Contemporary Educational Technology, 2024, **16**(4), ep538
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory

Salima Aldazharova [1]
0009-0004-6986-1885

Gulnara Issayeva [1]
0000-0002-4860-3797

Samat Maxutov [2]
0000-0002-4453-5512

Nuri Balta [2*]
0000-0002-6364-5346

[1] Abai Kazakh National Pedagogical University, Almaty, KAZAKHSTAN
[2] SDU University, Almaty, KAZAKHSTAN
[*] Corresponding author: baltanuri@gmail.com

Citation: Aldazharova, S., Issayeva, G., Maxutov, S., & Balta, N. (2024). Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory. *Contemporary Educational Technology, 16*(4), ep538. https://doi.org/10.30935/cedtech/15592

ARTICLE INFO

Received: 2 Aug 2024

Accepted: 23 Sep 2024

ABSTRACT

This study investigates the performance of GPT-4, an advanced AI model developed by OpenAI, on the force concept inventory (FCI) to evaluate its accuracy, reasoning patterns, and the occurrence of false positives and false negatives. GPT-4 was tasked with answering the FCI questions across multiple sessions. Key findings include GPT-4's proficiency in several FCI items, particularly those related to Newton's third law, achieving perfect scores on many items. However, it struggled significantly with questions involving the interpretation of figures and spatial reasoning, resulting in a higher occurrence of false negatives where the reasoning was correct, but the answers were incorrect. Additionally, GPT-4 displayed several conceptual errors, such as misunderstanding the effect of friction and retaining the outdated impetus theory of motion. The study's findings emphasize the importance of refining AI-driven tools to make them more effective in educational settings. Addressing both AI limitations and common misconceptions in physics can lead to improved educational outcomes.

Keywords: AI assisted learning, force concept inventory, GPT-4, physics education

INTRODUCTION

The integration of artificial intelligence (AI) into educational settings has gained significant attention in recent years (Yilmaz et al., 2023). Advanced AI models, such as GPT-4, have shown remarkable capabilities in generating human-like responses and reasoning. However, understanding how effectively these models can replicate human cognitive processes, particularly in complex domains like physics, is crucial for their application in enhancing educational outcomes. This study addresses the educational challenge of accurately assessing students' conceptual understanding in physics. By analyzing AI-generated responses, we aim to identify and investigate misconceptions, thereby informing targeted instructional strategies and improving the integration of AI-driven educational tools into curricula (Anderson et al., 2019). Furthermore, examining the reasoning patterns of AI can point out the potential of AI to promote higher-order learning and critical thinking skills.

The study focuses on analyzing the accuracy of responses, identifying false positives and false negatives, and understanding the reasoning patterns of AI. Similar previous studies (Kortemeyer, 2023; West, 2023) employed an earlier version of GPT, which could not read figures. Therefore, they generated text versions of the questions that included figures and then assessed the performance of GPT. However, this study used GPT-4, which can read and analyze figures directly.

The GPT-4's responses on the force concept inventory (FCI) is not only relevant to AI developers but also holds significant educational value for science educators. By examining responses, we can learn common misconceptions and reasoning patterns that both AI exhibit. This information is crucial for educators to develop targeted instructional strategies and improve the effectiveness of their teaching methods when using AI. For instance, if GPT-4 consistently identifies correct answers with flawed reasoning, educators can recognize similar patterns in student responses and address these misconceptions more effectively in their instruction. Furthermore, understanding the AI's performance on the FCI can help educators enhance the validity and reliability of this diagnostic tool.

AI in Education

AI technologies have been increasingly adopted in educational environments to enhance learning outcomes and personalize education. Studies have shown that AI can support various educational activities, including tutoring, grading, and providing feedback (Chen et al., 2020; Holmes et al., 2019). AI-driven educational tools can adapt to individual learning styles and paces, offering a unique educational experience that traditional methods often cannot achieve (Balta, 2024; Luckin et al., 2016). For instance, AI can analyze student performance data to identify learning gaps and suggest targeted interventions (Roll & Wylie, 2016). Furthermore, AI can facilitate collaborative learning by matching students with complementary strengths and weaknesses (Rosé et al., 2019).

Recent research suggests that AI can potentially enhance conceptual understanding and critical thinking in students by providing sophisticated reasoning and problem-solving support (Chen et al., 2020; de los Ángeles Domínguez-González, 2023; Holmes et al., 2019). These capabilities allow AI to go beyond mere information retrieval, fostering higher-order cognitive skills essential for mastering complex subjects like physics.

AI in Physics Education

AI's role in physics education is increasingly recognized as crucial for improving learning outcomes. AI technologies can provide personalized feedback, identify student misconceptions, and enhance the overall learning experience (Buabeng, 2018). Studies by Wink and Bonivento (2023) highlight the opportunities AI presents in physics education, such as adaptive learning environments and real-time feedback mechanisms. Tschisgale et al. (2023) explore integrating AI-based methods into qualitative research, suggesting that computational grounded theory can lead to new developments in student learning behaviors and misconceptions.

Mustofa et al. (2024) reviewed the literature on AI in physics problem-solving, emphasizing the potential of AI tools like ChatGPT to assist students in understanding complex concepts. Yeadon and Hardy (2024) provided a comprehensive review of AI's impact on physics education from GCSE to university levels, showing that AI can support various educational stages and needs. Lai and Cheong (2022) discussed the educational opportunities and challenges in augmented reality (AR) implementations in physics education. They highlighted how AR, combined with AI, can provide immersive learning experiences that enhance conceptual understanding and engagement. Dahlkemper et al. (2023) investigated how physics students evaluate AI responses on comprehension questions, focusing on the perceived scientific accuracy and linguistic quality of ChatGPT. Their findings suggest that while AI can be a valuable educational tool, it is crucial to consider student perceptions and trust in AI-generated content.

The integration of AI into specialized fields like medical physics and high energy physics also offers significant educational visions. Andersson et al. (2021) examined the impact of AI on the medical physics profession from a Swedish perspective, highlighting the need for ongoing education and training to keep pace with technological advancements. Zanca et al. (2021) discussed the focus on AI in medical physics, particularly its role in improving diagnostic accuracy and treatment planning. Kuzu (2021) presented a machine learning approach in high energy physics, demonstrating how AI can enhance data analysis and experimental accuracy. Boehnlein et al. (2021) provided an overview of AI and machine learning applications in nuclear physics, underlining the importance of interdisciplinary education to prepare students for careers in these cutting-edge fields. Wulff (2024) explored the role of language in physics education and how AI might enhance language-related research and instruction. This area of study is crucial as it addresses the communication barriers that can hinder effective teaching and learning in physics. Krupp et al. (2024) investigated the negative consequences of GPT-4 assisted problem-solving in physics education. Their findings point out the risks of unexamined acceptance of AI-generated solutions, which can lead to superficial understanding and dependence on AI for problem-solving.

