education sciences

MDPI

Article

Research on Intelligent Grading of Physics Problems Based on Large Language Models

Yuhao Wei [1], Rui Zhang [2,*], Jianwei Zhang [2], Dizhi Qi [2] and Wenqian Cui [2]

[1] Guohao College, Tongji University, Shanghai 200082, China; 2353242@tongji.edu.cn
[2] School of Physics Science and Engineering, Tongji University, Shanghai 200082, China; zhang@tongji.edu.cn (J.Z.); 2432907@tongji.edu.cn (D.Q.); 21666514@tongji.edu.cn (W.C.)
* Correspondence: 98733@tongji.edu.cn; Tel.: +86-021-6598-3380

Abstract: The automation of educational and instructional assessment plays a crucial role in enhancing the quality of teaching management. In physics education, calculation problems with intricate problem-solving ideas pose challenges to the intelligent grading of tests. This study explores the automatic grading of physics problems through a combination of large language models and prompt engineering. By comparing the performance of four prompt strategies (one-shot, few-shot, chain of thought, tree of thought) within two large model frameworks, namely ERNIEBot-4-turbo and GPT-4o. This study finds that the tree of thought prompt can better assess calculation problems with complex ideas (N = 100, ACC $\geq$ 0.9, kappa > 0.8) and reduce the performance gap between different models. This research provides valuable insights for the automation of assessments in physics education.

Keywords: automatic grading; large language models; prompt engineering; tree-of-thought prompt; physics education

check for updates

Academic Editor: Federico Corni

Received: 3 November 2024
Revised: 9 January 2025
Accepted: 15 January 2025
Published: 21 January 2025

Citation: Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W. (2025). Research on Intelligent Grading of Physics Problems Based on Large Language Models. *Education Sciences*, *15*(2), 116. https://doi.org/10.3390/educsci15020116

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The use of artificial intelligence in education can provide insightful technical support for scientific and efficient educational governance. Intelligent grading of student tests and assignments not only reduces teachers' workloads but also swiftly delivers student learning feedback data to both students and administrators. On one hand, this aids students in better reflecting on their learning and promptly adjusting their learning strategies. On the other hand, educational administrators can quickly identify potential issues in teaching and learning, make timely educational decisions, and enhance the quality of education and instruction. In recent years, the rapid development of generative artificial intelligence, especially large language models, has created a favorable opportunity for AI to assist in educational governance.

Large language models (LLMs) are deep learning models pre-trained on vast datasets and characterized by millions or more parameters. At the core of these models is transformer architecture, a neural network structure consisting of encoders and decoders with self-attention capabilities. The transformer model enables unsupervised learning from input data and processes entire data sequences in parallel (Vaswani et al., 2023). Among LLMs, the most prominent is the generative pre-trained transformer (GPT) developed by OpenAI. Since the launch of ChatGPT on 30 November 2022, the impressive text comprehension and generation abilities of LLMs have become widely recognized.

LLMs, particularly GPT-based systems, have emerged as transformative tools in education, excelling in content comprehension, generation, personalized learning, and automating complex educational tasks. In STEM education, LLMs have the potential

to revolutionize assessments by generating accurate and conceptually rich questions in subjects like physics, chemistry, and mathematics, utilizing advanced prompt engineering techniques, such as chain of thought strategies (Chan, 2025). Likewise, GPT-4 has exhibited exceptional performance, attaining pass rates that surpass those of other models on licensing exams across diverse linguistic and geographical contexts (Zong et al., 2024). GPT-powered systems such as EduCS (Song, 2023) and MCQGen (Hang et al., 2024) have demonstrated proficiency in generating high-quality multiple-choice questions (MCQs) for computer science curricula, leading to a substantial reduction in instructor workload. Chatbot-based systems like NewtBot, powered by GPT-3.5, have further enhanced physics education by acting as automated tutors, providing personalized feedback, and fostering student engagement (Lieb, 2024). Beyond content creation, GPT models have advanced physics education through automated problem-solving and evaluation tools that bridge theoretical knowledge with practical applications (Shamshin, 2024), and have demonstrated high consistency with human evaluations in assessing conceptual short-answer questions (Yan et al., 2024).

In addition to GPT, other LLMs have demonstrated remarkable reasoning and comprehension capabilities, such as enhanced representation through knowledge integration (ERNIE) and Claude-3-sonnet, which were developed by Baidu and Anthropic, respectively.

ERNIEBot has become a prominent LLM within China due to its local accessibility and advanced capabilities in processing Chinese language tasks. ERNIEBot has exhibited strong comprehension and generation abilities, making it a widely utilized model for educational purposes. It has enhanced physics learning through knowledge graphs and text classification, improving concept comprehension, and aiding students in understanding relationships between scientific principles (Shang et al., 2022). ERNIEBot has demonstrated exceptional performance in education, consistently surpassing GPT-4 in tasks like the Chinese National Medical Licensing Examination (Huang et al., 2024) and achieving 100% accuracy in identifying health-related rumors while providing culturally appropriate explanations (Luo et al., 2024). Furthermore, it facilitates the development of customized course recommendation systems, such as ICRA, enhancing both the precision of these systems and the user experience on MOOC platforms (Li et al., 2024).

