JOURNAL OF BALTIC SCIENCE EDUCATION

ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

GRADERS OF THE FUTURE: COMPARING THE CONSISTENCY AND ACCURACY OF GPT4 AND PRE-SERVICE TEACHERS IN PHYSICS ESSAY QUESTION ASSESSMENTS

Yubin Xu,
Lin Liu,
Jianwen Xiong,
Guangtian Zhu

Abstract. *As the development and application of large language models (LLMs) in physics education progress, the well-known AI-based chatbot ChatGPT4 has presented numerous opportunities for educational assessment. Investigating the potential of AI tools in practical educational assessment carries profound significance. This study explored the comparative performance of ChatGPT4 and human graders in scoring upper-secondary physics essay questions. Eighty upper-secondary students' responses to two essay questions were evaluated by 30 pre-service teachers and ChatGPT4. The analysis highlighted their scoring consistency and accuracy, including intra-human comparisons, GPT grading at different times, human-GPT comparisons, and grading variations across cognitive categories. The intraclass correlation coefficient (ICC) was used to assess consistency, while accuracy was illustrated through Pearson correlation coefficient analysis with expert scores. The findings reveal that while ChatGPT4 demonstrated higher consistency in scoring, human scorers showed superior accuracy in most instances. These results underscore the strengths and limitations of using LLMs in educational assessments. The high consistency of LLMs can be valuable in standardizing assessments across diverse educational contexts, while the nuanced understanding and flexibility of human graders are irreplaceable in handling complex subjective evaluations.*

Keywords: *Physics essay question assessment, AI grader, Human graders.*

Yubin Xu
South China Normal University, P. R. China
Lin Liu
Guangzhou Tieyi High School, P. R. China
Jianwen Xiong
South China Normal University, P. R. China
Guangtian Zhu
Jimei University, P. R. China

Introduction

Assessment plays a central role in education, evolving to encompass a broader range of objectives over the last few years. The focus has shifted from merely evaluating conceptual understanding to assessing higher-order thinking and competencies, fostering a more holistic 'assessment of learning' (Dochy et al., 2006). This new assessment of culture prioritizes the evaluation of analytical processes and problem-solving skills over factual knowledge and basic cognitive tasks. Physics essay questions provide an effective way to assess students' understanding of knowledge and their ability to apply knowledge innovatively. The TIMSS achievement tests included essay questions with free-response, researchers demonstrated through examples that free-response items offer enriched insight into students' thinking, their conceptual understanding and the nature of their misconceptions (Angell et al., 2002). However, free-response questions pose significant challenges in scoring consistency and accuracy (Case & Swanson, 1993; Jonsson & Svingby, 2007; Martinez, 1999). Influenced by the subjective consciousness of different examiners, the accuracy of assessing essay questions is not as high as that for multiple choice questions. Finding ways to improve the consistency and accuracy of essay question evaluation holds significant importance for guiding teaching.

With the rapid development of big data, artificial intelligence (AI), and large language models (LLMs), the applications of LLMs in the field of smart education have broad prospects (Gan et al., 2023). Incorporating LLMs into the field of education provides potential support for personalized learning (Lin et al., 2022) and adaptive assessment (McNamara et al., 2023), thereby improving the quality of education and the learning experience. Utilizing the vast amount of learning data accumulated in the field of education (Piety et

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

al., 2014), LLMs conduct in-depth analysis and data mining that can reveal learner patterns (Vermunt & Donche, 2017), evaluate learning outcomes (Aziz et al., 2012; Jia et al., 2021; Rudolph et al., 2023), and provide personalized recommendations (Bhargava & Ng, 2022). LLMs possess advantages in processing and analyzing large-scale data, suggesting the potential to improving the consistency and accuracy of essay question evaluation.

Essay questions are essential for evaluating deeper understanding and analytical skills in educational assessments but pose significant challenges in scoring consistency and accuracy. The rapid advancement of AI technologies, particularly LLMs, offers potential solutions to these challenges. Exploring the application of LLMs (e.g., ChatGPT4) in evaluating upper-secondary physics essay questions holds significant value, as a comparative study of LLM and human grading performance in terms of accuracy and consistency can provide novel insights for practical assessment practices.

Literature Review

Physics Essay Questions

The measurement of thinking skills is mostly performed using multiple-choice tests, but this method has the shortcoming of allowing students to anticipate the answers (Kubiszyn & Borich, 2024). Many students are able to give a correct answer to a multiple-choice question, but they do not understand the physics concepts and principles related to the questions (Henderson et al., 2001). Besides, multiple-choice questions do not provide information about the thinking process, only elicit the final answer from students (Kastner & Stangla, 2011). Questions related to thinking ability require the rater to identify the student's potential thinking pathways when scoring (Khan & Aljarallah, 2011). In this case, essay questions with free-response are suitable to represent students' high level cognitive skills (Birenbaum & Tatsuoka, 1987; Haladyna & Rodriguez, 2013; Haudek et al., 2012).

Physics education goals for 21st-century learning have evolved to emphasize higher end skills including reasoning, creativity, and open problem solving (Bao & Koenig, 2019). Essay questions in physics education are noted for their ability to encourage innovative thinking and a deeper conceptual understanding (Dudung & Oktaviani, 2020; Rusilowati et al., 2023). Using these questions helps teachers gauge the depth of students' conceptual understanding, which is valuable for adjusting teaching strategies and improving instruction (Risnita & Bashori, 2020). In fact, in the A-level physics examination in the UK, essay questions constitute the largest proportion of the exam. Essay questions account for more than 50 percent of the International General Certificate of Secondary Education (IGCSE) exams. Essay questions with free-response can test students' deep understanding of physical concepts and principles, their ability to solve practical problems, and explain physical phenomena (Quitadamo & Kurtz, 2007).

Designing objective and accurate scoring criteria is an important prerequisite for the accurate evaluation of essay questions. A widely used taxonomy of cognitive processes, Bloom's taxonomy (Anderson & Krathwohl, 2001), serves as a framework for instruction and assessment. Our research is based on Bloom's taxonomy of educational objectives and combined with expert analysis. We have designed the scoring criteria of essay questions adopted in this study. The criteria cover numerous cognitive domain categories, including knowledge and understanding, application and analysis, synthesis, evaluation and creation.

Large Language Models in Physics Education

Over the past decades, AI for education has received a great deal of interest and has been applied in various educational scenarios (Wang et al., 2024). As a scoring tool, LLMs are capable of generating logically consistent answers across disciplines, balancing both depth and breadth (Susnjak & McIntosh, 2024). Students using ChatGPT by keeping or refining the results from LLMs as their own answers tend to perform better than average (Malinka et al., 2023). Using fine-tuned LLMs to generate human-like responses, these results provide real-time assistance to students by helping them solve challenging questions, correcting errors, and offering explanations or hints for areas of confusion (Ouyang et al., 2022). Researchers have also found the great potential of LLMs to help teachers create high-quality educational materials (Leiker et al., 2023). These studies show that ChatGPT is a useful tool that can significantly aid teachers in question generation and evaluation.

The powerful text processing capabilities of LLMs offer significant potential for automating assessments. LLMs can play a crucial role as tools in the processes of question generation, solving, and evaluation. In the question generation phase, LLMs can assist teachers and professional question designers by using algorithms and big data analysis to extract useful information from a vast array of academic papers, textbooks, and online resources

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

(BaiDoo-Anu & Owusu Ansah, 2023; Elkins et al., 2023; Terwiesch, 2023). This aids in creating comprehensive, moderately difficult, and discriminative questions more efficiently. During the question solving phase, LLMs can provide solutions from multiple perspectives, thereby helping question designers refine their questions and the framing of problems (Lehmann et al., 2024). In the evaluation phase, LLMs can automatically assess whether student responses meet the requirements based on predefined scoring criteria, enabling objective and accurate evaluations (Pinto et al., 2023; Xiao et al., 2024; Yancey et al., 2023).

However, there are still many limitations as AI continues to develop in the field of education. Existing work has found that when used as a marking tool, LLMs' concordance with human markers averages at 50.8%, with notable inaccuracies in marking straightforward questions (Yeadon & Hardy, 2024). Another study reveals that GPT has shown considerable potential for grading freeform student work in physics; however, while AI-assigned grades have a strong correlation with manually assigned grades, they are currently not reliable enough for summative assessments (Kortemeyer, 2023b). A case study explores how humanlike GPT's responses are in an introductory physics course; the result is that GPT would narrowly pass the course while exhibiting many of the preconceptions and errors typical of a beginning learner (Kortemeyer, 2023a). Therefore, when using LLMs as a scoring tool for physics essay questions, we should avoid over-relying on these tools at the expense of human judgment. Exploring the integration of both is a crucial area of research.