The Force Concept Inventory

Physics education often relies on diagnostic tools to assess students' understanding of fundamental concepts. FCI is a widely used instrument designed to evaluate students' grasp of Newtonian mechanics (Hestenes et al., 1992). It consists of multiple-choice questions that probe common misconceptions and core principles in physics. The FCI's effectiveness lies in its ability to reveal students' reasoning processes and identify specific areas of misunderstanding (Hake, 1998). Research has shown that the FCI can effectively measure conceptual gains in students, particularly when used in conjunction with interactive engagement methods (Hake, 1998; Halloun & Hestenes, 1985; Hestenes & Halloun, 1995).

One of the key strengths of the FCI is its ability to identify specific misconceptions that students hold. Misconceptions in physics are deeply rooted, intuitive beliefs that often contradict scientific principles. For example, many students believe that a constant force is required to keep an object in motion, which contradicts Newton's first law of motion. The FCI helps educators determine these misconceptions and address them directly through targeted instruction and interactive activities (Hestenes et al., 1992; McDermott & Redish, 1999).

The FCI questions cover a range of topics in Newtonian mechanics, including kinematics, dynamics, forces and interactions, and Newton's third law. Each question on the FCI is designed to test specific aspects of these topics, often presenting scenarios that are counterintuitive to students' everyday experiences (Balta & Eryılmaz, 2017). For example, a question might ask about the forces acting on an empty office chair resting on a floor, which can reveal whether students correctly understand the concept of normal force and equilibrium.

Reasoning in Physics Education

Reasoning is a critical component of learning and understanding physics. Students develop reasoning skills through a combination of instruction, practice, and the integration of new information with existing knowledge (Chi et al., 1989). Effective reasoning in physics involves the ability to apply conceptual knowledge to solve problems, analyze scenarios, and evaluate outcomes. Research has shown that students often struggle with reasoning in physics due to persistent misconceptions and intuitive beliefs that conflict with scientific principles (Hammer, 1996).

Studies have demonstrated that targeted instructional interventions can improve students' reasoning abilities in physics. For instance, interactive engagement methods and formative assessments have been shown to enhance students' conceptual understanding and reasoning skills (Docktor & Mestre, 2014; Hake, 1998). By focusing on reasoning, educators can better address the underlying misconceptions that hinder students' learning and foster a deeper understanding of physics concepts.

Recent research also emphasizes the role of metacognitive strategies in improving physics reasoning. When students are taught to reflect on their thinking processes and evaluate their understanding, they are more likely to develop effective problem-solving skills and a deeper comprehension of physics principles (Yerushalmi et al., 2017; Zohar & Dori, 2012). These findings suggest that incorporating metacognitive training into physics education can be a powerful tool for enhancing reasoning and overall learning outcomes. Additionally, studies have shown that integrating computer-based simulations and interactive learning environments can significantly enhance students' reasoning skills in physics by providing immediate feedback and enabling exploration of complex concepts (Smith & Knight, 2021; Van der Veen & Van den Berg, 2021).

AI and Physics Conceptual Understanding

Research on AI's ability to understand and reason about physics concepts has shown hopeful results. AI models can be trained to solve physics problems and generate explanations that align with human reasoning patterns (Lample & Charton, 2019). However, the extent to which AI models can accurately imitate human cognitive processes, particularly in identifying and addressing misconceptions, remains an area of active investigation (Bengio et al., 2020). Studies have demonstrated that AI can perform well on standard physics problems, but its ability to handle conceptual questions like those on the FCI varies (Schoenfeld, 2018). AI models often surpass in computation but may struggle with the conceptual concepts that require deeper understanding (Bengio et al., 2020).

Recent studies have demonstrated significant improvements in GPT-4's reasoning capabilities in physics compared to its predecessors. According to West (2023), GPT-4 has shown near expert-level competence in responding to FCI questions, achieving a much higher accuracy rate than GPT-3.5. This improvement is attributed to GPT-4's enhanced ability to process and apply conceptual knowledge rather than relying solely on rote memorization or pattern recognition. The study also highlights that GPT-4's responses are more stable and less prone to variability when subjected to different prompts or perturbations, indicating a more robust understanding of the underlying physics concepts (West, 2023). A comprehensive literature review by Mahligawati et al. (2023) indicates the current state of AI in physics education, providing an in-depth analysis of the benefits and challenges associated with its implementation.

Despite the advancements, GPT-4 still exhibits notable limitations. GPT-4's explanations can sometimes contain minor errors or inconsistencies, suggesting that while it has a strong grasp of many concepts, it still lacks the depth of understanding found in human experts. While it performs well on most FCI questions, it struggles with certain types of conceptual reasoning, particularly those involving spatial relationships and directionality. For example, GPT-4 has difficulty consistently applying the correct principles in problems that require understanding the cardinal directions of motion (West, 2023).

The ability of AI to interpret and reason about graphs in kinematics is crucial for assessing its overall physics competence. Polverini and Gregorcic (2024) explored GPT-4's performance on the test of understanding graphs in kinematics. Their findings indicate that GPT-4 performs comparably to high school students in interpreting kinematics graphs but exhibits significant differences in the distribution of correctness and the reasoning processes displayed. While GPT-4 was successful in proposing productive strategies for solving the tasks, it had difficulties correctly interpreting the visual aspects of graphs, suggesting the need for critical approaches when using AI as a tutor or assistive tool in physics education (Polverini & Gregorcic, 2024).

Kortemeyer (2023) examined whether an AI agent could pass an introductory physics course by evaluating GPT's responses to typical course assessments. The study found that while GPT could pass the course, it exhibited many preconceptions and errors common among beginning learners. The AI's performance was comparable to that of a student with a rudimentary understanding of physics, capable of generating plausible answers but often stumbling on detailed calculations and deeper conceptual understanding (Kortemeyer, 2023).

Integrating AI into physics education offers both challenges and opportunities. Studies such as those by Lai and Cheong (2022) and Dahlkemper et al. (2023) demonstrate the potential for AI to enhance learning environments and provide valuable feedback to students. However, it is essential to consider the limitations and biases that may arise from relying heavily on AI, as noted by Wang (2020) and Ge and Hu (2020).

Comparing AI and Human Learners

One of the primary areas of comparison between AI and human learners is in their reasoning patterns. Human learners typically develop reasoning skills through a combination of instruction, practice, and the integration of new information with existing knowledge (Chi et al., 1989). In contrast, AI models like GPT-4 rely on vast amounts of data and sophisticated algorithms to generate responses. This fundamental difference often results in distinct reasoning pathways.