Claude is an LLM that has garnered attention for its sophisticated reasoning and ethical considerations. This model presents a compelling alternative to GPT in educational applications. In the context of language education, Claude-2 has undergone extensive evaluation concerning its automated essay scoring capabilities. It has been demonstrated to possess excellent reliability and validity in the assessment of English language learner essays, comparable to GPT-4 (Pack et al., 2024). In the field of computing education, Claude has been evaluated based on concept inventories, demonstrating its adaptability and strong performance in evaluating students' understanding of computational concepts (Ali et al., 2024). Furthermore, it outperformed GPT in identifying legal arguments from judges' decisions, achieving accuracy rates of up to 90% when replicating nuanced legal reasoning (Burgess et al., 2024). Similarly, in robotics education, Claude excelled in generating context-aware code for human–robot interaction tasks, achieving a 95% success rate—compared to lower rates by other LLMs, including GPT-3.5—and highlighting its suitability for programming education (Sobo et al., 2024).

In the realm of physics education, LLMs, particularly GPT, have demonstrated a strong grasp of physics concepts and can successfully solve some problems. For example, GPT-4 was evaluated using the force concept inventory (FCI), performing at the level of human experts, with scores ranging between 90% and 100% (West, 2023). In kinematics, GPT-4 performed comparably to high school students, providing mostly correct answers with accurate textual descriptions of graphs (Polverini & Gregorcic, 2024).

However, LLMs exhibit significant limitations when tasked with more complex problems. GPT-4, for instance, was evaluated on 40 problems from a college-level engineering physics course. While it successfully solved 62.5% of the well-specified problems, its accuracy dropped to just 8.3% for under-specified problems (K. D. Wang et al., 2024). On the AP Physics 2 exam, GPT-3.5 scored around 30%, while GPT-4 managed just over 60% (OpenAI et al., 2024). Additionally, GPT-3.5 and GPT-4 scored 49.4% and 38.6%, respectively, on Durham University's physics exams (2018–2022), which places them on par with weaker students (Yeadon & Halliday, 2023). In a college-level introductory physics course, ChatGPT was predicted to receive a grade of 1.5, which is sufficient for course credit but below the threshold required for earning a bachelor's degree (Kortemeyer, 2023a). These findings highlight that LLMs' performance is constrained by their limited ability to fully grasp and reason through complex physics problems.

Prompt engineering is a technique used to enhance LLM performance by providing additional context or task-specific instructions without altering the underlying model parameters. Compared to fine-tuning, prompt engineering is cost-effective and relatively simple to implement. Today, there are many types of prompts. For new tasks where extensive training is not feasible, zero-shot prompts provide only instructions to guide the LLM's response (Radford et al., 2019), while few-shot prompts provide examples for models of learning to learn (Brown et al., 2020). One of the most widely used prompt techniques is the chain of thought (CoT) prompt, which guides the model to produce a series of logically connected outputs, significantly enhancing LLMs' reasoning abilities (Wei et al., 2023). A simpler variant, the automatic chain of thought (Auto-CoT) prompt uses the phrase "let's think step by step" to encourage sequential reasoning without complex instructions (Zhang et al., 2022).

Several CoT variants have since emerged, including tree of thought (ToT), graph of thought (GoT), thread of thought (ThoT), etc. The ToT prompt enables LLMs to make deliberate decisions by evaluating multiple reasoning paths, self-evaluating choices, and backtracking or looking ahead when necessary (S. Yao et al., 2023). The GoT prompt captures non-linear jumps in reasoning by linking thought units through a structured node-edge relationship (Y. Yao et al., 2024). The ThoT prompt segments inputs into manageable parts, allowing LLMs to filter out irrelevant information and concentrate on key details (Zhou et al., 2023).

In education, timely and accurate feedback is essential for students to adjust their learning progress, but scoring assignments can be time-consuming for instructors. LLMs, with the aid of prompt engineering, offer a promising solution to reduce this workload and provide timely, objective feedback. GPT-4 demonstrated a level of accuracy comparable to that of human experts in the evaluation of clinical notes (Burke et al., 2024), while models such as GPT-4 and Gemini Pro exhibited consistency and precision in the grading of short-answer assignments for undergraduate courses in the field of medical education (Grévisse, 2024). Rubric-based methods further enhanced grading of long answer questions, capturing nuanced responses more effectively than traditional methods (Sonkar et al., 2024). Furthermore, GPT-4 and Gemini have been employed to streamline college examination grading (J. X. Lee & Song, 2024), while GPT-4o has demonstrated reliability in the evaluation of practical skills such as CPR (L. Wang, 2024). In computer science education, models such as GPT-4 and Gemini Pro have exhibited strong alignment with human grading, accurately assessing criteria such as correctness, style, and adherence to best practices (Mohamed, 2025). Further, integrating LLMs into autograding systems has enabled the generation of comprehensive test suites for introductory programming courses, uncovering ambiguities in problem statements and offering detailed evaluation frameworks (Alkafaween et al., 2025). Beyond conventional grading, LLMs have introduced

innovations such as Code Generation Based Grading (CGBG), which transforms natural language responses into executable code for evaluation, and created comprehensive test suites that surpass instructor-designed ones in identifying edge cases (Smith & Zilles, 2024). These technologies also address large-scale assessment challenges in MOOCs (Lagakis, 2024) and foster collaboration in project-based learning environments through tools like chatbot-driven "standups", enhancing both efficiency and student engagement (Menezes et al., 2024).

Prompt engineering has demonstrated considerable efficacy in enhancing the performance of LLMs for automated scoring. Specifically, the CoT prompt has demonstrated considerable potential for improving the performance of LLMs for automatic assessment tasks in physics. For example, GPT-4 achieved an R-squared correlation of 0.84 with human graders when assessing synthetic physics problems, making it a viable tool for formative assessments (Kortemeyer, 2023b). Additionally, with a few-shot prompt, GPT-3.5 was used to generate feedback on students' written responses to conceptual physics questions, and approximately 70% of the AI-generated feedback required minimal or no modification before being shared with students (Wan & Chen, 2024).