Consistency and Accuracy of Assessment

The aim of this study was to compare the consistency and accuracy of scoring physics essay questions between ChatGPT4 and human graders to explore potentially more efficient grading methods. The intraclass correlation coefficient (ICC) measures the extent of agreement and consistency among raters for two or more numerical or quantitative variables (Bujang & Baharum, 2017). The accuracy of the physics essay questions scored by ChatGPT4 and human graders is illustrated using Pearson correlation coefficient analysis with expert scores.

The ICC is also designed to measure the degree of reliability, consistency, and stability (Bartko, 1976; Shrout & Fleiss, 1979). The widely accepted evaluation criteria for the ICC are as follows: ICC < 0.4 indicates poor consistency; $0.4 <$ ICC < 0.59 indicates fair consistency; $0.6 <$ ICC < 0.74 indicates good consistency; and ICC > 0.75 indicates excellent consistency (Cicchetti, 1994).

In addition, this study employed the Bland-Altman method for consistency testing. The Bland-Altman method evaluates the agreement between two measurement methods by creating a Bland-Altman plot (Watson & Petrie, 2010). The basic concept involves calculating the mean and difference of two measurements, then computing the mean difference and its 95% confidence interval (mean $\pm$ 1.96 $\times$ standard deviation). A scatter plot is then created with the mean on the x-axis and the difference on the y-axis, including the 95% confidence interval. If most points lie within the 95% confidence interval, it indicates good agreement between the two methods. This study used the Bland-Altman method to examine the consistency between grading using LLMs and human grading, providing a visual representation of their consistency. By analyzing specific cases outside the 95% confidence interval, the study further investigates the underlying reasons for discrepancies.

When the sample size is too small, using ICC and Pearson Coefficient Correlation analysis to describe the consistency and accuracy between groups can lead to significant biases. For consistency testing with small samples, non-parametric tests are generally recommended (Egbert & LaFlair, 2018). This study aimed to explore the intraclass correlation of three sets of data, therefore, the Friedman test was selected. For small samples that do not meet the normality assumption, the root mean square error (RMSE) is used to describe the consistency between two sets of data, thereby explaining the accuracy of the study data (mean scores from teachers and GPT) relative to the standard data (expert scores).

Research Questions

This research focused on the application of LLM (ChatGPT4) in scoring upper-secondary school physics essay questions, assessing its viability as a tool to enhance educational assessments. The grading performance of ChatGPT4 against that of pre-service physics teachers was contrasted, focusing on scoring consistency and accuracy. The following research questions were formulated:

RQ1: Does ChatGPT4 demonstrate consistency in grading the same set of physics essay questions multiple times, and how does this consistency compare to that of human graders assessing the same responses?

RQ2: How does the accuracy of ChatGPT4's grading of physics essay questions compare to human grading?

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

RQ3: How does ChatGPT4's consistency and accuracy performance in grading physics essay questions across different cognitive categories compare to human grading?

Research Methodology

General Description

This study employed a quantitative research design to investigate the grading consistency and accuracy of GPT and human graders from multiple perspectives. Specifically, in the spring semester of 2024, ChatGPT4 and 30 human graders evaluated the responses from 80 upper-secondary students to two essay questions of physics. The analysis encompassed intra-human grading variations, temporal consistency of GPT grading, human-GPT comparisons, and grading differences across cognitive categories. The research received ethical approval, and all participants provided informed consent before participating.

Materials

Two physics essay questions (Essay Question 1 and Essay Question 2) were developed as test materials, the three experts formulated scoring criteria for the questions and provided unified expert scores for the eighty student responses. The specific content of these questions is outlined in Question Cards 1 and Question Cards 2, along with brief analyses of their characteristics and the knowledge content assessed. According to Bloom's taxonomy of educational objectives, the scoring criteria encompass multiple cognitive categories, including knowledge and understanding, application and analysis, synthesis, evaluation and creation. The scoring criteria for both questions were formulated under expert guidance, based on the point-based scoring method.

Question Card 1: Escape Lanes

Escape lanes enable vehicles with brake failure to decelerate and stop safely, away from other traffic. As shown in the left figure, these escape lanes are inclined upward, covered with small gravel, and end in a pile of discarded tires. Another type of escape lane utilizes a vehicle arresting system, as shown in the right figure. This system consists of a series of steel nets along the escape lane to stop the vehicle. These nets are connected to barriers on both sides of the lane with long steel belts that exceed their elastic limits upon vehicle deceleration.

Please discuss the following questions:

1-a Explain the principle behind stopping a vehicle using the escape lane shown in the left figure.
1-b Describe the advantages of constructing a vehicle arresting system escape lane compared to the escape lane shown in the left figure.
1-c Explain why it is necessary for the steel belts to exceed their elastic limit.

The design of Essay Question 1 drew inspiration from the 2016 A-level exam in the UK and was tailored to suit Chinese high school students. This question integrates multiple physics concepts, including kinetic energy, potential energy, friction, and elastic and plastic deformation (1-a, knowledge and understanding). It requires students to apply these concepts to real-world situations, analyzing the characteristics and advantages of both escape lanes (1-b, application and analysis). The question assesses students' ability to synthesize and use relevant knowledge to explain physical problems in complex contexts (1-c, synthesis).

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Question Card 2: Automated Logistics

In an automated logistics warehouse, workers frequently perform loading operations next to a high-speed conveyor belt. Recently, it has been observed that workers near the conveyor belt sometimes experience mild electric shocks under certain weather conditions.

Please discuss the following issues:

2-a Describe why workers feel electric shocks when close to the conveyor belt during loading operations and explain how weather conditions can affect this phenomenon.

2-b Assuming the conveyor belt is made of non-conductive material and prone to becoming electrically charged during high-speed motion, analyze the physical principles behind the charging of the conveyor belt and discuss the impact of this charge accumulation on the surrounding environment or equipment.

2-c Discuss factors that may intensify or mitigate the electric shocks experienced by workers and propose several potential methods to reduce or eliminate static accumulation on the conveyor belt.

Essay Question 2 is open-ended, with no single correct answer. Students are required to describe the relationship between static electricity and weather conditions on charge accumulation (2-a, knowledge and understanding), combining it with the actual production scenario, examine students' ability to apply and analyze this knowledge (2-b, application and analysis). Finally, students are encouraged to propose solutions from multiple perspectives, stimulating innovative thinking and providing a reasonable evaluation of these solutions (2-c, evaluation and creation).

Based on Bloom's taxonomy of educational objectives and expert guidance, experts have classified the six sub-questions in Essay Questions 1 and 2 into cognitive categories, as detailed in Table 1. Categorizing the types of questions facilitates subsequent discussions on the comparative performance of GPT and human grading across various cognitive categories.

Table 1
Questions' Cognitive Categories

Cognitive Categories	Item Number
Knowledge and Understanding	Q1a , Q2a
Application and Analysis	Q1b , Q2b
Synthesis	Q1c
Evaluation and Creation	Q2c

The design of the scoring criteria for the essay questions was grounded in expert opinions and relevant theories of educational assessment. Three experts—a university professor, a physics instructor from the tested class, and a graduate student specializing in physics education research—collaborated to analyze and discuss the responses of 80 students, and formulated scoring criteria for two essay questions. Together, they provided a unified expert scoring for these student responses, which served as the benchmark for evaluating the accuracy of subsequent assessments by GPT and pre-service teachers. A point-based scoring method was employed, in which each answer was compared to a standard, with points awarded based on the appropriateness of the response. Both Essay Ques-

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

tion 1 and Essay Question 2 adhered to point-based scoring principles. The scoring criteria for both questions were detailed in Tables 2 and 3 respectively.