Studies have shown that while AI can correctly answer a significant portion of FCI questions, it often does so using different reasoning pathways compared to students. For example, AI models may use statistical patterns and correlations found in the training data, which can sometimes lead to correct answers without a deep understanding of the underlying principles (Geiger et al., 2021). In contrast, human learners typically build their understanding through conceptual frameworks and experiential learning.

Another important aspect of comparing AI and human learners is the nature and frequency of misconceptions and errors. Human learners often hold persistent misconceptions about fundamental physics concepts, such as the nature of force and motion (Hestenes et al., 1992). These misconceptions can be difficult to overcome and require targeted instructional interventions. AI models, on the other hand, can also exhibit errors, but these are usually different from those seen in human learners. For instance, AI models might generate responses that are statistically likely but conceptually incorrect. Analyzing these errors can reveal the limitations of AI reasoning and suggest areas for improvement.

A critical factor in comparing AI and human learners is the stability and consistency of their responses. Human learners can exhibit variability in their answers based on factors such as stress, fatigue, and confidence levels (Schunk & Pajares, 2002). In contrast, AI models like GPT-4 can provide consistent responses given the same prompts, which can be advantageous in educational settings. However, this consistency can also be a drawback if the AI's underlying misconceptions are not addressed, leading to repeated errors.

AI models have shown varying levels of performance on conceptual versus computational tasks. Research has demonstrated that AI often excels in computational tasks that require precise calculations and the application of formulas (Lample & Charton, 2019). However, conceptual tasks that require a deep understanding of physical principles and the ability to synthesize information across different contexts pose a greater challenge. For instance, while GPT-4 can accurately solve physics problems that involve straightforward calculations, it may struggle with FCI questions that probe deeper conceptual understanding. A study by Polverini and Gregorcic (2024) found that while GPT-4 performed comparably to high school students, it exhibited significant differences in the distribution of correctness and reasoning processes.

The quality of AI responses can be significantly influenced by the design of prompts. Prompt engineering involves crafting specific queries that guide the AI towards more accurate and relevant answers. Polverini and Gregorcic (2024) demonstrated that carefully crafted prompts could enhance GPT-4's performance on physics problems. By specifying the context and asking the AI to behave like a physics teacher, they were able to elicit more accurate and detailed responses.

False Positives and False Negatives in Physics Education

The accurate assessment of students' understanding is vital for diagnosing misconceptions and guiding instructional strategies. False positives and false negatives in educational assessments can significantly impact the effectiveness of teaching and learning. Hestenes et al. (1992), the creators of the FCI, initially highlighted the importance of accurately diagnosing students' misconceptions in physics. They emphasized that understanding the nature of these misconceptions is crucial for developing effective instructional interventions. The study demonstrated that while the FCI is effective in identifying common misconceptions, the tool's accuracy can be affected by false positives and false negatives.

McDermott and Redish (1999) reviewed the state of physics education research, emphasizing the persistent nature of misconceptions in students' understanding of physics concepts. They noted that misconceptions are often deeply rooted and resistant to traditional instructional methods, leading to frequent occurrences of false positives when students' incorrect beliefs align with superficially correct answers. Hammer (1996) explored students' beliefs about force and motion, identifying specific misconceptions that lead to false positives and false negatives in assessments. Hammer (1996) found that students often hold intuitive beliefs that conflict with Newtonian physics, such as the idea that a constant force is needed to maintain motion. These misconceptions can result in correct answers based on incorrect reasoning (false positives) or incorrect answers despite correct reasoning processes (false negatives).

Geiger et al. (2021) conducted a comparative analysis of misconceptions in physics held by both human students and AI models. The study found that AI models, including GPT-3 and GPT-4, often produced false positives by overgeneralizing patterns learned from training data. For example, an AI might correctly identify that an object in motion stays in motion but provide incorrect reasoning based on patterns rather than a deep understanding of Newtonian mechanics. Another study by Kortemeyer (2023) evaluated whether an AI agent could pass an introductory physics course. The study highlighted the prevalence of false negatives in AI-generated responses, where the AI provided correct answers with incorrect reasoning. These false negatives were attributed to the AI's lack of well understanding of physics concepts, leading to correct answers being marked as incorrect due to flawed reasoning. The AI's tendency to produce false positives and false negatives was particularly obvious in questions involving the interpretation of graphs, where it failed to correctly analyze visual information despite providing seemingly logical reasoning (Polverini & Gregorcic, 2024).

While previous studies have demonstrated AI's proficiency in solving computational problems and offering instructional support in physics, a significant gap exists in its ability to handle conceptual assessments such as those found in the FCI. These assessments require a deep understanding of fundamental principles and often involve complex visual data. Current AI models, including GPT-4, have demonstrated difficulties in such areas, often producing false negatives where correct reasoning is provided alongside incorrect answers. This study seeks to address these gaps by investigating GPT-4's performance on FCI questions, focusing on its reasoning patterns and the occurrence of false positives and false negatives. Depending on the above literature review from different dimensions, we formulated following research question: What is the performance of GPT-4 on FCI questions in terms of correctness, reasoning, false positives, and false negatives?

METHOD

Participants and Instrument

There were no human participants in this study, but rather the focus was on the AI system. The AI model used in this study was GPT-4, developed by OpenAI (2023). GPT-4 was tasked with answering the FCI questions across multiple sessions.

The primary instrument used for this study was the FCI, a well-established diagnostic tool in physics education designed to assess students' understanding of Newtonian mechanics. The FCI consists of 30 multiple-choice questions that probe various misconceptions and fundamental principles in physics (Hestenes et al., 1992).

Data Collection

For each item of the FCI, we first captured an image of the item from the PDF document using the Snipping Tool. Subsequently, we uploaded the image version of the item to GPT-4 along with a straightforward prompt that sought both a response and a concise explanation. The items were posed utilizing the GPT-4 version that was released to the public on March 14, 2023. The entire test was presented to GPT-4 across 1200 distinct chat sessions, repeating each of the 30 items 40 times. To prevent GPT-4 from referencing its own previous responses for answers, each item was consistently introduced in a fresh conversation. The performance of GPT-4 was documented, focusing on correct answers, false positives, false negatives, and reasoning.

In this study, we chose to repeat each of the 30 FCI items 40 times. This decision was based on the need to obtain a statistically significant representation of GPT-4's performance while controlling for variability across sessions. Previous research has indicated that AI models, including GPT-4, can provide different responses to the same question depending on session conditions. By selecting 40 trials, we aimed to ensure a reliable measure of performance trends, including false positives, false negatives, and reasoning quality.