In other areas of education, the use of prompts in LLM-based scoring has also been explored. For instance, GPT-4 was employed to assess three explanation questions in middle school Earth science, utilizing a combination of a CoT prompt and active learning to achieve classification accuracies (ACCs) of no less than 0.80 (Cohn et al., 2024). In another study, six prompt strategies were used to guide GPT-4 in grading students' explanations of physical and chemical phenomena. CoT prompts, combined with item stems and scoring rubrics, significantly improved the accuracy of results, with a 13.44% and 3.7% increase compared to zero-shot and few-shot prompts, respectively (G.-G. Lee et al., 2024). In the field of mathematics, GPT-4 demonstrated proficiency in assessing open-ended questions, with prompt engineering such as CoT, self-consistency and ToT prompts enhancing its handling of complex reasoning (U. Lee et al., 2024).

Most studies so far have focused on short-answer or explanation-based questions, which typically have more fixed responses. However, many physics problems require students to present detailed calculation processes and may have multiple solution methods. To improve LLM performance in such scenarios, we have designed specialized prompts. While traditional CoT prompts follow a single chain of reasoning, we will also explore the effects of ToT prompts. Based on the above considerations, we propose the following research questions: (1) How effective are LLMs with prompts in grading complex physics calculation problems? (2) How do different prompt types influence the effectiveness of scoring physics calculation problems?

2. Materials and Methods

2.1. The Generation of Samples

To conduct a comprehensive study, we selected six college-level physics questions, each requiring students to provide a detailed solution process. These questions covered various topics, including mechanics, electromagnetism, and modern physics. All questions were adapted from homework exercises. Each question could be solved using at least two distinct methods. The specifics of these questions are presented in Table 1. To ensure the accuracy of the reference answers, we consulted an expert in physics education, who reviewed and validated them.

Table 1. Description of chosen problems.

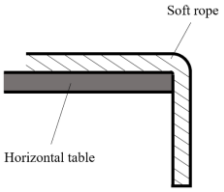
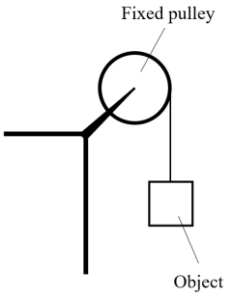
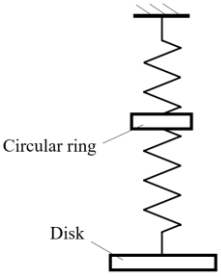
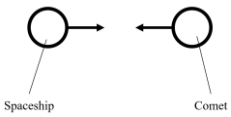
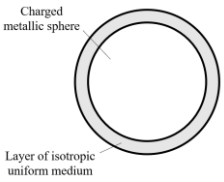
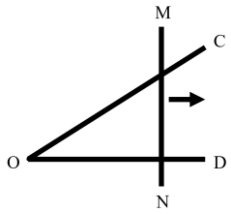
Problem Number	Knowledge Concept	Problem	Schematic Diagram
Q1	Newton's Second Law, Work-Energy Theorem	A soft rope with uniform thickness is partially placed on a smooth horizontal table, while the other part hangs off the table's edge. There is a small circular corner with a very small radius between two perpendicular surfaces at the table's edge, allowing the rope's velocity direction to change continuously as it passes over the corner. The total length of the rope is L, and the hanging part is initially b long. The rope's initial velocity is zero. Find the velocity of the rope at the moment it completely leaves the table (assume the rope does not stretch).	
Q2	Law of Rigid Body Rotation, Conservation of Mechanical Energy	A fixed pulley has a mass of $M = 2.0$ kg, a radius of $R = 0.10$ m, and a non-stretchable light rope. One end of the rope is fixed to the pulley, while the other end is attached to an object of mass $m = 5.0$ kg. If the moment of inertia of the pulley is $\frac{1}{2}MR^2$, its initial angular velocity $\omega_0 = 10.0$ rad/s (counterclockwise), and $g = 10\ \mathrm{m \cdot s^{-2}}$, find: (1) the magnitude and direction of the angular acceleration of the pulley; (2) the maximum height the object reaches.	
Q3	Simple Harmonic Motion, Conservation of Mechanical Energy	A disk of mass M is hanging from a light spring with a stiffness coefficient of k. A thin circular ring of mass m is dropped freely from a height h above the disk and sticks to the disk, causing them to vibrate together. Let $t = 0$ be the moment of impact, and the coordinate origin be chosen at the equilibrium position of the vibration, with the positive direction downward. Write the expression for the vibration.	
Q4	Time Dilation Effect, Lorentz Transformation	A spaceship and a comet are moving towards each other, with speeds of 0.6 c and 0.8 c relative to the ground, respectively. From the ground's perspective, they will collide in 5 s. Questions: (1) What is the velocity of the comet as seen from the spaceship? (2) How much time will pass on the spaceship's clock before the collision?	
Q5	Gauss's Law in a Medium, Gauss's Law in a Vacuum, Superposition of Electric Fields	A charged metallic sphere with a charge of Q_0 and radius R_1 is surrounded by a layer of isotropic uniform medium with a relative permittivity ε_r, extending to a radius R_2. Find the magnitude of the electric field E inside the medium and the surface charge density on the inner surface of the medium.	

