. SSRN. https://doi.org/10.2139/ssrn.4312418