Data Analysis

Collected data focused on the accuracy of answers, the quality of reasoning, and the occurrence of false positives and false negatives. First, for the overall performance, the number of correct answers provided by GPT-4 was tallied. Performance trends were identified to highlight areas of strength and weakness. Second, the frequency and distribution of false positives and false negatives were documented. Third, detailed analysis was conducted for specific FCI items where GPT-4 demonstrated notable performance patterns. For example, items where GPT-4 consistently failed or succeeded were examined closely. Fourth, the reasoning provided by GPT-4 qualitatively analyzed to identify common themes, conceptual misunderstandings, and differences in cognitive approaches. Special attention was given to the AI's ability to interpret figures and diagrams, as this was identified as a consistent challenge area.

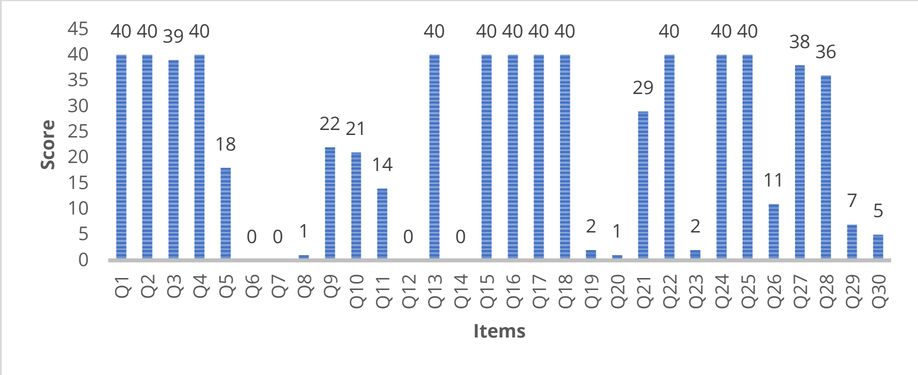

Figure 1. Performance of GPT4 on FCI (Source: Authors)

RESULTS

GPT4's Performance

The study started with an assessment of GPT-4's capability to respond to 30 questions from the FCI. Individual questions were posed one by one in different sessions. However, there are group of question in FCI with same description (for example 5-6, and 8-11 items). They were given together but at different sections each time. **Figure 1** summarizes the performance of GPT-4 across FCI items, offering a visual representation of its accuracy and areas of difficulty.

In this rigorous evaluation, GPT-4 demonstrated a high level of proficiency on several items, accurately responding to questions Q1, Q2, Q4, Q13, Q15, Q16, Q17, Q18, Q22, Q24, and Q25 without mistakes. Its performance was not perfect, however, as it encountered minor challenges, such as a single error in question Q3, two errors in question Q27, and a more noticeable struggle with question Q28, where four mistakes were recorded. Conversely, certain questions proved to be consistent stumbling blocks for GPT-4, with items Q6, Q7, Q12, and Q14 posing significant challenges and receiving no correct answers in any trial.

GPT4's False Positives and False Negatives

When the answer is incorrect and the reasoning is correct, it is called false negative. Similarly, correct answer and incorrect reasoning results in a false positive. We observed only one false positive while many false negatives. In items 7, 12, and 14 all provided answers are false negatives. Among these, in items 7 and 12 GPT-4 always indicated the same wrong choice (that was B for item 7 and C for item 12) while the reasoning was correct. However, in items 14 all reasonings were correct but different wrong answers were provided. For example, in item 14 the incorrect choices varied between choice B (11) and C (29).

False negatives were observed in 14 items (6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 26, 27, 28, and 30). Items 26, 27, and 28 each contained only one false negative while others included false negatives varying between eight and 30. For example, Item 23 had only two correct answers (B) while the rest were false negatives with wrong choices varied between A (22), C (5), D (8), and E (3). GPT4's answers are provided in **Appendix A** with false negatives yellow and false positives green colored. An example of false negative in item 21 is as follows:

> The correct answer is (B). When the rocket engine is turned on at point "b" and produces a constant thrust perpendicular to the initial direction of motion, the rocket will begin to curve in the direction of the thrust. The constant sideways thrust will cause the rocket to follow a curved path that is concave towards the direction of the thrust (to the right, from the point of view of the diagram). This is due to the fact that the rocket is already moving to the right, and the thrust adds a constant acceleration upwards, resulting in a parabolic trajectory (20th trial).

This reasoning is correct, however, path indicated in B is not a curved path, thus, the answer is incorrect.

In 40 trials of each item, false positives were observed only once among the entire responses, that is in 26th item. The item 26 is also the sole item without a figure that includes a false negative. Items 25, 26, and 27 are a series of items related to a woman horizontally pushing a box on a horizontal rough surface with

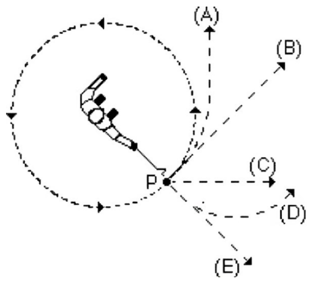

Figure 2. FCI 7th item (Source: Hestenes et al., 1992)

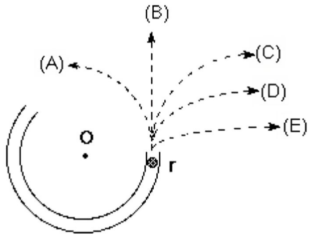

Figure 3. FCI 6th item (Source: Hestenes et al., 1992)

constant velocity. The 26th item asks the movement of the box if the woman doubles the constant horizontal force. GPT-4 provides the correct answer (with a continuously increasing speed) with a wrong reasoning as follows:

> The correct answer is (E). If the woman doubles the force, according to Newton's second law, the acceleration of the box will double, because the mass remains the same ($F = ma$). If the box starts from rest, it will continue to accelerate as long as the force is applied, which means its speed will continuously increase.

Ignoring the friction force and claiming that acceleration will double is incorrect, however, finally stating that speed will continuously increase is correct. Since "with a continuously increasing speed" is the correct choice, we accepted this reasoning as a false positive.

GPT4's Problems with Figures

A distinct area of difficulty for GPT-4 involved interpreting figures, particularly those depicting trajectories. The study determined several questions where GPT-4 failed to correctly analyze illustrations, leading to uniform inaccuracies in its responses. For instance, item 7, which involved predicting the trajectory of a steel ball post-string breakage (**Figure 2**), saw GPT-4 consistently opt for an incorrect path despite valid reasoning based on Newton's first law of motion.