Table 1. *Cont.*

Problem Number	Knowledge Concept	Problem	Schematic Diagram
Q6	Faraday's Law of Electromagnetic Induction, Motional EMF, Induced EMF	A metal frame COD, bent at an angle θ, lies in a plane perpendicular to a magnetic field. A conductor rod MN moves to the right with a constant velocity v along the frame and is perpendicular to side OD. The velocity v is also perpendicular to MN. At $t = 0$, $x = 0$. The magnetic field is uniform and $B = B_0 \cos(\omega t)$. Find the magnitude of the induced EMF ε_i in the frame.	

We chose Claude-3-sonnet to ensure the balanced distribution of sample responses, as the accuracy rate of students' solutions is relatively high in our university, which is not ideal for evaluating the performance of the LLMs. To generate samples, we designed CoT prompts based on the problem answers and set the temperature to the default value. In addition, to simulate a diverse range of student errors, one expert in physics education identified common mistakes students might make, and we created corresponding CoT prompts to reflect these errors. The CoT prompts template for sample generation is illustrated in Table 2. Claude-3-sonnet then generated 100 sample answers for each question, with 50% of the samples being correct (positive) and the other 50% incorrect (negative). Each sample was generated independently to ensure that outputs did not influence one another.

Table 2. Templates of the CoT prompts for sample generation.

Kind of Samples	Negative Samples	Positive Samples
Prompt templates	The following is a college-level physics problem: "<Problem>" A reference answer in LaTeX format is as follows: "<Reference Answer>" The thought process of the sample solution is as follows: "<CoT of Reference Answer>" <Set Errors and Reasons> The following is his thought process: "<Erroneous CoT>" Please mimic the student's thought process to provide the complete solution to this problem in LaTeX format.	The following is a college-level physics problem: "<Problem>" A reference answer in LaTeX format is as follows: "<Reference Answer>" The thought process of the sample solution is as follows: "<CoT of Reference Answer>" Student's thought process is the same as reference answer. Please mimic the student's thought process to provide the complete solution to this problem in LaTeX format.

For each question, correct sample answers were generated based on the solution ideas of the two reference answers. The causes of errors for incorrect samples are listed in Table 3. To illustrate this process, we will focus on Q1. Two types of correct answer samples and three types of incorrect answer samples were designed based on recommendations from physics experts. The derivation of the correct answer samples involved two approaches: one employed Newton's second law to calculate the rope's acceleration expression, followed by integration to determine the final velocity; the other applied the conservation of energy principle, converting gravitational potential energy into kinetic energy to solve for the final velocity. The three incorrect approaches simulated common student mistakes: one modeled students who misunderstood the need for integration and directly averaged the initial and final accelerations to compute the result; another considered only the vertical hanging part of the rope, neglecting the horizontal section on the table; and the last focused solely on the horizontal section, ignoring the vertical hanging part. The CoT prompts for incorrect samples altered the corresponding steps of the CoT prompts for correct samples.

Table 3. Causes of errors for the incorrect samples of problems.

Problem Number	Causes of Errors for Incorrect Samples
Q1	The initial and final accelerations are to be averaged directly for computation. The vertical hanging part of the rope is to be considered, while the horizontal section on the table is to be neglected. The horizontal section on the table is to be considered, while the vertical hanging part is to be neglected.
Q2	The treatment of the tension in the rope is equivalent to the object's gravity. Both the misjudgment of the direction of velocity and the assumption that the direction of acceleration is equivalent to the direction of the fixed pulley's motion constitute primary errors. The rotational kinetic energy of the pulley is neglected while calculating mechanical energy, with only the object's kinetic energy being considered.
Q3	The collision was elastic; energy loss during the collision was neglected. The free end of the spring was mistaken for the equilibrium point. Computational errors were made in calculating the phase or angular frequency.
Q4	The ground reference frame was confused with the spaceship reference frame. It was assumed that the situation observed simultaneously on the ground for the spaceship and the comet also happens simultaneously in the spaceship reference frame.
Q5	The medium was treated as a vacuum. The relative permittivity was mistaken for the absolute permittivity. The medium was assumed to be a conductor. The electric displacement field was confused with the electric field strength.
Q6	The induced electromotive force was confused with the motional electromotive force. The motional electromotive force was mistaken for the induced electromotive force. Errors were made in the time derivative of magnetic flux, and the time variation in magnetic field intensity was neglected.

2.2. Scoring Criteria

The evaluation framework employed for this research study encompassed two primary criteria: the accuracy of the value and the validity of the underlying methodology. The validity of a sample is contingent upon the accuracy of both its value and its process. The validity of a prediction is contingent upon its congruence with the actual outcome, both in terms of value and process. With respect to the evaluation of the process, a correct score was assigned when the sample utilized the appropriate formula or theorem. Conversely, if the scoring for either the value or the process was found to be incorrect, the prediction was deemed to be incorrect.

2.3. Scoring Process

During the evaluation processes, LLMs were tasked with performing binary classification on both the calculation result and processing of the samples, guided by four types of prompts. The same part of the four prompts is shown in Figure 1, and differences among these four prompts are illustrated in Table 4. In all cases, the LLMs were assigned the role of a college physics professor, provided with the problem description, and instructed to output the result in JSON format. For one-shot and few-shot prompts, the LLMs were given one or two reference answers as examples before generating their output. For CoT and ToT prompts, the LLMs were guided to think step by step rather than providing an immediate response.