Table 2
Scoring Criteria of Escape Lanes

Scoring Principles	1. Point-by-point scoring (bolded content in the scoring criteria, **1 point per item**): Full points are awarded if all mentioned points are included within the correct answers (similar suffices) and the terminology used is accurate; no points are awarded for missing or incorrect points. 2. Responses must be in **complete sentences**: If only correct keywords are listed without coherent logical expression, one point is deducted for each sub-question. 3. For open-ended questions: Points are awarded if the **direction of the answer is correct (conforms to scientific accuracy)** and the content is rational. Full points are given if at least two points are addressed, and all mentioned points fall within the correct answer range.	
Item	Scoring Criteria	Score
1-a	1. The escape lane in Figure 1 utilizes the principle of **converting kinetic energy into gravitational potential energy and internal energy**, effectively stopping the vehicle. As the vehicle ascends, **it works against gravity, thereby gradually transforming kinetic energy into gravitational potential energy**. (2 points awarded. Points are also awarded if the student answers from the perspective of force and acceleration: the component of gravity acts downwards along the slope, and the friction from the gravel is significant. These two forces produce an acceleration opposite to the direction of velocity, causing the vehicle to decelerate.) 2. Friction between the tires and the gravel not only transforms the vehicle's kinetic energy into internal energy but also **increases the magnitude of the frictional force**, allowing the vehicle to stop more quickly. (1 point) 3. The discarded tires at the end of the lane **serve as a buffer**, extending the collision time and reducing the interaction force. (1 point)	4
1-b	Reference Answer: 1. The steel net and barriers are designed to withstand impact forces and can be adjusted to accommodate different vehicle weights and speeds. 2. This system more effectively converts the vehicle's kinetic energy into internal energy. 3. Vehicle arresting systems generally occupy less space and are relatively simple to maintain. 4. Vehicle arresting systems have low terrain requirements, making them suitable for areas with significant geographical constraints. (3 points required for a full answer, 1 point each. Points awarded for correct direction and scientific validity of the answer.)	3
1-c	1. When an object undergoes deformation due to an external force and returns to its original state after the deformation ceases, it is called elastic deformation; if the object does not fully return to its original state, it is called plastic deformation. The steel belts exceed their elastic limit and undergo plastic deformation. (1 point) 2. By converting kinetic energy into internal energy, the steel belts absorb the vehicle's kinetic energy without recoiling, thereby decelerating the vehicle and ensuring it can come to a stop. (2 points) (Points are awarded for correct direction and scientific validity of the answer.)	3

Table 3
Scoring Criteria of Automated Logistics

Scoring Principles	1. Point-by-point scoring (bolded content in the scoring criteria, **1 point per item**): Full points are awarded if all mentioned points are included within the correct answers (similar suffices) and the terminology used is accurate; no points are awarded for missing or incorrect points. 2. Responses must be in **complete sentences**: If only correct keywords are listed without coherent logical expression, one point is deducted for each sub-question. 3. For open-ended questions: Points are awarded if the **direction of the answer is correct (conforms to scientific accuracy)** and the content is rational. Full points are given if at least two points are addressed, and all mentioned points fall within the correct answer range.	
Item	Scoring Criteria	Score
2-a	1. Cause of electric shocks: The surface of the conveyor belt **accumulates static electricity**. When workers approach or touch the conveyor belt, the accumulated charge **discharges** (points are awarded if students describe the discharge phenomenon) onto the workers, causing a sensation of electric shock. (2 points) 2. Impact of weather conditions: **Dry weather exacerbates** the accumulation of static electricity, while **humid weather reduces** it (mentioning either dry or humid suffices) because moisture **enhances the conductivity of the air**, aiding in the neutralization of charges. (2 points)	4

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

2-b	3. Physical Principle: When the conveyor belt moves at high speeds, **triboelectric charging occurs due to friction**, leading to the **transfer of electrons** from one object to another (points are awarded for identifying frictional electrification and explaining electron transfer, 2 points). 4. Impact: Interference with electronic devices, sparking or even fire hazards, and causing mild electric shocks to humans. (Answering any one of these correctly and scientifically awards 1 point)	3
2-c	1. **Exacerbating** Factors: Dry weather, humidity levels, types of materials (some materials are more prone to generate static electricity), speed and material of the conveyor belt. 2. **Mitigating** Factors: Increasing air humidity, using conveyor belts made from conductive materials, using anti-static materials for conveyor belts, regular cleaning and maintenance of the conveyor belts to reduce dust accumulation. 3. **Solutions**: Maintaining appropriate humidity levels within the warehouse, providing workers with anti-static clothing, using anti-static sprays or coatings, regular inspections and maintenance of the conveyor belts to reduce friction-generated static electricity, and installing static eliminators near the conveyor belts. (3 points required for a full answer, 1 point each. Points awarded for correct direction and scientific validity of the answer.)	3

The scoring criteria for the two essay questions initially clarify the general scoring principles by specifying point-by-point marking to ensure fairness in grading. Each scoring point corresponds to the knowledge or skill being assessed, emphasizing that responses must be logically coherent and composed in complete sentences. For open-ended questions, students can earn points if their responses are correct in direction and scientifically reasonable. Through the use of bolding, clear and specific scoring points are provided, serving as a reference for pre-service teachers and GPT reviewers.

Participants

The participants in this research included eighty upper-secondary school students, thirty pre-service teachers, and three experts—a university professor, a physics instructor from the tested class, and a graduate student specializing in physics education research. Students from two upper-secondary classes in China, all preparing for their college entrance examination with physics as a mandatory subject, were selected for this study. The test utilized a paper-based format, with students given 20 minutes to complete their answers, which were then collected immediately afterward. A total of 40 valid responses for each of the two essay questions were collected, resulting in 80 responses. The answers from the two classes were assigned random codes ranging from 1 to 40.

The human grading was conducted by 30 pre-service physics teachers, all specializing in physics education. Each pre-service teacher was responsible for grading the responses of eight different students: four from Class 1, who completed Essay Question 1, and four from Class 2, who completed Essay Question 2. For example, Teachers 1, 2, and 3 (T1, T2, T3) each graded the responses of eight students numbered from both classes (students S1 to S4 in Class 1 and s1 to s4 in Class 2). Subsections Q1a, Q1b, Q1c, and Q2a, Q2b, Q2c correspond to the three sub-questions of Essay Questions 1 and 2 respectively. Similarly, Teachers 4, 5, and 6 (T4, T5, T6) graded responses from students numbered 5-8 (S5 to S8 and s5 to s8); Teachers 7, 8, and 9 (T7, T8, T9) graded those from students numbered 9-12 (S9 to S12 and s9 to s12), and so forth, up to Teachers 28, 29, and 30 (T28, T29, T30), who graded the responses from students numbered 37-40 (S37 to S40 and s37 to s40). As a result, each of the 80 students' responses was graded by three different pre-service teachers. The grouping of teachers' grading is illustrated in Figure 1 below, and the scoring sheet (excerpt-Group 1) is detailed in Table 4.

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

Figure 1
The Grouping of Pre-service Teachers' Grading Sessions

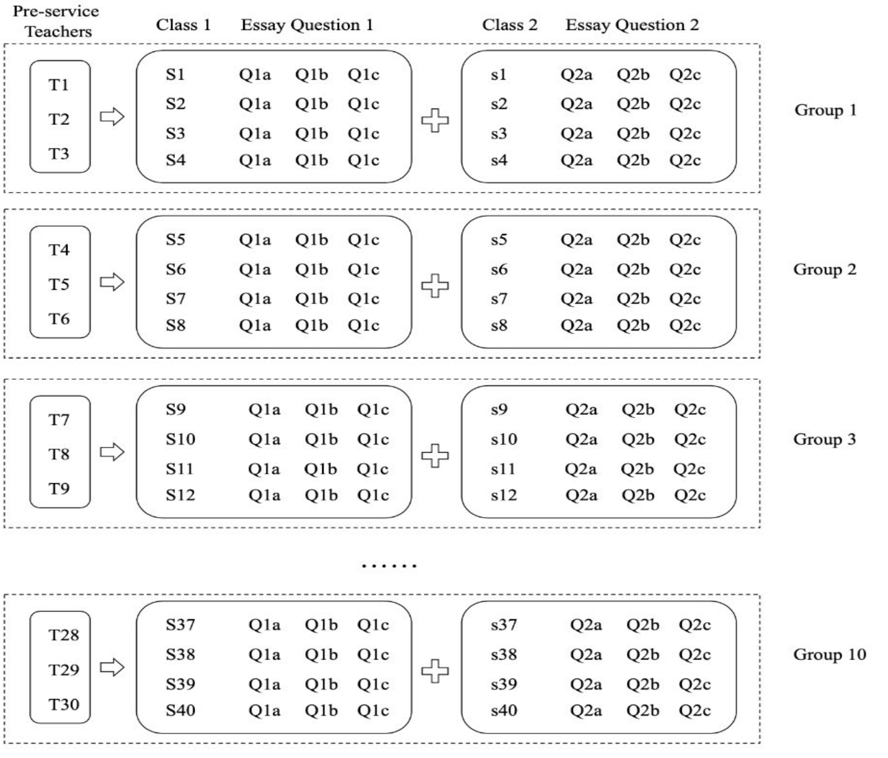


Table 4
Scoring Sheet (Excerpt-Group 1)