In all 40 trials GPT-4 indicated C as the correct path for the 7th item. The answer is incorrect; however, the reasoning was correct in all trials. Below is the reasoning during the 3rd trial:

> When the string breaks, the steel ball will no longer be under the centripetal force that was keeping it in the circular path. Instead, it will move in a straight line tangent to the circle at the point where the string breaks, due to its inertia. This path corresponds to option (C), which is a straight line that represents the direction the ball was moving at the moment the string broke. This is consistent with Newton's first law of motion, which states that an object in motion will stay in motion with the same speed and in the same direction unless acted upon by an unbalanced force. Since there is no more force acting on the ball once the string breaks, it will continue in a straight line path.

Another evidence for GPT-4's incorrect analyses of figures is seen in item 6 (**Figure 3**). In this item, a ball is shot at high speed into a frictionless channel in the shape of a segment of a circle which is anchored to a frictionless horizontal tabletop. Which path in **Figure 3** would the ball most closely follow after it exits the channel is asked.

In this question GPT-4 sees the curved path C or E as a straight path! The answer is incorrect, however, the reasoning is correct. Below is the answer provided in the 18th trial:

> After the ball exits the channel at "r," the ball will continue to move in a straight line in the direction of the tangent to the circle at point "r" due to inertia (Newton's first law of motion). There are no forces acting on the ball in the horizontal direction (since the table is frictionless and air resistance is negligible), so it will not follow a curved path after exiting the channel. Therefore, the ball will move in a straight line, which corresponds with path (C).

GPT4's Problems with Horizontal Planes

GPT-4 sometimes omits the horizontal table under objects, argues the downward effect of gravity, and claims that gravity curves the ball's path. An example of this reasoning was provided on the 20th trial in item 6:

> Once the ball exits the channel, it will follow a path determined by its velocity at the point of exit and the gravitational force acting on it. Because the table is frictionless, no horizontal force will be acting on the ball after it leaves the channel, meaning it will not accelerate in the horizontal direction. The only force acting on it will be gravity, which will pull it down in a vertical direction. The combination of its horizontal velocity and vertical acceleration due to gravity will cause the ball to follow a parabolic trajectory, which is option (E). This path is consistent with projectile motion.

Another example in which GPT-4 ignores the horizontality is seen in item 8. In this item, a hockey puck is sliding at constant speed on a frictionless horizontal surface. The puck receives a swift horizontal kick in a direction perpendicular to its motion. The question asks which path the puck would most closely follow. GPT-4's incorrect reasoning, where it ignores the reaction of the table and includes only gravity to curve the puck's path is, as follows:

> The correct path the puck would most closely follow after receiving the kick is (E). This is because the puck already has a horizontal velocity (v_0) and then receives an additional kick upward. The result of these two motions—horizontal and vertical—is a resultant vector that causes the puck to follow a parabolic trajectory due to gravity, assuming no other forces act upon it after the kick (13th trial).

Conceptual Mistakes by GPT4

The most obvious conceptual mistake by GPT-4 is its belief that as speed increases, the friction force also increases. This is evident in item 26. For example, the reasoning in the 35th trial is as follows:

> Doubling the constant horizontal force would initially result in an unbalanced force acting on the box, which, according to Newton's second law of motion ($F = ma$), would cause the box to accelerate. However, as the box accelerates, the force of friction may increase until it balances the applied force again, leading to a new constant speed. The new constant speed would not necessarily be double the original speed, because the relationship between force and speed is not direct; it's the relationship between force and acceleration that's direct. Therefore, the correct answer is:
>
> (B) with a constant speed that is greater than the speed "v_0" in the previous question, but not necessarily twice as great.

Another conceptual mistake, repeated five times in item 11, is GPT-4's claim that a frictionless surface results in no normal force from the surface. This can be seen in the 31st trial's reasoning as follows:

> Along the frictionless path you have chosen in question 8, the main force(s) acting on the puck after receiving the kick is/are (B) a downward force of gravity. On a frictionless surface, the only force acting on the puck would be gravity, which acts downwards. There would be no horizontal forces

> because friction is absent, and no upward forces from the surface if we are considering it to be an idealized frictionless surface.

Another conceptual mistake exists in providing answers to item 29. GPT-4 claims that there is a net downward force of air pressure on the chair. The following reasoning is from the 12th trial:

> For question 29, the correct answer is (D) 1, 2, and 3. An office chair at rest is still subject to the force of gravity acting downward. The floor exerts an upward normal force to support the chair. And there is also a net downward force exerted by the air, which is the weight of the air above the chair (even if minimal).

Impetus Theory in GPT-4

Another conceptual mistake is that GPT-4 thinks there is impetus, one of the Aristotelian principles of motion (Jung, 2020). This conception was assessed in item 5, which asks which forces are acting on the ball when it is within the frictionless channel. Some of the results, including the impetus concept, provided by GPT-4 are as follows:

> A force in the direction of motion, which would be the ball's inertia carrying it forward along the channel (15th trial).

> The force in the direction of motion is indeed acting on the ball because the ball is moving in that direction. In a frictionless environment, the only forces acting on the ball would be those due to gravity and its inertia/motion (17th trial).

> The ball is moving along the channel, so there is indeed a force acting in the direction of the motion due to the ball's inertia (19th trial).

Another instance of the impetus idea is revealed in item 30, where despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court. The question asks which of the forces, (1) the downward force of gravity, (2) a force by the "hit," and (3) a force exerted by the air, are acting on the tennis ball after it has left contact with the racquet and before it touches the ground. Several examples including the impetus concept provided by GPT-4 are as follows:

> The force from the hit, which is the initial force that set the ball into motion, still influences the ball's trajectory until other forces neutralize it (7th trial).

> The force of the hit (2) is what initially propelled the ball into motion and the effects of this force continue to act on the ball until other forces nullify it (17th trial).

> The ball is also experiencing a forward force from the hit, which propels it over the net (22nd trial).

> The force of the hit propels the ball forward and is the reason the ball continues to move after contact with the racket has ceased (25th trial).

Another instance of the impetus claim is found in item 11. For example, in the 18th trial, GPT-4 claims that there is a horizontal force acting on the puck due to the kick delivered. The associated reasoning is as follows:

> The correct answer is (B), a downward force of gravity and a horizontal force in the direction of motion. The main forces acting on the puck are the constant horizontal force applied by the kick and the constant downward force of gravity. There's no upward force from the surface since the puck is on a frictionless surface, which means there's no normal force acting opposite to gravity.

Interesting Behaviors of GPT4

Sometimes GPT-4 exhibits interesting behavior by presenting both correct and incorrect reasoning together. This is evident, for example, in the response to item 8. The following is the result from the 32nd trial:

> The correct answer is (E). When the puck receives a horizontal kick, it gains a horizontal velocity component while maintaining its initial vertical velocity due to gravity. The combination of these two perpendicular velocity components (horizontal from the kick and vertical from gravity) results in a diagonal path. The puck will follow a parabolic trajectory because of the influence of gravity pulling it downwards as it moves horizontally, which is shown in option (E).