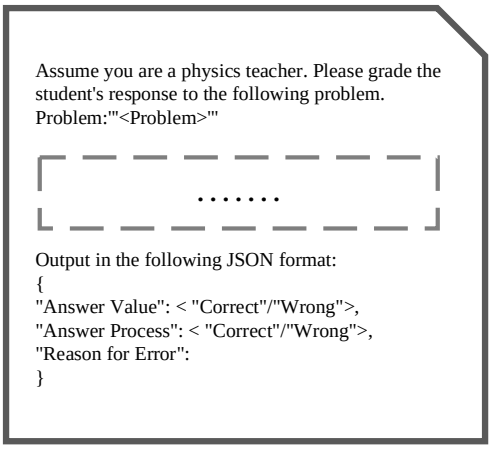


Figure 1. The same part of four prompts.

When evaluating the value, the task was primarily focused on whether the sample's numerical or symbolic answers matched the given answer, relying heavily on the LLM's ability to extract text accurately. In ToT prompts, we applied the depth-first search algorithm, commonly used in tree structures. First, the LLMs were instructed to search for key steps, such as formulas or intermediate results, for one of the solution methods. If these key points were found, they were compared with the reference answer. If not, the LLMs would proceed to search for the key steps of the second method. Finally, the LLMs would output the scoring results. In contrast, CoT prompts involved only one method, so there was no need for a depth-first search.

Table 4. The different parts in the four prompts.

Prompt	Template
One-Shot	The LaTeX text for the reference answer is: "'<Reference Answer 1>"' The student's response is as follows: "'<Sample>"' Please evaluate the student's answer by considering the reference answer.
Few-Shot	The LaTeX text for the "Solution 1" reference answer is: "'<Reference Answer 1>"' The LaTeX text for the "Solution 2" reference answer is: "'<Reference Answer 2>"' The student's response is as follows: "'<Sample>"' Please evaluate the student's answer by considering both reference answers.
CoT	The LaTeX text for the reference answer is: "'<Reference Answer 1>"' The student's response is as follows: "'<Sample>"' Please evaluate the student's answer by considering the reference answer, and think through the following steps: "'<CoT Prompt>"'
ToT	The LaTeX text for the "Solution 1" reference answer is: "'<Reference Answer 1>"' The LaTeX text for the "Solution 2" reference answer is: "'<Reference Answer 2>"' The student's response is as follows: "'<Sample>"' Please evaluate the student's answer by considering both reference answers, and think through the following steps: "'<ToT Prompt>"'

2.4. Evaluation

For evaluation, we used accuracy (ACC) to assess classification performance and the kappa coefficient to measure agreement between predictions and actual results. The formula for ACC is:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$$

Here, T and F indicate whether the LLM's prediction is classified as positive or negative, respectively, and P and N indicate whether the actual result is positive or negative, respectively. The kappa coefficient is calculated as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Here, p_0 is the observed agreement, or the proportion of items where both raters agree, and p_e is the expected agreement, representing chance-based agreement.

3. Results

3.1. ERNIEBot-4-Turbo's Grading Effect

In this experiment, we used ERNIEBot-4-turbo to score the sample answers using four different prompts. The ACC and kappa coefficients for each prompt are presented in Table 5 and Figure 2. The highest ACC and kappa values were both 1.00, indicating that ERNIEBot-4-turbo can accurately score some questions with specific prompts. For each question, the responding highest ACC and kappa was no less than 0.90 and 0.80, respectively. Consequently, we posit that ER-NIEBot-4-turbo demonstrates a high level of proficiency in addressing complex physics calculation problems.

Table 5. ACCs and kappa coefficients of different prompts.

Question	Prompt	ACC	Kappa
Q1	One-Shot	0.84	0.68
	Few-Shot	0.87	0.74
	CoT	0.81	0.62
	ToT	0.97	0.94
Q2	One-Shot	0.52	0.04
	Few-Shot	0.62	0.24
	CoT	0.68	0.36
	ToT	0.99	0.98
Q3	One-Shot	0.79	0.58
	Few-Shot	0.82	0.64
	CoT	0.93	0.84
	ToT	0.98	0.96
Q4	One-Shot	0.86	0.72
	Few-Shot	0.95	0.90
	CoT	0.97	0.94
	ToT	1.00	1.00
Q5	One-Shot	0.90	0.80
	Few-Shot	0.93	0.86
	CoT	0.95	0.90
	ToT	0.91	0.82

Table 5. *Cont.*

Question	Prompt	ACC	Kappa
Q6	One-Shot	0.82	0.64
	Few-Shot	0.84	0.68
	CoT	0.83	0.66
	ToT	0.90	0.80