Teacher	Students Class 1	Q1a Score	Q1b Score	Q1c Score	Students Class 2	Q2a Score	Q2b Score	Q2c Score
T1	S1	4	1	1	s1	3	2	2
	S2	3	1	0	s2	2	3	2
	S3	3	1	1	s3	3	2	1
	S4	4	2	1	s4	0	1	0
T2	S1	3	1	2	s1	3	1	2
	S2	2	0	0	s2	1	2	3
	S3	3	1	3	s3	2	2	1
	S4	2	2	3	s4	1	1	0
T3	S1	3	0	1	s1	3	2	2
	S2	3	0	0	s2	0	2	1
	S3	3	1	2	s3	3	2	1
	S4	4	2	2	s4	1	1	0

Data Analysis

1. Stage I: Consistency and accuracy of pre-service teacher scoring

According to the grouping in Figure 1, the responses to Essay Question 1 from four students in Class 1 and Essay Question 2 from four students in Class 2 form one group. The answers of the 80 students were divided into ten groups (1-4, 5-8, 9-12, 13-16, 17-20, 21-24, 25-28, 29-32, 33-36, and 37-40). Each group's answers were inde-

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

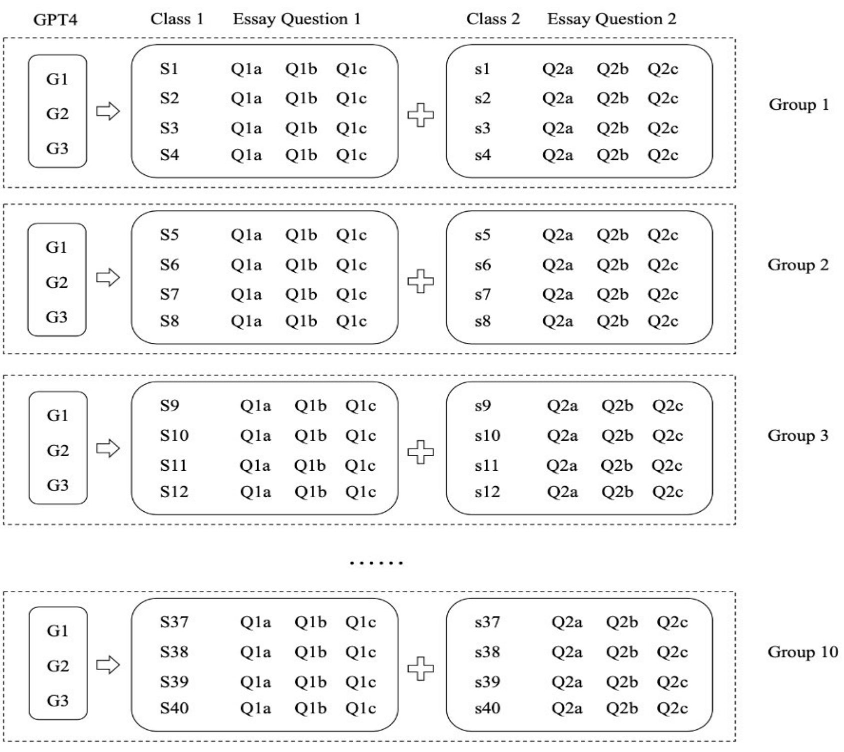
pendently scored by three pre-service teachers. The Intraclass Correlation Coefficient (ICC) was used to assess the consistency of scoring results among three teachers. The Pearson correlation coefficient was used to analyze the accuracy of the pre-service teachers' scoring results by comparing the average scores of the same group of pre-service teachers with the scores given by experts.

2. Stage II: Consistency and accuracy of ChatGPT4 scoring

ChatGPT4 was utilized as a grading tool, with the general scoring principles and detailed criteria explained to GPT4, and it was used to score all 80 student responses. The prompt for GPT comprised three parts: specifying requirements, clarifying the questions and scoring criteria, and detailing the scoring process with a brief explanation of the results. Initially, GPT was oriented with a defined role and task: "Assume the role of an experienced scorer of physics essay questions," and "You are now tasked with scoring the responses from a class of 40 students to essay questions." Subsequently, the essay questions were presented to GPT for comprehension, followed by a notification that the scoring criteria would be provided subsequently. Once GPT had comprehended the questions, the scoring criteria were delivered in two stages: first, the general scoring principles were elucidated, followed by the provision of detailed scoring guidelines. GPT was instructed to not only score the student responses but also to provide a brief explanation of the scores. Upon receiving these instructions, GPT requested the upload of student responses and commenced scoring. This procedure was repeated for two separate essay questions. A step-by-step guide has been attached in the appendix, enabling readers to easily follow the process of using GPT to evaluate students' responses.

GPT4 was assigned to score the same set of student responses at three separate times, generating three distinct sets of scoring data for all students. The results of GPT's three scoring sessions were labeled as G1, G2, and G3. Following the same grouping method as for the pre-service teachers, student responses were divided into ten groups, each receiving three evaluation results from GPT, as illustrated in Figure 2. The Intraclass Correlation Coefficient (ICC) was used to analyze the consistency of GPT's three scoring sessions within each group, to investigate whether GPT maintained consistency in scoring the same answers multiple times. This identical grouping method also facilitated a comparative analysis of the consistency of GPT's scores over different times compared to the consistency of scores from different teachers.

Figure 2
The Grouping of GPT4's Grading Sessions



https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

The mean of the three scores from GPT4 in each group was compared with expert scores using Pearson correlation coefficient analysis to assess accuracy. Given the identical grouping format, the accuracy of the scoring results between pre-service teachers and GPT4 could also be compared.

3. Stage III: Consistency of the mean scores between pre-service teachers and ChatGPT4

To conduct a more detailed analysis of the differences in scoring consistency between pre-service teachers and GPT4, Bland-Altman plots were applied to observe the distribution of the two sets of scoring results within the 95% confidence interval. This method allowed for targeted analysis of specific reasons for discrepancies outside the 95% confidence interval. According to the grouping in the Scoring Sheet, the average scores from three teachers were calculated in each group, as well as the average of the three scores from GPT4 in the same groups. These pairs of average data were compared to analyze their consistency. Each group included scores from eight students, four from Essay Question 1 and four from Essay Question 2, resulting in a total of 24 data points per group. The consistency of mean scores between pre-service teachers and GPT4 was analyzed using the Bland-Altman method.

4. Stage IV: Performance of consistency and accuracy between pre-service teachers and ChatGPT4 across different cognitive categories

Based on Bloom's taxonomy, the six sub-questions included in the two essay questions were classified into four cognitive categories: knowledge and understanding, application and analysis, synthesis, and evaluation and creation. Due to the reduced sample size within each category after categorizing the questions, the Friedman test was applied to evaluate the intragroup consistency of teachers' and GPT's scores for the categories with very small sample sizes, specifically synthesis, evaluation and creation. For the categories with relatively larger sample sizes—specifically, knowledge and understanding, and application and analysis—the Intraclass Correlation Coefficient (ICC) was continued to be used to analyze consistency. To evaluate the accuracy of scorings across different cognitive categories in a small sample, the root mean square error (RMSE) was calculated for both pre-service teachers and GPT against expert scores. By comparing the RMSE values, the accuracy of their scorings across different cognitive categories was assessed.

Research Results

The responses from 80 students were divided into 10 groups. For each group, three pre-service teachers and three GPT4 scoring sessions evaluated the responses. Across the 10 groups, results indicated that GPT4 exhibited higher consistency, whereas pre-service teachers demonstrated superior accuracy.

Consistency Comparison between Pre-service Teachers and ChatGPT4 Scoring

To assess the relative consistency within the group of pre-service teachers and compare it with the consistency across three GPT4 scoring sessions, the Intraclass Correlation Coefficient (ICC) was calculated for both datasets using the same set of student responses. Table 5 displays the ICC results for both datasets. According to Table 5, among the ten groups of student responses, five groups exhibited an ICC greater than 0.6 for pre-service teachers' scores, while seven groups demonstrated the same for GPT4's scores. Comparative analysis within the same groups revealed that in seven out of ten groups, GPT4's ICC surpassed that of the pre-service teachers. This suggests that GPT4 generally exhibits higher consistency, demonstrating better stability when given scoring criteria.

Table 5
Consistency of Scoring of Teachers and GPT4

Group	Students' Number	ICC-Human	ICC-GPT4
Group1	1-4	.721	.799
Group2	5-8	.336	.549
Group3	9-12	.618	.643
Group4	13-16	.593	.540
Group5	17-20	.670	.719
Group6	21-24	.634	.631
Group7	25-28	.787	.632

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Group	Students' Number	ICC-Human	ICC-GPT4
Group8	29-32	.471	.613
Group9	33-36	.593	.655
Group10	37-40	.419	.507

Note. In the table, entries highlighted in **bold** indicate higher levels of consistency.