As seen in the above reasoning, GPT-4 initially provides the correct answer but then makes a mistake by integrating gravity into its explanation.

Another interesting behavior is that GPT-4 initially makes a mistake and then corrects itself. For example, in item 9, the 40th trial is as follows:

> The speed of the puck just after it receives the kick is (C) equal to the arithmetic sum of the speeds “v_0” and “v_x”. Since the puck was already moving with speed “v_0” and receives an additional horizontal speed “v_x” from the kick, the total speed immediately after the kick can be found by using the Pythagorean theorem for the resultant velocity vector because the two velocities are perpendicular to each other. The arithmetic sum is not correct because speed is a scalar quantity that doesn’t involve direction.

It initially claims the arithmetic sum of the speeds, then corrects itself by incorporating the Pythagorean theorem, and finally explains why the initial analysis was not correct by stating that speed is a scalar quantity that doesn’t involve direction.

GPT4’s Extremely Successful Instances

In some of the mechanics concepts, GPT-4 is extremely successful. For example, except for four incorrect responses in item 28, it provided correct and informative results in items 4, 15, 16, and 28, which include concepts related to Newton’s third law. For instance, in the second trial of the fourth item, GPT-4 provided the following reasoning:

> The correct answer is (E) the truck exerts the same amount of force on the car as the car exerts on the truck. This is because according to Newton’s Third Law of Motion, whenever two objects interact, they exert equal and opposite forces on each other. So, when a large truck collides head-on with a small compact car, the force exerted by the truck on the car is equal in magnitude and opposite in direction to the force exerted by the car on the truck. The difference in damage between the truck and the car is due to the differences in their masses and the resultant accelerations from the forces, not because of a difference in the magnitude of the forces.

Similarly, in 15th and 16th items, it provided following reasonings:

> For question 15, the correct answer is (A): the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car. This is based on Newton’s third law of motion, which states that for every action, there is an equal and opposite reaction. So the force the car exerts on the truck is matched by an equal force exerted by the truck on the car.

> For question 16, the correct answer is also (A): the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car. Even when the car and truck have reached a constant cruising speed, Newton’s third law still applies. The forces between them remain equal and opposite, regardless of whether they are accelerating or moving at a constant speed.

DISCUSSION

The assessment of GPT-4’s performance on the FCI indicates its strengths in certain areas of physics while also revealing significant areas for improvement. The analysis of false positives, false negatives, conceptual mistakes, and difficulties with figures and horizontal planes stresses the need for further refinement of GPT-4’s understanding of physics concepts. Nonetheless, its success in correctly applying Newton’s laws demonstrates its potential as a valuable tool for physics education, provided that its limitations are acknowledged and addressed.

Our findings align with previous research on AI and education. Studies have shown that AI models, including GPT-4, can perform well on standard problems but struggle with conceptual understanding (Bengio et al., 2020; Schoenfeld, 2018). This study confirms these results, highlighting GPT-4's limitations in handling conceptual questions involving figures and spatial reasoning.

Additionally, research by Mahligawati et al. (2023), and Wink and Bonivento (2023) indicate the potential of AI to revolutionize physics education by providing related instructional support and feedback, helping students overcome common misconceptions and deepen their understanding of core principles. However, studies by Polverini and Gregorcic (2024) and Kortemeyer (2023) also highlight the challenges AI faces in interpreting visual information and applying conceptual knowledge, which this study corroborates.

An unexpected result was the frequency of false negatives in GPT-4's responses. While GPT-4's reasoning was often correct, its answers were incorrect due to misinterpretations of the questions or figures. This finding suggests that GPT-4's underlying algorithms may not fully grasp the important points of physical concepts, especially those involving visual elements. This result is significant as it points out the limitations of current AI models in educational settings, particularly in subjects requiring deep conceptual understanding.

Examining GPT-4's responses to the FCI can provide strong evidence for the validity and reliability of this widely-used diagnostic tool. By analyzing the AI's performance, we can identify patterns in both correct and incorrect answers, which may point out underlying issues in the question design or common misconceptions that are not adequately addressed. For instance, if GPT-4 consistently selects incorrect answers with plausible reasoning, this could indicate ambiguities in the questions or reveal deeper misconceptions that students might also hold. Enhancing the educational instruments' validity and reliability through such analysis can lead to more accurate assessments of students' conceptual understanding, ultimately informing more effective instructional strategies and interventions. Furthermore, this approach aligns with the growing body of research emphasizing the importance of robust assessment tools in education (Hestenes et al., 1992; McDermott & Redish, 1999). As AI continues to evolve, benefiting from its capabilities to refine educational assessments holds significant potential for improving learning outcomes (Holmes et al., 2019).

A question that may arise from readers is whether there were common topics on which GPT-4 performed well or poorly. First, the items that resulted in completely incorrect answers were all questions that included drawings (Q6, Q7, Q12, and Q14). Referring to the original FCI article (Table I), question 6 and question 7 pertain to the same Newtonian concept (impulsive force), while questions 12 and 14 are related to different concepts. Second, again referring to the original FCI article, the items on which GPT-4 performed well or poorly did not group according to either Newtonian concepts (Table I in the original article) or the taxonomy of misconceptions (Table II in the original article).

On the other hand, in 40 trials, the absence of correct answers in items Q6, Q7, Q12, and Q14 is a significant issue for AI-assisted learning. This raises substantial doubts about the validity of GPT-4. Additionally, in 40 trials, GPT-4 was 100% successful in items Q1, Q2, Q4, Q13, Q15, Q16, Q17, Q18, Q22, Q24, and Q25. This varied performance highlights the complex nature of GPT-4's understanding of physics concepts, revealing patterns of strengths and weaknesses in its reasoning capabilities.

CONCLUSIONS

This study investigated the comparative performance of GPT-4 on the FCI to analyze false positives, false negatives, and reasoning patterns. Key findings include: First, GPT-4 exhibited a high level of proficiency in several FCI items, particularly those related to Newton's third law, achieving perfect scores on many items. However, it struggled significantly with questions involving the interpretation of figures and spatial reasoning. Second, GPT-4 demonstrated a higher occurrence of false negatives (correct reasoning but incorrect answers) compared to false positives (incorrect reasoning but correct answers). Third, GPT-4 displayed several conceptual errors, such as misunderstanding the effect of friction and retaining the outdated impetus theory of motion. The study reveals that while GPT-4 can provide accurate answers, it often lacks deep conceptual understanding, leading to false negatives. This suggests that while AI can be a valuable tool in education, it still requires significant refinement to match human problem-solving skills fully.

