

ChatGPT reflects student misconceptions in physics

Sav Wheeler and Rachel Scherr

School of STEM, University of Washington Bothell, 17927 113th AVE NE, Bothell, WA, 98011

With many students turning to machine learning models such as ChatGPT for homework help, diagnosing the accuracy of the information these technologies purport across different fields is more important than ever. In this study, we find that ChatGPT's responses to standard physics questions are consistent with an accurate theoretical understanding, but are often incorrect when applied to a given system, reflecting many of the same misconceptions as students.

I. INTRODUCTION

Released in November of 2022, ChatGPT is a trained AI (artificial intelligence) model based on the large language model GPT-3.5, with a focus on training for a dialogue-based conversational format [1]. The AI gained widespread popularity soon after its release, including among students, who frequently used the AI for academic help [2, 3]. The potential for student use of GPT as a writing tool has sparked fears among educators of a new wave of plagiarism and cheating [3], though text-checker tools have since been developed to mitigate this [4, 5]. ChatGPT has gained academic notoriety for passing or excelling at several standardized tests such as the SAT, GRE, and multiple AP tests [6]. However, these benchmark tests may not be representative of field-specific student understandings, and in several cases, ChatGPT has not only failed an exam, but performed worse on average than student responses [7].

Haensch et al.'s [8] evaluation of social media content relating to ChatGPT, in an effort to gauge student views on use of the program, suggests that students generally view ChatGPT positively in its applications and use for writing essays, generating code, and answering questions. Alongside this, their study suggests an under-representation on social media of ChatGPT's capacity to generate misinformation in these applications. Bridging the gap between the content-critical academic perspective and opportunistic student perspectives on this technology will be essential, going forward, in ensuring students know the limitations of the tools at their disposal, and develop a degree of awareness of how generative AI technology can affect their learning for better or worse.

In this study, we analyze the performance of ChatGPT on the Force Concept Inventory (FCI) in order to gauge how this AI might affect student learning of key physics concepts by reflecting common-sense misconceptions. While this inventory is not necessarily representative of the complete range of topics and concepts in introductory physics courses, it contains a breadth of essential foundational concepts such as Newtonian laws, kinetics, and momentum. We focus on the categorical performance of ChatGPT on sets of questions in contrast to overall performance as well as a qualitative analysis of specific prompt responses that may be consistent with common student misconceptions.

The purpose of this study is to quantify ChatGPT's potential to inform or mislead students who use the AI. The quantitative analysis can provide a general overview of the accuracy of GPT's responses, while the qualitative analysis can provide a more in-depth look at what particular problem-solving skills or frameworks ChatGPT's responses appear to employ, and whether these resources are consistent with students' own.

II. METHOD

We prompted ChatGPT with questions based on the FCI. Many of the questions were based on diagrams, so written

transcriptions of the diagrams were given for any question which depended on one, including any questions with answers that were a part of the diagram or other representative visuals. All prompts are included with ChatGPT's responses on the website listed in the Data section of this paper.

Five standard trials were run for the exam, where all questions for a trial were given in the same chatroom, including question numbers and lettered options. Three "variant" trials were also conducted, which tested different experimental changes to the testing method. Variant trial 1 had ChatGPT answer the questions individually in separate chatrooms. Variant trials 2 and 3 included a preliminary "role-play" prompt to see if ChatGPT performs better or worse when taking on a given role, which were respectively:

- Variant 2: "Pretend you are a first-year physics student taking an exam. Answer the following questions with your letter selection. I'll start:"
- Variant 3: "Pretend you are a physics professor creating an answer key for an exam. Write the correct letter answers to the following questions. I'll start:"

III. RESULTS

Table I shows ChatGPT's letter answers to each FCI question prompt, with incorrect responses labeled in bold. Cells labeled "N" indicate responses where no letter answer was given. Exact prompts used and ChatGPT's written responses to each prompt can be found on the following website: <https://lucinder.github.io/CapstoneStudy/StudyData.html>

A. Overall Performance

ChatGPT scored generally poorly on the Force Concept Inventory ($\bar{x} = 51\%$, $\sigma = 8.5\%$), with a distribution of predicted scores falling within 44%-59% with 99% confidence. An FCI score of 60% is considered to be an entry threshold to Newtonian physics [9]. Notably, GPT had a slightly lower median score ($\hat{x} = 48\%$) than mean score, indicating a slight right skew in scores. Scores on variant trials were 10.3% worse on average than scores on standard trials.

B. FCI Cluster Performance

The original publication of the FCI outlines six "clusters" of concepts, as well as indicating which questions contain answers that rely on an understanding of each concept [10]. Based on ChatGPT's responses to questions containing answers corresponding to one or more of these concept clusters, we recorded the metrics of ChatGPT's performance on different clusters.