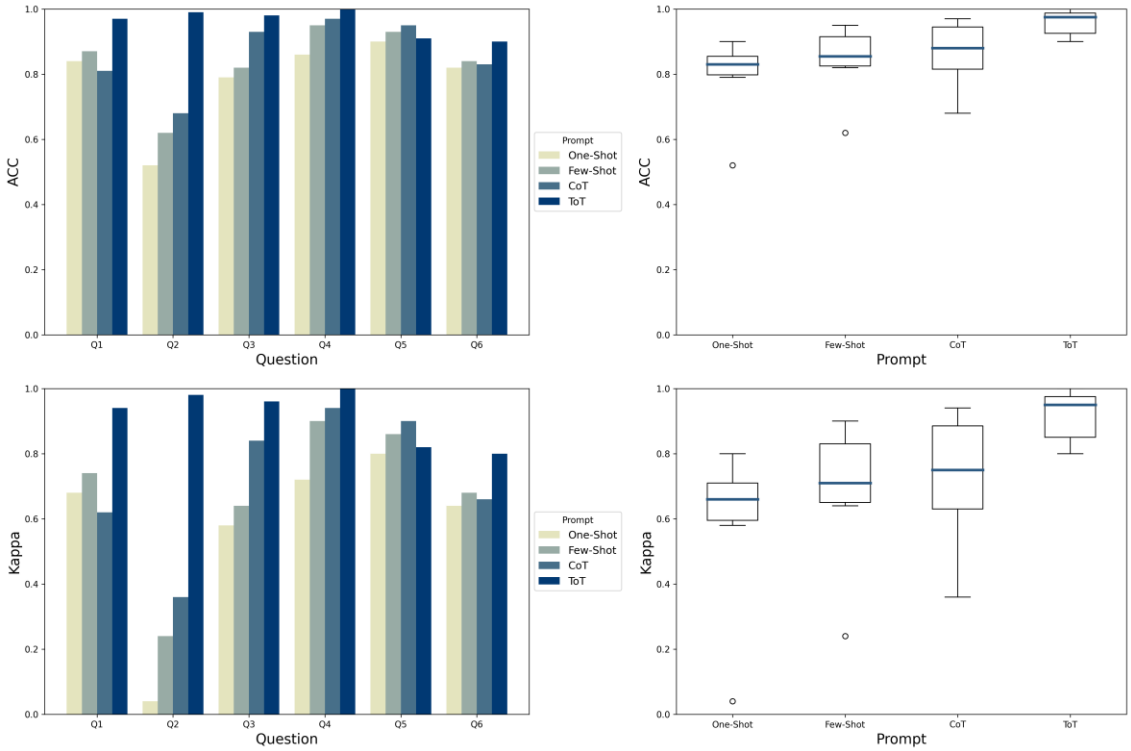


Figure 2. ACCs and kappa coefficients of different prompts.

3.2. Comparison of the Effect of Different Prompts

As demonstrated in Table 5 and Figure 2, the Tot prompt had the highest ACC and kappa coefficient across all four prompts, achieving the best performance in five out of six questions. This demonstrates that the ToT prompt is superior in terms of both classification accuracy and consistency. Regardless of the specific question, the ACC for ToT prompts remained no less than 90%, showcasing its robust classification ability. Similarly, the kappa coefficient for the ToT prompt exceeded 0.80 in all cases, indicating a high level of agreement between predictions and the ground truth.

Conversely, the one-shot prompt had the lowest ACC and kappa scores, indicating the weakest classification ability and consistency. This is likely because the one-shot prompt provides the least amount of contextual information, resulting in limited improvement in the LLM's reasoning capability compared to CoT or ToT prompts.

The performance of the one-shot prompt, few-shot prompt, and CoT prompt also changed more across different questions, as shown in Table 6. The variance in ACC and kappa coefficients for these prompts is nearly an order of magnitude greater than for the ToT prompt. For instance, in Q2, the performance of these three prompts is noticeably worse than that of ToT, while in other questions, the gap between these three and the ToT prompt narrows.

Table 6. Variance of different prompts.

Prompt	Variance of ACC	Variance of Kappa
One-Shot	0.0155	0.0622
Few-Shot	0.0116	0.0466
CoT	0.0101	0.0397
ToT	0.0015	0.0061

4. Discussion

4.1. Comparison of the Effect of Prompts on GPT-4o and ERNIEBot-4-Turbo

We also compared the performances of ERNIEBot-4-turbo and GPT-4o on Q6 using the four prompts separately. The ACC and kappa coefficients for each prompt are shown in Table 7. Overall, the ACC across all experiments exceeded 80%, and the kappa coefficients were consistently above 0.60, regardless of the model or prompt used. This indicates that all combinations yielded good classification performance and moderate agreement.

Across all prompts, GPT-4o outperformed ERNIEBot-4-turbo, demonstrating superior performance in this task. In particular, CoT and ToT prompts had a more pronounced effect on improving ERNIEBot-4-turbo's performance. The results suggest two distinct performance classes: one-shot and few-shot prompts show a larger gap between the two models, while the CoT and ToT prompts narrow the performance difference. Given that both CoT (Wei et al., 2023) and ToT (S. Yao et al., 2023) are designed to enhance LLMs' reasoning abilities, we believe these prompts improve ERNIEBot-4-turbo's reasoning more significantly compared to GPT-4o, which already has stronger reasoning capabilities.

Table 7. The ACCs and kappa coefficients of different models.

Prompt	ACC		Relative Difference in ACC	Kappa		Relative Difference in Kappa
	ERNIEBot	GPT		ERNIEBot	GPT	
One-Shot	0.82	0.89	0.09	0.64	0.78	0.22
Few-Shot	0.84	0.92	0.10	0.68	0.84	0.24
CoT	0.83	0.85	0.02	0.66	0.7	0.06
ToT	0.9	0.93	0.03	0.8	0.86	0.07

4.2. How Effective Are LLM Prompts in Grading Complex Physics Calculation Problems?

The results from these experiments suggest that LLMs, particularly when guided by prompts, are well-suited for scoring tasks in physics education. This demonstrates the significant potential of LLMs in automated scoring, largely attributed to the power of both the models and the prompts used. However, given the importance of grading accuracy for students' academic performance and the fact that there is still room for improvement, further research and refinement are necessary.