The potential of AI to promote higher-order learning extends beyond its traditional role as an information retrieval tool. In this study, we explored how GPT-4, an advanced AI model, can contribute to deeper conceptual understanding in physics. While AI is often perceived primarily as a means to enhance recall and provide factual information, our analysis of GPT-4's performance on the FCI indicates its capacity to engage with complex reasoning and problem-solving tasks. These findings suggest that AI-driven tools can be designed to support and enhance conceptual learning, making them valuable assets in educational settings aimed at fostering deeper understanding rather than mere memorization (Anderson et al., 2001). This aligns with the broader educational goal of promoting critical thinking and problem-solving abilities among students.

The FCI, while widely used, may not fully capture the breadth of GPT-4's conceptual understanding and reasoning skills. Additionally, the study relied on the accuracy of AI's responses to the FCI without considering its performance on other types of assessments. GPT-4's performance may not fully represent the capabilities of AI in education, as different models and versions may exhibit varying levels of proficiency.

AI can provide immediate feedback and alternative explanations that human teachers might not consider, while human learners can offer insights into the nuanced understanding of concepts that AI lacks. This synergy can lead to more effective learning experiences.

Future research can identify the discrepancies between AI and student responses, revealing potential ambiguities or weaknesses in the questions. This process will allow for refinement and better alignment with educational goals.

Author contributions: SN: conceptualization, data curation, and writing – review & editing; **GI:** methodology, writing – review & editing; **SM:** writing – original draft; **NB:** Formal analysis, and supervision. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: The authors declared that this study was exempt from ethics committee approval since no human or animal data was used in this study.

Declaration of interest: The authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Anderson, M., Anderson, S. L., & Armen, C. (2019). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 40*(4), 45–52.

Balta, N. (2024). A short review of AI in education: Perspectives from the Web of Science database. *The European Educational Researcher*, *7*(2), 41–43. https://doi.org/10.31757/euer.723

Balta, N., & Eryılmaz, A. (2017). Counterintuitive dynamics test. *International Journal of Science and Mathematics Education, 15*, 411–431. https://doi.org/10.1007/s10763-015-9694-6

Bengio, Y., Lavoie, P., & Vincent, P. (2020). Learning neural networks to solve differential equations. *Journal of Machine Learning Research, 21*(1), 3485–3510.

Boehnlein, A., Diefenthaler, M., Fanelli, C., Hjorth-Jensen, M., Horn, T., Kuchera, M. P., Lee, D., Pang, L.-G., Poon, A., Sato, N., Schram, M., Scheinker, A., Smith, M. S., Wang, X.-N., & Ziegler, V. (2021). Artificial intelligence and machine learning in nuclear physics. *arXiv*. https://doi.org/10.1103/RevModPhys.94.031003

Buabeng, I. (2018). Physics classroom interactions: Teaching strategies and practices. *Journal of Research in Science, Mathematics and Technology Education*, *1*(3), 311–328. https://doi.org/10.31756/jrsmte.134

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1989). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research, 19*(1), Article 010142. https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

de los Ángeles Domínguez-González, M., Hervás-Gómez, C., Díaz-Noguera, M. D., & Reina-Parrado, M. (2023). Attention to diversity from artificial intelligence. *The European Educational Researcher*, *6*(3), 101–115. https://doi.org/10.31757/euer.633

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics-Physics Education Research, 10*(2), Article 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Ge, Z., & Hu, Y. (2020). Innovative application of artificial intelligence (AI) in the management of higher education and teaching. *Journal of Physics: Conference Series, 1533*(3), Article 032089. https://doi.org/10.1088/1742-6596/1533/3/032089

Geiger, P., Willner, J., & Kuhn, D. (2021). Misconceptions in physics: A comparative analysis of human and AI reasoning. *Physics Education Research, 23*(2), 231–245.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74. https://doi.org/10.1119/1.18809

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53*(11), 1043–1055. https://doi.org/10.1119/1.14030

Hammer, D. (1996). Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions. *Journal of the Learning Sciences, 5*(2), 97–127. https://doi.org/10.1207/s15327809jls0502_1

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–151. https://doi.org/10.1119/1.2343497

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Jung, E. (2020). Impetus. In H. Lagerlund (Ed.), *Encyclopedia of medieval philosophy: Philosophy between 500 and 1500* (pp. 832–835). Springer. https://doi.org/10.1007/978-94-024-1665-7_239

Kortemeyer, G. (2023). 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

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance–Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. *Frontiers in Artificial Intelligence and Applications, 386*, 199–212. https://doi.org/10.3233/FAIA240195

Kuzu, S. Y. (2021). Artificial intelligence based machine learning approach in high energy physics. *International Journal of Innovative Engineering Applications, 5*(2), 176–180. https://doi.org/10.46460/ijiea.929292

Lai, J. W., & Cheong, K. H. (2022). Educational opportunities and challenges in augmented reality: Featuring implementations in physics education. *IEEE Access, 10*, 43143–43158. https://doi.org/10.1109/ACCESS.2022.3166478

Lample, G., & Charton, F. (2019). Deep learning for symbolic mathematics. *arXiv*. https://doi.org/10.48550/arXiv.1912.01412

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Mahligawati, F., Allanas, E., Butarbutar, M. H., & Nordin, N. A. N. (2023). Artificial intelligence in physics education: A comprehensive literature review. *Journal of Physics: Conference Series, 2596*(1), Article 012080. https://doi.org/10.1088/1742-6596/2596/1/012080

McDermott, L. C., & Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics, 67*(9), 755–767. https://doi.org/10.1119/1.19122

Mustofa, H. A., Bilad, M. R., & Grendis, N. W. B. (2024). Utilizing AI for physics problem solving: A literature review and ChatGPT experience. *Lensa: Jurnal Kependidikan Fisika, 12*(1), 78–97. https://doi.org/10.33394/j-lkf.v12i1.11748

OpenAI. (2023). GPT-4: Technical report. *OpenAI*. https://cdn.openai.com/papers/gpt-4.pdf

Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the test of understanding graphs in kinematics. *Physical Review Physics Education Research, 20*(1), Article 010109. https://doi.org/10.1103/PhysRevPhysEducRes.20.010109

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education, 26*(2), 582–599. https://doi.org/10.1007/s40593-016-0110-3

Rosé, C. P., Resnick, L., Goldman, P., & Sherin, B. L. (2019). The future of AI in education: Integrating technology and human judgment. In R. Sharpe, H. Beetham, & S. de Freitas (Eds.), *Rethinking learning in the digital age* (pp. 265–293). Springer.