Accuracy Comparison Between Pre-service Teachers and ChatGPT4 Scoring

In this study, three experts uniformly scored the answers of 80 students, and these scores served as the benchmark for assessing accuracy. To compare the accuracy of the scoring results between pre-service teachers and GPT4, the mean scores from three pre-service teachers and three GPT4 scoring sessions were calculated for each student group. Subsequently, the Pearson correlation coefficient (r) analysis was conducted to compare these mean scores with the expert scores. Table 6 presents the results of the Pearson correlation coefficient (r) between the mean scores of both groups and the expert scores. According to Table 6, the scoring from both the pre-service teachers and GPT4 showed a significant positive correlation with expert scores. Analyzing the scoring results for each group, it was observed that the Pearson correlation coefficients (r) between the mean scores of the pre-service teachers and the expert scores were consistently higher. This indicates that the accuracy of the pre-service teachers' scoring is superior to that of GPT4 when scoring criteria are provided.

Table 6
Accuracy of Scoring of Teachers and GPT4

Group	Students' Number	Pearson correlation coefficient (r)	
		Human and expert	GPT4 and expert
Group1	1-4	.950**	.798**
Group2	5-8	.799**	.678**
Group3	9-12	.886**	.709**
Group4	13-16	.747**	.503*
Group5	17-20	.869**	.833**
Group6	21-24	.927**	.865**
Group7	25-28	.943**	.913**
Group8	29-32	.843**	.795**
Group9	33-36	.948**	.854**
Group10	37-40	.881**	.664**

**Significant at the .01 level ($p < .01$)
*Significant at the .05 level ($p < .05$)

Based on the correlation between the mean scores of pre-service teachers and experts, two groups were selected for focused analysis: Group 1, which exhibited the highest consistency, and Group 4, which exhibited the lowest. Scatter plots illustrating the expert scores, pre-service teachers' mean scores, and GPT mean scores for these groups are shown in Figure 3 and Figure 4, respectively. Since eight students in each group answered three sub-questions, 24 scale marks were arranged on the horizontal axis of the figure. Each scale mark resents one student's score for one sub-question. For example, the first mark scale 1 on the horizontal axis in Figure 1 is the score of student S1 answering sub-question Q1a. In order to effectively observe the discrepancies between the mean scores graded by pre-service teachers, GPT4 and experts, the 24 scale marks were arranged in descending order of maximum expert score. Data points where discrepancies with the expert scores exceeded a difference of 1 were highlighted in larger dots.

The results showed that in Group 1, none of the 24 mean scores from pre-service teachers deviated by more than 1 from the expert scores. However, among GPT's 24 mean scores, three exceeded this deviation. In Group

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

4, two of the pre-service teachers' mean scores exceeded a difference of 1 from the expert scores, while eight of GPT's mean scores surpassed this threshold. Overall, GPT's mean scores were more likely to show "outlier data" with deviations exceeding 1 from the expert scores, suggesting a greater likelihood of significant discrepancies between GPT and experts when scoring certain student responses.

Figure 3
Scatter Plots of Group 1

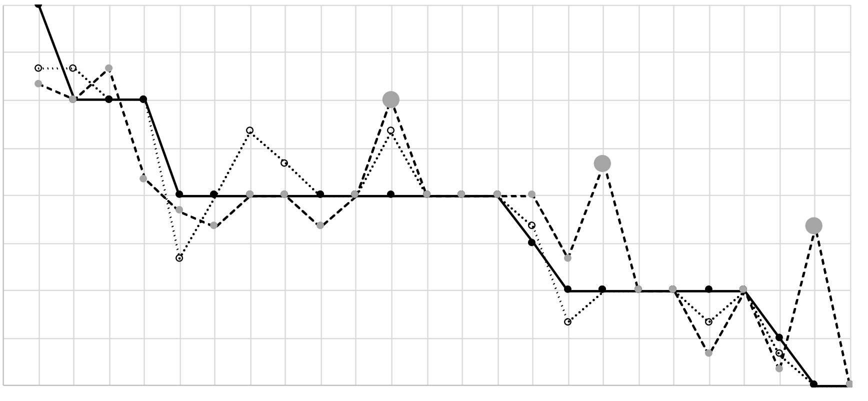
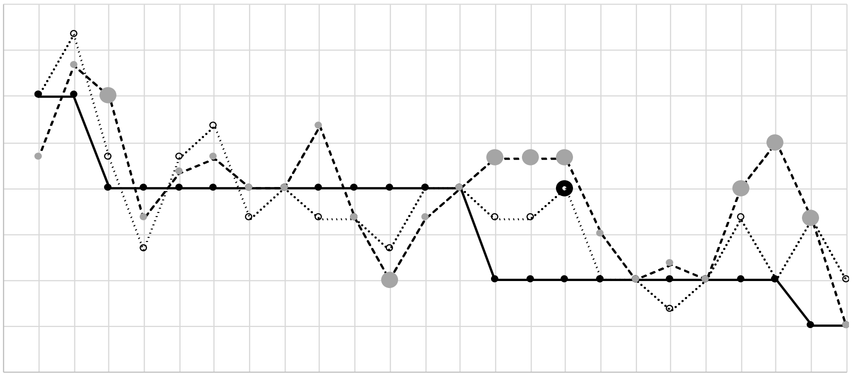


Figure 4
Scatter Plots of Group 4



Consistency Comparison Between Pre-Service Teachers' and ChatGPT4 Mean Scores

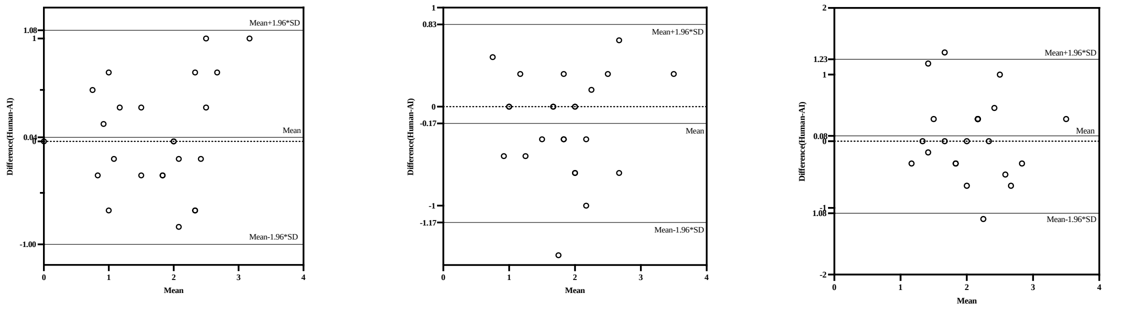
In Section B of the results, it was noted that the mean scores from teacher evaluations exhibited higher accuracy compared to those of GPT4, with GPT4 scores frequently displaying "outlier data" relative to expert scores. This section explores the consistency of mean scores from teacher evaluations and those from GPT4, identifies points of excessive inconsistency, and analyzes the reasons for these deviations by comparing them to expert scores. In each of the ten groups, both teachers and GPT4 provided 24 mean scores (8 students answering 3 sub-questions). The Bland-Altman method was utilized to analyze the consistency of their mean scores. Given the small amount of data (n = 24, less than 50) in the research, the Shapiro-Wilk test was opted to assess the normality of the difference between the scores from teachers and GPT4. Out of the ten groups analyzed, nine passed the normality test, indicating a normal distribution of difference between human scores and GPT scores (except for Group 1). Employing the Bland-Altman analysis method, Bland-Altman plots were generated for the mean scores of ten sets of pre-service teachers and GPT4. The plots for three groups—Group 3, with good consistency; Group 4, with moderate consistency; and Group 2, with poor consistency—are presented in Figure 5.

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Figure 5

The Bland-Altman Plots of Three Example Groups



Group3 Group4 Group2

Based on the Bland-Altman plots for the ten groups, it was observed that the scores for Groups 3, 7, and 10 all fell within the 95% confidence interval, indicating good consistency between the pre-service teachers and GPT grading. For Groups 4, 6, and 8, 4.2% (1 out of 24) of the scores lay outside the 95% confidence interval. The differences in mean scores between teachers and GPT for these groups fell within the expected 5% range, and thus the consistency was considered acceptable. However, the remaining four groups exhibited 8.3% (2 out of 24) of scores falling outside the 95% confidence interval, exceeding the expected 5% range. Considering the influence of the characteristics and quality of students' responses on consistency, a detailed analysis of data points falling outside the 95% confidence interval was conducted, yielding the results presented in Table 7.