4.3. How Do Different Prompt Types Influence the Effectiveness of Scoring Physics Calculation Problems?

Although the ToT prompt achieved the best overall performance, its superiority over the one-shot prompt in Q5 was limited, and it was outperformed by the few-shot and CoT prompts. Q5 involved calculating electric field strength and charge density in a medium, which could be approached either through Gauss's law in a vacuum or Gauss's law directly in the medium. The equations for Gauss's law in a vacuum and Gauss's law in the medium are almost the same, which misled ERNIEBot-4-turbo's grading.

In some cases, few-shot outperforms CoT, such as in Q1 and Q6, while in other cases, CoT outperforms few-shot. This variability is influenced by the specific characteristics of the two solution methods for each question. When the two methods share similar key formulas, LLMs might mistakenly interpret these formulas as identical, causing few-shot to perform worse than CoT, as CoT improves the LLM's reasoning ability. For example, Q3 involves sinusoidal oscillation, and although the two solution methods are largely similar (one analyzing momentum and the other energy), few-shot outperforms CoT in this case because of the LLM's ability to learn from examples when the solution methods diverge significantly.

4.4. Limitations and Future Outlook

The samples in this study are limited to six college-level physics questions. While these six questions represent key topics in college physics, they are not exhaustive. Additionally, only two models, ERNIEBot-4-turbo and GPT-4o, were tested during a specific time period (July 2024). As LLMs continue to evolve, the effectiveness of various prompts may also change in the future.

Furthermore, applying this method directly to education on a large scale is currently unrealistic. Each ToT prompt was manually designed, and constructing a database of questions paired with corresponding ToT prompts would be highly time-consuming and labor-intensive. Given the large number of potential physics questions and the continuous creation of new problems, it is impractical to rely on manual prompt creation. Moreover, the quality of prompts can vary significantly depending on the experience and expertise of the designers. Although prompts can enhance LLM performance, their effectiveness can only be fully evaluated and improved through iterative testing, as demonstrated in this study.

Future research should explore the possibility of instructing LLMs to generate ToT prompts automatically. LLMs could be tasked with assessing the quality of these prompts and going through multiple iterations of them until the desired outcomes are achieved, without relying on human feedback, through techniques like prompt engineering or training agents (Radha et al., 2024). This automated process could allow for more efficient and accurate prompt generation through continuous feedback from the LLM itself. Additionally, self-consistency prompts could be incorporated into the scoring process alongside CoT or ToT prompts. Using self-consistency prompts involves asking LLMs to perform a task multiple times and then comparing the results across iterations to improve prediction accuracy (X. Wang et al., 2023).

5. Conclusions

This study demonstrates the potential of LLMs for use in the automated grading of physics calculation problems, particularly when guided by prompts. Among the four evaluated prompts, the ToT prompt achieved the highest accuracy and consistency overall, significantly enhancing the performance of the LLMs in grading compared with other prompts. While GPT-4o demonstrated superior performance to ERNIEBot-4-turbo in this scoring task, the use of CoT and ToT prompts can narrow the gap between the two models. Although the samples used in this study came from physics courses, the methods adopted here can also be helpful for other courses. With the development of LLMs, educational artificial intelligence can pave the way for better educational governance.

Author Contributions: Conceptualization, R.Z. and Y.W.; methodology, R.Z.; validation, Y.W. and D.Q.; formal analysis, Y.W.; investigation, Y.W.; data curation, Y.W. and D.Q.; writing—original draft preparation, Y.W.; writing—review and editing, W.C.; visualization, Y.W.; project administration, R.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tongji University grant number 1370104330.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the reported results are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ali, M., Mai, Y., Rao, P., & Xie, B. (2024, August 13–15). *Using benchmarking infrastructure to evaluate LLM performance on CS concept inventories: Challenges, opportunities, and critiques*. ACM Conference on International Computing Education Research (ICER), Melbourne, Australia. [CrossRef]

Alkafaween, U., Albluwi, I., & Denny, P. (2025). Automating autograding: Large language models as test suite generators for introductory programming. *Journal of Computer Assisted Learning*, *41*(1), e13100. [CrossRef]

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *arXiv*, arXiv:2005.14165. [CrossRef]

Burgess, P., Williams, I., Qu, L., & Wang, W. (2024). Using generative AI to identify arguments in judges' reasons: Accuracy and benefits for students. *Law, Technology and Humans*, *6*, 5. [CrossRef]

Burke, H. B., Hoang, A., Lopreiato, J. O., King, H., Hemmer, P., Montgomery, M., & Gagarin, V. (2024). Assessing the ability of a large language model to score free-text medical student clinical notes: Quantitative study. *JMIR Medical Education*, *10*, e56342. [CrossRef]

Chan, K. W. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education*, *8*, 100344. [CrossRef]

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 23182–23190. [CrossRef]

Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, *24*(1), 1060. [CrossRef] [PubMed]

Hang, C. N., Tan, C. W., & Yu, P.-D. (2024). MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Education Society Section*, *12*, 102261–102273. [CrossRef]

Huang, L., Hu, J., Cai, Q., Fu, G., Bai, Z., Liu, Y., Zheng, J., & Meng, Z. (2024). The performance evaluation of artificial intelligence ERNIE bot in Chinese national medical licensing examination. *Postgraduate Medical Journal*, *100*, qgae062. [CrossRef]

Kortemeyer, G. (2023a). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, *19*(1), 010132. [CrossRef]

Kortemeyer, G. (2023b). Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research*, *19*(2), 020163. [CrossRef]

Lagakis, P. (2024). *Automated grading in coding exercises using large language models*. Springer.