Schoenfeld, A. H. (2018). On reasoning and sense making in mathematics and science: Themes and highlights. *International Journal of STEM Education, 5*(1), 3–13.

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield, & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 15–31). Academic Press. https://doi.org/10.1016/B978-012750053-9/50003-6

Smith, T. I., & Knight, R. D. (2021). Using computer simulations to improve physics learning. *Journal of Science Education and Technology, 30*(3), 346–358.

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research, 19*(2), Article 020123. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

Van der Veen, J. T., & Van den Berg, E. (2021). Enhancing conceptual understanding with computer simulations in physics education. *Physics Education, 56*(1), Article 015011.

Wang, L. (2020). Artificial intelligence and career development of college teachers: Challenge and countermeasures. *Journal of Physics: Conference Series, 1550*(2), Article 022030. https://doi.org/10.1088/1742-6596/1550/2/022030

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

Wink, R., & Bonivento, W. M. (2023). Artificial intelligence: New challenges and opportunities in physics education. In M. Streit-Bianchi, M. Michelini, W. Bonivento, & M. Tuveri, M. (Eds.), *New challenges and opportunities in physics education. Challenges in physics education* (pp. 427–434). Springer. https://doi.org/10.1007/978-3-031-37387-9_27

Wulff, P. (2024). Physics language and language use in physics–What do we know and how AI might enhance language-related research and instruction. *European Journal of Physics, 45*(2), Article 023001. https://doi.org/10.1088/1361-6404/ad0f9c

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

Yerushalmi, E., Cohen, E., & Singh, C. (2017). Assessing and improving student reasoning in physics. *Physical Review Physics Education Research, 13*(1), Article 010121.

Yilmaz, H., Maxutov, S., Baitekov, A., & Balta, N. (2023). Student's perception of Chat GPT: A technology acceptance model study. *International Educational Review, 1*(1), 57– 83. https://doi.org/10.58693/ier.114

Zanca, F., Avanzo, M., Colgan, N., Crijns, W., Guidi, G., Hernandez-Giron, I., Kagadis, G. C., Diaz, O., Zaidi, H., Russo, P., Toma-Dasu, I., & Kortesniemi, M. (2021). Focus issue: Artificial intelligence in medical physics. *Physica Medica: European Journal of Medical Physics, 83*, 287–291. https://doi.org/10.1016/j.ejmp.2021.05.008

Zohar, A., & Dori, Y. J. (2012). *Metacognition in science education: Trends in current research*. Springer. https://doi.org/10.1007/978-94-007-2132-6

APPENDIX A: GPT4'S ANSWERS

Table A1. GPT4's answers

Q	CA	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20
Q1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q2	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q3	C	C	C	C	C	C	E	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q4	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
Q5	B	E	C	E	E	B	D	B	D	D	B	E	B	C	E	D	E	C	B	E	B
Q6	B	E	E	C	E	C	E	C	C	C	C	C	C	E	C	E	C	E	C	C	E
Q7	B	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q8	B	E	E	E	E	E	E	E	E	E	E	C	E	E	E	E	E	E	E	E	E
Q9	E	D	E	E	E	E	E	E	E	E	C	E	E	E	C	E	E	C	C	E	D
Q10	A	A	A	A	A	A	A	A	A	A	A	B	C	C	B	C	D	A	E	C	A
Q11	D	E	E	E	E	E	E	E	E	E	E	D	D	C	D	A	C	D	B	D	C
Q12	B	C	C	C	C	C	C	C	D	C	D	C	C	C	C	C	C	C	C	C	C
Q13	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
Q14	D	B	B	C	B	C	C	B	C	C	B	C	B	C	C	C	C	B	C	C	C
Q15	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q16	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q17	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q18	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q19	E	C	C	C	C	C	C	C	C	D	C	D	D	D	C	E	D	C	C	E	D
Q20	D	A	A	B	B	A	B	C	A	C	B	B	B	B	B	B	B	B	A	B	B
Q21	E	E	E	C	E	B	E	D	E	E	E	E	D	D	E	E	E	D	E	B	B
Q22	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q23	B	A	A	E	A	E	A	C	A	A	A	D	E	A	D	D	A	C	A	C	D
Q24	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q25	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q26	E	B	B	E	B	B	B	B	E	B	B	E	E	C	E	B	C	C	E	E	C
Q27	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	B	C	C
Q28	E	E	E	E	E	E	C	E	E	E	E	E	E	E	E	E	E	E	E	E	E
Q29	B	D	D	B	D	D	D	D	B	D	D	D	D	D	B	D	B	D	D	D	B
Q30	C	E	E	C	E	D	E	E	C	E	A	D	D	E	E	E	C	E	E	E	E

Q	CA	G21	G22	G23	G24	G25	G26	G27	G28	G29	G30	G31	G32	G33	G34	G35	G36	G37	G38	G39	G40
Q1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q2	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q3	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q4	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
Q5	B	E	B	B	B	D	A	B	B	B	E	D	C	C	B	C	B	B	B	B	B
Q6	B	C	C	E	D	C	E	C	C	C	C	C	E	E	C	E	C	C	C	E	E
Q7	B	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q8	B	C	E	E	C	C	E	C	E	E	C	E	E	E	E	E	B	E	E	E	E
Q9	E	E	C	C	E	E	E	E	B	D	E	C	C	C	C	B	E	C	E	C	C
Q10	A	B	D	A	A	A	A	A	A	A	C	D	B	D	D	A	C	E	D	D	A
Q11	D	A	B	D	D	D	D	D	D	B	B	B	D	A	C	C	D	D	C	A	C
Q12	B	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q13	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
Q14	D	C	C	C	C	C	C	C	C	C	C	B	B	C	C	C	B	C	C	B	C
Q15	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q16	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q17	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q18	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q19	E	D	C	C	C	C	C	C	C	C	D	C	C	C	C	D	C	D	C	C	D
Q20	D	C	B	D	B	B	B	B	C	B	B	B	B	A	C	B	B	A	B	C	B
Q21	E	E	E	E	E	E	E	E	D	D	E	B	E	E	E	E	E	E	E	E	E
Q22	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Q23	B	A	D	C	A	A	A	D	B	B	C	D	D	A	A	A	A	A	A	A	A
Q24	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Q25	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Q26	E	C	C	C	E	C	D	E	D	C	E	C	B	C	E	B	C	D	B	C	A
Q27	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	B	C
Q28	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	C	E	E
Q29	B	B	D	D	D	D	D	D	D	B	D	D	D	D	D	D	D	D	D	D	D
Q30	C	C	E	E	E	E	D	D	E	E	E	E	E	E	D	D	E	C	D	D	D

Note. Yellows are false negatives; Greens are false positives; Q: Question; CA: Correct answer; G: GPT.