Table 7

Inconsistent Data Points Between Pre-service Teachers and GPT4

Group	Students' Number	Question Item	Mean Scores		
			Human	GPT4	Expert
Group1	S2	Q1c	.00	1.67	.00
	s2	Q2a	1.00	2.33	1.00
Group2	s6	Q2a	2.33	1.00	1.00
	s8	Q2a	1.67	2.83	2.00
Group4	s16	Q2a	1.00	2.50	1.00
Group5	S17	Q1b	2.00	.83	2.00
	S19	Q1c	1.00	2.33	1.00
Group6	S24	Q1b	.67	1.50	1.00
Group8	S31	Q1b	1.00	2.00	1.00
Group9	S34	Q1c	2.33	1.00	2.00
	S35	Q1a	1.67	2.83	2.00

The data points that exceeded the 95% confidence interval were categorized to analyze specific cases. From a grading perspective, the mean scores from pre-service teachers were generally closer to the expert scores. The only exception was for Student 6(s6) in Group 2 answering sub-question Q2a, where the grading accuracy of the pre-service teachers was lower than that of GPT4. Further analysis of this particular case revealed that one of the three pre-service teachers made an error in judging the student's response. The student's statement that 'humid weather might exacerbate this phenomenon (electric shocks)' was mistakenly judged as 'correct,' whereas in reality, dry weather intensifies static electricity situations. This indicates that even human grading is susceptible to rare misjudgments, thus justifying the practice of averaging scores from multiple graders for subjective questions to enhance reliability.

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Aside from this isolated case, GPT4 exhibited lower grading accuracy in the remaining cases. This aligns with the findings derived from the scatter plots presented in Section B of the results. Further analysis of these cases revealed that some inaccuracies arose from misjudgments by GPT4, while others stemmed from the students' responses being relatively vague, leading to misunderstandings by GPT4 of the intended meaning. This suggests that pre-service teachers are generally better at understanding students' language and expressions, and that GPT4 is more prone to misinterpretations.

Consistency and Accuracy Performance Across Different Cognitive Categories Between Pre-Service Teachers and ChatGPT4

Based on Bloom's Taxonomy, the six sub-questions of Essay Question 1 and Essay Question 2 are categorized into four cognitive categories, as shown in Table 3. The data from pre-service teachers and GPT4 scores according to these cognitive categories were analyzed to assess the consistency and accuracy of scoring between pre-service teachers and LLMs across different categories.

In terms of scoring consistency, within the knowledge and understanding category, GPT4 demonstrated higher consistency than pre-service teachers, achieving higher ICC values in seven of ten groups. In the application and analysis category, both GPT4 and pre-service teachers showed strengths and limitations in their consistency, with GPT4 achieving higher ICC values in six out of ten groups compared to the teachers. The ICC values for both groups are presented in Table 8.

Table 8
Scoring Consistency of Different Cognitive Categories

Category	Knowledge and Understanding		Application and Analysis	
Group	ICC-Human	ICC-GPT	ICC-Human	ICC-GPT
Group1	.672	.785	.741	.699
Group2	.331	.529	.380	.481
Group3	.448	.481	.713	.601
Group4	.707	.604	.545	.572
Group5	.751	.772	.636	.718
Group6	.245	.618	.588	.670
Group7	.696	.258	.537	.447
Group8	.249	.641	.427	.599
Group9	.502	.522	.364	.752
Group10	.355	.144	.837	.295

Note. In the table, entries highlighted in **bold** indicate higher levels of consistency.

In the categories of synthesis, evaluation and creation, the Friedman test for small samples was utilized to examine the consistency among raters within groups for both teachers and GPT. The Chi-square values for teachers and GPT are displayed in Table 9. In the synthesis category, among the ten groups, only in Group 4 did pre-service teachers have a Chi-square value significantly lower than that of GPT. The mean Chi-square value for teachers across ten groups was higher than that of GPT. A higher Chi-square value indicates greater variability among scorings, suggesting better consistency in GPT's scorings compared to teachers. In the evaluation and creation category, of the ten groups, five had Chi-square values where GPT exceeded teachers, and in two groups, the values were equal. The mean Chi-square value for GPT across these groups was slightly higher than that for teachers. Hence, in the evaluation and creation categories, the consistency of GPT was slightly inferior to that of the teachers.

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Table 9

Chi-square Value of Human and GPT

Category	Synthesis		Evaluation and Creation	
Group	Human	GPT	Human	GPT
Group1	5.636	4.000	2.000	**4.000**
Group2	5.600	2.000	8.000	2.667
Group3	5.600	4.000	5.600	2.000
Group4	2.000	**6.615**	2.000	2.000
Group5	2.932	3.000	.667	**3.000**
Group6	4.000	3.000	-	1.500
Group7	4.667	2.000	1.000	**6.500**
Group8	4.667	.800	6.000	**6.615**
Group9	3.000	2.000	5.692	**7.600**
Group10	6.500	-	5.286	5.286
Mean:	4.460	3.046	4.027	4.117

Note. In the table, entries highlighted in **bold** indicate lower levels of consistency.

To assess scoring accuracy, the root mean square error (RMSE) between the mean scores of pre-service teachers and expert scores was calculated, and the RMSE between the mean scores of GPT4 and expert scores was compared, as shown in Table 10. Lower RMSE values indicate smaller differences between the two data sets. The results showed that teachers' scoring accuracy consistently surpassed that of GPT across all categories. The mean RMSE across all ten groups was lower for teachers than for GPT, indicating that the average discrepancy between teachers' scores and expert scores was smaller. Examining the data across the ten groups in the four categories, the number of groups where GPT's RMSE was lower than that of the teachers was 2, 3, 3, and 2, respectively. Therefore, overall, teachers' scoring accuracy was higher than GPT's in every dimension.

Table 10

RMSE Between Expert and Two Groups (Human/GPT)

Category	Knowledge and Understanding		Application and Analysis		Synthesis		Evaluation and Creation	
Group	Human	GPT	Human	GPT	Human	GPT	Human	GPT
group1	.441	.770	.186	.243	.333	.837	.000	.236
group2	.580	.717	.471	.659	.553	**.527**	.667	.707
group3	.471	.640	.264	.610	.553	.672	.373	.527
group4	.391	.894	.429	**.400**	1.002	1.217	.333	**.167**
group5	.624	**.589**	.471	**.453**	.289	.687	.408	.672
group6	.264	.612	.425	**.391**	.373	**.344**	.000	.527
group7	.333	.400	.333	.486	.289	**.167**	.373	.601
group8	.624	**.550**	.500	.530	.408	.812	.391	**.354**
group9	.592	.743	.289	.349	.373	.833	.250	.583
group10	.295	.507	.132	.471	.441	.866	.553	.743
Mean	.461	.642	.350	.459	.461	.696	.335	.512

Note. In the table, entries highlighted in **bold** indicate lower RMSE of GPT

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Discussion

AI-based assessment tools have demonstrated significant potential for providing timely feedback on students' understanding of physics concepts (Verawati & Nisrina, 2024). AI-driven automatic scoring systems for physics essay tests enable instructors to concentrate on personalized teaching strategies rather than repetitive grading tasks (Kurniawan et al., 2024). However, people sometimes question the reliability of these AI tools, as noted by Ding et al. (2023), and this skepticism can hinder their full adoption in educational contexts where consistency and accuracy are essential.

This study comparatively examined the consistency and accuracy of scoring physics essay questions, exploring the potential of LLMs in this domain. Overall, human grading displayed superior accuracy, with experts and non-experts providing similar evaluations based on standardized criteria. Human scoring failed to show high consistency in some groups, influenced by two factors. First, pre-service teachers' differing understandings and applications of scoring criteria, even when all raters followed the same standards. The inherently subjective nature of essay assessment further complicates the process, making it challenging to ensure both consistent and accurate grading (Harahap & Djamas, 2019; Gilovich et al., 2002). Second, the characteristics of student answers, as more open or ambiguous responses could lead to more subjective scoring by different teachers, reducing consistency. In some cases, teachers' limited expertise in essay assessment techniques or insufficient familiarity with instructional strategies for physics essay questions contributed to reduced grading accuracy (Maison et al., 2020).