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. [CrossRef]

Lee, J. X., & Song, Y.-T. (2024, July 5–7). *College exam grader using LLM AI models*. 2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 282–289), Beijing, China. [CrossRef]

Lee, U., Kim, Y., Lee, S., Park, J., Mun, J., Lee, E., Kim, H., Lim, C., & Yoo, Y. J. (2024). Can we use GPT-4 as a mathematics evaluator in education?: Exploring the efficacy and limitation of LLM-based automatic assessment system for open-ended mathematics question. *International Journal of Artificial Intelligence in Education*. [CrossRef]

Li, B., Hou, Y., Dong, J., Yang, B., & Wang, X. (2024). ICRA: A study of highly accurate course recommendation models incorporating false review filtering and ERNIE 3.0. *PLoS ONE*, *19*(12), e0313928. [CrossRef]

Lieb, A. (2024, May 11–16). *Student interaction with newtbot: An LLM-as-tutor chatbot for secondary physics education*. CHI EA'24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA.

Luo, Y., Miao, Y., Zhao, Y., Li, J., Chen, Y., Yue, Y., & Wu, Y. (2024). Comparing the Accuracy of Two Generated Large Language Models in Identifying Health-Related Rumors or Misconceptions and the Applicability in Health Science Popularization: Proof-of-Concept Study. *JMIR Formative Research*, *8*, e63188. [CrossRef] [PubMed]

Menezes, T., Egherman, L., & Garg, N. (2024, July 8–10). *AI-grading standup updates to improve project-based learning outcomes*. ITiCSE 2024: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V.1, Milan, Italy.

Mohamed, K. (2025). Hands-on analysis of using large language models for the auto evaluation of programming assignments. *Information Systems*, *128*, 102473. [CrossRef]

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2024). GPT-4 technical report. *arXiv*, arXiv:2303.08774. [CrossRef]

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. [CrossRef]

Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the test of understanding graphs in kinematics. *Physical Review Physics Education Research*, *20*(1), 010109. [CrossRef]

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radha, S. K., Jelyani, Y. N., Ghukasyan, A., & Goktas, O. (2024). Iteration of thought: Leveraging inner dialogue for autonomous large language model reasoning. *arXiv*, arXiv:2409.12618. [CrossRef]

Shamshin, A. (2024). Implementation of artificial intelligence in an online platform for solving and checking the solutions of physical problems. In *Information technology for education, science, and technics. ITEST 2024. Lecture notes on data engineering and communications technologies* (Vol. 222). Springer. [CrossRef]

Shang, J., Huang, J., Zeng, S., Zhang, J., & Wang, H. (2022, May 4–6). *Representation and extraction of physics knowledge based on knowledge graph and embedding-combined text classification for cooperative learning*. 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China.

Smith, D. H., IV, & Zilles, C. (2024, July 8–10). *Code generation based grading: Evaluating an auto-grading mechanism for "explain-in-plain-english" questions*. 2024 on Innovation and Technology in Computer Science Education V.1, Milan, Italy.

Sobo, A., Mubarak, A., Baimagambetov, A., & Polatidis, N. (2024). Evaluating LLMs for code generation in HRI: A comparative study of ChatGPT, gemini, and claude. *Applied Artificial Intelligence*, *39*(1), 2439610. [CrossRef]

Song, T. (2023). Automatic generation of multiple-choice questions for CS0 and CS1 curricula using large language models. In *International conference on computer science and education*. Springer.

Sonkar, S., Ni, K., Tran Lu, L., Kincaid, K., Hutchinson, J. S., & Baraniuk, R. G. (2024). Automated Long Answer Grading with RiceChem Dataset. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (Vol. 14829, pp. 163–176). Springer Nature. [CrossRef]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. *arXiv*, arXiv:1706.03762. [CrossRef]

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*(1), 010152. [CrossRef]

Wang, K. D., 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*, 1330486. [CrossRef]

Wang, L. (2024). Suitability of GPT-4o as an evaluator of cardiopulmonary resuscitation skills examinations. *Resuscitation*, *204*, 110404. [CrossRef] [PubMed]

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *arXiv*, arXiv:2203.11171. [CrossRef]

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, arXiv:2201.11903. [CrossRef]

West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. *arXiv*, arXiv:2303.17012. [CrossRef]

Yan, Z., Zhang, R., & Jia, F. (2024, June 14–17). *Exploring the potential of large language models as a grading tool for conceptual short-answer questions in introductory physics*. 2024 9th International Conference on Distance Education and Learning (pp. 308–314), Guangzhou, China. [CrossRef]

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv*, arXiv:2305.10601. [CrossRef]

Yao, Y., Li, Z., & Zhao, H. (2024). Beyond chain-of-thought, effective graph-of-thought reasoning in language models. *arXiv*, arXiv:2305.16582. [CrossRef]

Yeadon, W., & Halliday, D. P. (2023). Exploring durham university physics exams with large language models. *arXiv*, arXiv:2306.15609. [CrossRef]

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic chain of thought prompting in large language models. *arXiv*, arXiv:2210.03493. [CrossRef]

Zhou, Y., Geng, X., Shen, T., Tao, C., Long, G., Lou, J.-G., & Shen, J. (2023). Thread of thought unraveling chaotic contexts. *arXiv*, arXiv:2311.08734. [CrossRef]

Zong, H., Wu, R., Cha, J., Wang, J., Wu, E., Li, J., Zhou, Y., Zhang, C., Feng, W., & Shen, B. (2024). Large language models in worldwide medical exams: Platform development and comprehensive analysis. *Journal of Medical Internet Research*, *26*, e66114. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.