LLMs scoring showed superior consistency but lower accuracy compared to human scoring. Although prior research suggests that AI enhances the diversity, scientific rigor, and accuracy of assessment methods (Hamid et al., 2022; Zhai et al., 2021), the findings in this research indicated that GPT4's accuracy still falls short of human raters. GPT4's superior consistency stems from its immunity to subjective emotions or cognitive biases, but its lack of extensive training with relevant datasets limits its accuracy. Certain AI scoring systems developed for specific educational contexts can automatically process assignments with a low error rate (Bewersdorff et al., 2023; Ouyang et al., 2022). However, GPT4, which has not undergone extensive training with relevant data on these two questions, exhibited slightly lower performance in terms of accuracy. GPT4 was more prone to misjudgments, whereas human raters excelled at interpreting and evaluating complex responses, effectively capturing nuances that GPT4 often missed.

For instance, item 1-c in question 1 requires students to "*Explain why it is necessary for the steel belts to exceed their elastic limit.*" One student's response was, "*After exceeding the elastic limit, the steel belts will no longer elongate. The rigidity force increases, which helps stop the car. If the elastic limit is not exceeded, the steel belts will continue to elongate, resulting in a weaker elastic force that might fail to stop the car.*" Based on the grading criteria, the expert awarded this response a score of 0. Similarly, the average score given by three teachers was also 0 (0/0/0). However, GPT4's average score across three evaluations was 1.67 (2/2/1). The student's response contained accurate phrasing but lacked correct logic. Both teachers and experts agreed that the student's answer was logically unclear and did not address the fundamental reason why the steel belts must exceed their elastic limit—specifically, the conversion of mechanical energy into internal energy. In contrast, GPT4 scored the response higher and justified its scoring as follows:

- *"After exceeding the elastic limit, the steel belts will no longer elongate"* — This statement captures the key idea of "no recoil," which could justify 1 point.
- *"The rigidity force increases, which helps stop the car"* — This statement suggests the effect of exceeding the elastic limit and can be loosely interpreted as kinetic energy converting to internal energy, which may warrant another point.

This example highlights that when dealing with ambiguous responses, GPT tends to focus on surface-level information in the text rather than deeper logical reasoning, leading to a higher probability of misjudgment.

Bland-Altman analysis comparing the consistency of scores between pre-service teachers and GPT4 revealed varied levels of consistency across groups. Case analyses, when compared to expert scores, found that both pre-service teachers and GPT4 were capable of misjudgments, though more frequent with GPT4. When student answers were vague, pre-service teachers were more adept at understanding and handling the nuances in the responses than GPT4.

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Pre-service teachers and GPT4 demonstrated different strengths and limitations across various cognitive categories. In categories such as 'knowledge and understanding,' 'application and analysis,' and 'synthesis,' GPT scored with greater consistency. However, in the 'evaluation and creation' category, GPT's internal consistency was slightly lower than that of teachers. Overall, GPT showed better consistency, likely due to the absence of subjective factors following the given criteria, making its scoring more stable. However, in the 'evaluation and creation' category, which involves higher complexity, human scoring exhibited less variability within groups. In terms of accuracy, pre-service teacher scores outperformed GPT across all four categories. This indicates that compared to GPT, pre-service teachers are better at capturing subtle nuances in answers and more adept at flexibly applying scoring criteria. Faced with answers that are more complex, diverse, and open-ended, which may exceed the processing capabilities of LLMs, the advantages of human scoring—both in consistency and accuracy—become more evident.

Conclusions and Implications

Through a series of analyses on consistency and accuracy, it is evident that GPT4 provides stable scoring in most cases once scoring criteria are established. This stability is a significant advantage of using GPT4 for grading essay questions. Human grading often falls prey to subjective influences, particularly with subjective question types like essays, where scoring criteria might lack strict delineations of right and wrong, leaving space for interpretation. Even when all graders follow the same criteria, their interpretations and applications can vary.

Pre-service teachers' scores closely align with expert scores, illustrating the limitations of LLMs scoring—specifically, its challenges with deeply understanding content, applying knowledge in new contexts, or performing complex analyses as accurately as humans. This underscores that in certain scenarios, the expertise and judgment of human scorers remain irreplaceable. Moreover, the quality and characteristics of student answers can influence the accuracy of scores in complex questions. Additionally, if the training data for LLMs does not sufficiently cover the diversity or complexity of answers, their scoring accuracy may not reach human levels.

Based on these discussions, significant challenges remain in assessing complex learning outcomes in educational assessment, necessitating ongoing exploration and innovation in assessment methods. The research findings reveal that for the assessment of higher-order thinking skills, such as evaluation and creation of cognitive categories, human grading plays an irreplaceable role in educational assessments. Therefore, a practical approach might involve combining AI and human grading. This strategy could leverage AI to improve grading efficiency while ensuring accuracy through human review, thus capitalizing on the strengths of both and enhancing the overall consistency and accuracy of scores.

Limitations

This study acknowledged that the limited representativeness of the data (only two essay questions) and the small sample size (only two classes of students from China) may restrict the generalizability of its conclusions. Moreover, the rapid evolution of LLMs and related technologies suggests that the results of this study may become outdated as advancements continue. While the current analysis has produced intriguing findings, future research could yield more comprehensive insights by expanding the sample size to different countries and leveraging more sophisticated LLM tools. Furthermore, prioritizing the collection of diverse and relevant data and optimizing LLM training processes will be essential to enhancing the practical efficiency of educational assessments.

Acknowledgements

This study is supported by 2024 Fujian Province Undergraduate Teaching and Education Research Program. The authors would also like to express their gratitude to all the pre-service teachers and students who participated in the physics essay question tests for their assistance in data collection.

https://doi.org/10.33225/jbse/25.24.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Declaration of Interest

The authors declare no competing interest.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman.

Angell, C., Kjaernsli, M., & Lie, S. (2002). Exploring students' responses on free-response science items in TIMSS. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from others* (Vol. 8, pp. 159–187). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47219-8_8

Aziz, A. A., Yusof, K. M., & Yatim, J. M. (2012). Evaluation on the effectiveness of learning outcomes from students' perspectives. *Procedia-Social and Behavioral Sciences*, *56*, 22–30. https://doi.org/10.1016/j.sbspro.2012.09.628

BaiDoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bao, L., & Koenig, K. (2019). Physics education research for 21st century learning. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1), Article 2. https://doi.org/10.1186/s43031-019-0007-8

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762–765. https://doi.org/10.1037/0033-2909.83.5.762

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*, Article 100177. https://doi.org/10.1016/j.caeai.2023.100177

Bhargava, P., & Ng, V. (2022). Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12317–12325. https://doi.org/10.1609/aaai.v36i11.21496

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*(4), 385–395. https://doi.org/10.1177/014662168701100404

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Science*, *12*(1), 1–11.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine*, *5*(2), 107–115. https://doi.org/10.1080/10401339309539601

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, *20*(1), Article 63. https://doi.org/10.1186/s41239-023-00434-1

Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present and future trends* (pp. 191–206). Elsevier.

Dudung, A., & Oktaviani, M. (2020). Mathematical connection ability: An analysis based on test forms. *International Journal of Advanced Science and Technology*, *29*(6), 4694–4701.

Egbert, J., & LaFlair, G. T. (2018). Statistics for categorical, nonparametric, and distribution-free data. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The palgrave handbook of applied linguistics research methodology* (pp. 523–539). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-59900-1_23

Elkins, S., Kochmar, E., Serban, I., & Cheung, J.C.K. (2023). How useful are educational questions generated by large language models?. In Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C. (Eds.), *International Conference on Artificial Intelligence in Education* (pp. 536–542). Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2304.06638

Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023). Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)* (pp. 4776-4785). IEEE. https://doi.org/10.1109/BigData59044.2023.10386291

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Hamid, T., Chhabra, M., Ravulakollu, K., Singh, P., Dalal, S., & Dewan, R. (2022). A review on artificial intelligence in orthopaedics. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 365-369). IEEE. https://doi.org/10.23919/indiacom54597.2022.9763178

Harahap, D. E., & Djamas, D. (2019). Preliminary study on development of assessment performance instruments on physics learning to improve students' critical thinking ability. In *Journal of Physics: Conference Series* (Vol. 1185, No. 1, p. 012128). IOP Publishing. https://doi.org/10.1088/1742-6596/1185/1/012128

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education*, *11*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084

Henderson, C., Heller, K., Heller, P., Kuo, V. H., & Yerushalmi, E. (2001). Instructors' Ideas about Problem Solving–Setting Goals. In Cummings, K., Franklin, S., & Marx, J. (Eds.), *Proceedings of AAPT Physics Education Research Conference* (pp. 79-82). AAPT. https://doi.org/10.1119/perc.2001.pr.006

https://doi.org/10.33225/jbse/24.23.187

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). All-in-one: Multi-task learning bert models for evaluating peer assessments. *ArXiv*. https://doi.org/10.48550/arXiv.2110.03895

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences*, *12*, 263–273. https://doi.org/10.1016/j.sbspro.2011.02.035

Khan, M. U., & Aljarallah, B. M. (2011). Evaluation of modified essay questions (MEQ) and multiple choice questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. *International Journal of Health Sciences*, *5*(1), 39–43.

Kortemeyer, G. (2023a). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, *19*(1), Article 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kortemeyer, G. (2023b). Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research*, *19*(2), Article 020163. https://doi.org/10.1103/PhysRevPhysEducRes.19.020163

Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.

Kurniawan, W., Riantoni, C., Lestari, N., & Ropawandi, D. (2024). A hybrid automatic scoring system: Artificial intelligence-based evaluation of physics concept comprehension essay test. *International Journal of Information and Education Technology*, *14*(6), 876–882. https://doi.org/10.18178/ijiet.2024.14.6.2113

Lehmann, J., Meloni, A., Motta, E., Osborne, F., Recupero, D. R., Salatino, A. A., & Vahdati, S. (2024). Large language models for scientific question answering: An extensive analysis of the SciQA benchmark. In A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, & P. Lisena (Eds.), *The semantic web* (Vol. 14664, pp. 199–217). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60626-7_11

Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023). Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. *ArXiv*. https://doi.org/10.48550/arXiv.2306.01815

Lin, H., Wan, S., Gan, W., Chen, J., & Chao, H. C. (2022). Metaverse in education: Vision, opportunities, and challenges. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 2857-2866). IEEE. https://doi.org/10.1109/BigData55660.2022.10021004

Maison, D., Aatalini, K., DA, H., Kurniawan, W., Suryani, A., Lumbantoruan, A., & Dewi, U. P. (2020). Science process skill in science program higher education. *Universal Journal of Educational Research*, *8*(2), 652–661. https://doi.org/10.13189/ujer.2020.080238

Malinka, K., Peresíni, M., Firc, A., Hujnák, O., & Janus, F. (2023). On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education* (Vol. 1, pp. 47-53). Association for Computing Machinery. https://doi.org/10.1145/3587102.3588827

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

McNamara, D. S., Arner, T., Butterfuss, R., Fang, Y., Watanabe, M., Newton, N., McCarthy, K. S., Allen, L. K., & Roscoe, R. D. (2023). iSTART: Adaptive comprehension strategy training and stealth literacy assessment. *International Journal of Human–Computer Interaction*, *39*(11), 2239–2252. https://doi.org/10.1080/10447318.2022.2114143

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. *ArXiv*. https://doi.org/10.48550/arXiv.2203.02155

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 193-202). Association for Computing Machinery. https://doi.org/10.1145/2567574.2567582

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering* (pp. 293-302). Association for Computing Machinery. https://doi.org/10.1145/3613372.3614197

Quitadamo, I. J., & Kurtz, M. J. (2007). Learning to improve: Using writing to increase critical thinking performance in general education biology. *CBE—Life Sciences Education*, *6*(2), 140–154. https://doi.org/10.1187/cbe.06-11-0203

Risnita, R., & Bashori, B. (2020). The effects of essay tests and learning methods on students' chemistry learning outcomes. *Journal of Turkish Science Education*, *17*(3), 332–341. https://doi.org/10.36681/tused.2020.30

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, *6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Rusilowati, A., Negoro, R. A., Aji, M. P., & Subali, B. (2023). Development of waves critical thinking test: Physics essay test for high school student. *European Journal of Educational Research*, *12*(4), 1781–1794. https://doi.org/10.12973/eujer.12.4.1781

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Susnjak, T., & McIntosh, T. R. (2024). ChatGPT: The end of online exam integrity?. *Education Sciences*, *14*(6), Article 656. https://doi.org/10.3390/educsci14060656

Terwiesch, C. (2023). *Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course*. University of Pennsylvania.

Verawati, N. N. S. P., & Nisrina, N. (2024). The role of artificial intelligence (AI) in transforming physics education: A narrative review. *Lensa: Jurnal Kependidikan Fisika*, *12*(2), 212-228. https://doi.org/10.33394/j-lkf.v12i2.13523

Vermunt, J. D., & Donche, V. (2017). A learning patterns perspective on student learning in higher education: State of the art and moving forward. *Educational Psychology Review*, *29*(2), 269–299. https://doi.org/10.1007/s10648-017-9414-6

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large language models for education: A survey and outlook. *ArXiv*. https://doi.org/10.48550/arXiv.2403.18105

Watson, P. F., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology*, *73*(9), 1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). Human-AI collaborative essay scoring: A dual-process framework with LLMs. *ArXiv*. https://doi.org/10.48550/arXiv.2401.06431

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short l2 essays on the cefr scale with gpt-4. In Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., Tack,A., Yaneva, V., Yuan, Z., & Zesch, T. (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576-584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Yeadon, W., & Hardy, T. (2024). The impact of AI in physics education: A comprehensive review from GCSE to university levels. *Physics Education*, *59*(2), Article 025010. https://doi.org/10.1088/1361-6552/ad1fa2

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*(1), Article 8812542. https://doi.org/10.1155/2021/8812542

Appendix

General Steps to make GPT evaluate students' responses

Step 1: [Define Role and Task]
Assume you are an experienced evaluator of physics essay questions. Your task is to evaluate the responses of 40 students to a specific physics essay question. I will provide you with the question content, the scoring rubric, and the answers from 40 students.
(GPT Response: Understood. Please provide the following details: Question content, including the topic, background information, and the specific questions students need to answer. Scoring rubric, including grading criteria, point distribution for each section, and key aspects to focus on. Student responses: Provide the answers of the 40 students one by one, and I will evaluate them based on the rubric, giving scores and explanations for each. You may provide the materials step by step, and I will complete the scoring efficiently and thoroughly!)

Step 2: [Provide Question Content]
Here is the question content: XXX. Please familiarize yourself with the question; I will provide the scoring rubric shortly.
(GPT Response: Understood. I have reviewed the question on XXX. Please proceed with the scoring rubric or indicate specific aspects you'd like me to focus on when evaluating the student responses.)

Step 3: [Provide General Guidelines for Scoring]
First, you need to understand the general scoring principles: XXXX.
(GPT Response: Understood. These are detailed scoring principles, focusing on comprehensive coverage of points, clarity of expression, and scientific accuracy. I am now familiar with the general requirements for scoring. If you are ready to provide the detailed scoring points, we can begin the evaluation process.)
[Provide Detailed Criteria]
Under the guidance of the general principles, here are the detailed scoring criteria: XXXX.
(GPT Response: Got it. I fully understand the detailed scoring criteria and the key points for each question. I am now ready to evaluate the students' answers based on these criteria. You may start providing the student responses.)

Step 4: [Specify Requirements for Scoring Results]
Excellent. Next, I will provide you with the answers from 40 students. Please score them and provide a brief explanation for your scoring decisions.
(GPT Response: Certainly. Please provide the text of the student responses. I will score them according to the detailed criteria you provided. You may start sending the responses.)

After the Previous Steps, Teachers Can Provide Text Versions of Students' Answers to GPT.

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Received: *October 27, 2024* Revised: *December 04, 2024* Accepted: *January 18, 2025*

Cite as: Xu, Y., Liu, L., Xiong, J., & Zhu, G. (2025). Graders of the future: Comparing the consistency and accuracy of GPT4 and pre-service teachers in physics essay question assessments. *Journal of Baltic Science Education*, *24*(1), 187–207. https://doi.org/10.33225/jbse/25.24.187

Yubin Xu PhD Candidate, National Demonstration Center for Experimental Physics Education, School of Physics, South China Normal University, Guangzhou, Guangdong, 510006, P. R. China.
E-mail: yubinxu.phy@foxmail.com
ORCID: https://orcid.org/0009-0009-5967-3270

Lin Liu MS, Guangzhou Tieyi High School, Guangzhou, Guangdong, 510006, P. R. China.
E-mail: A15119266827@163.com

Jianwen Xiong PhD, Professor, National Demonstration Center for Experimental Physics Education, School of Physics, South China Normal University, Guangzhou, Guangdong, 510006, P. R. China.
E-mail: jwxiong@scnu.edu.cn

Guangtian Zhu *(Corresponding author)* PhD, Associate Professor, Teachers College, Jimei University, Xiamen, Fujian, 361021, P. R. China.
E-mail: zhuguangtian@gmail.com
ORCID: https://orcid.org/0000-0002-3677-8143