Table II shows ChatGPT's average scores on each FCI cluster along with their number of questions (n_q), standard deviations, and 99% confidence intervals of scores. Fig. 1

TABLE I. ChatGPT’s letter responses to each FCI question. Incorrect responses are indicated in bold.

Trial	Question No.																												Score (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Correct Response	C	E	A	B	D	B	E	A	D	B	E	B	A	A	C	B	C	B	B	E	D	D	D	E	B	B	A	C	C
1	C	E	A	B	D	E	B	A	D	B	E	B	B	A	B	C	B	B	C	D	C	A	B	E	B	A	A	C	C
2	C	E	A	A	C	D	C	A	B	B	E	B	B	A	B	B	B	B	D	A	B	B	E	A	A	A	C	B	
3	C	E	A	A	D	C	B	A	D	B	E	B	A	A	B	N	B	B	N	E	C	B	B	A	C	A	A	D	B
4	C	E	A	B	D	B	E	A	B	B	E	B	A	A	B	B	B	B	A	A	E	A	D	E	A	A	A	C	C
5	C	E	A	D	C	D	C	A	B	B	E	B	A	A	B	C	B	B	A	E	D	A	B	B	A	A	A	C	C
Variant 1	C	A	D	D	C	D	B	A	D	B	C	B	A	A	C	B	B	B	C	B	D	D	B	D	A	E	A	D	C
Variant 2	C	E	A	A	C	C	A	A	D	B	E	B	A	A	B	C	B	B	N	E	E	A	B	C	A	C	C	C	B
Variant 3	C	E	A	A	C	D	B	A	D	B	E	B	A	A	B	N	B	B	N	A	C	A	B	C	A	A	A	C	B

TABLE II. Key figures from ChatGPT’s cluster performance

Concept Cluster	n_q	\bar{x}	σ	99% CI
Kinematics	7	23%	15%	[9.0%, 37%]
Newton’s First Law	8	63%	13%	[50%, 75%]
Newton’s Second Law	4	19%	29%	[0.0%, 45%]
Newton’s Third Law	4	88%	19%	[70%, 100%]
Superposition Principle	4	63%	13%	[50%, 75%]
Types of Force	12	50%	10%	[41%, 59%]

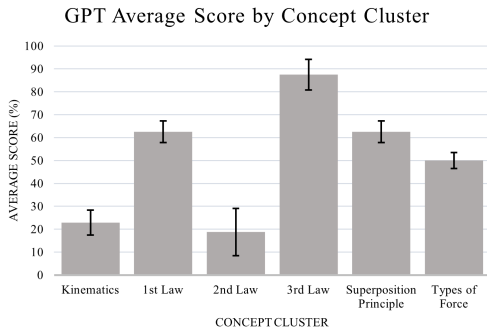


FIG. 1. ChatGPT’s average score in each concept cluster.

also shows the average score on questions in each concept cluster by ChatGPT, with margins of error indicated by thin lines on bars. ChatGPT notably performed poorest on questions that probed concepts of Kinematics and Newton’s Second Law.

IV. ANALYSIS

ChatGPT’s consistent incorrect responses may result in students receiving incorrect or confusing information when using ChatGPT as a learning tool for physics. The follow-

ing sections discuss the potential causes and implications of ChatGPT’s incorrect responses, and how educators can adapt to widespread student use of ChatGPT with these limitations in mind.

In some cases, ChatGPT’s responses to FCI questions represent accurate theoretical understanding of concepts in physics, and its explanations may be helpful to students. For example, when asked to compare the collision forces between a large truck and small compact car, ChatGPT explains,

According to Newton’s Third Law of Motion, for every action, there is an equal and opposite reaction. Therefore, during the collision between the truck and the car, both vehicles will exert equal and opposite forces on each other. The truck exerts the same amount of force on the car as the car exerts on the truck.

In other cases, though, ChatGPT responses are likely to mislead learners. ChatGPT’s responses to kinematics questions reflected common student difficulties distinguishing velocity and acceleration, as well as a tendency to add velocities as scalars rather than vectors [10]. For instance, when asked about a constant-velocity hockey puck that receives a horizontal “kick” in the direction horizontally perpendicular to its existing path in a question probing students’ ability to interpret motion as vectors, ChatGPT explained,

Since the surface is frictionless, the horizontal component of the velocity of the puck should remain constant throughout its motion. Therefore, the speed of the puck just after it receives the kick should be the arithmetic sum of the speed it had before the kick and the speed it acquires from the kick. Thus, the answer is the arithmetic sum of speeds v_0 and v .

In another response to the hockey puck scenario, ChatGPT interpreted the instantaneous “kick” to be a long-lasting force that produces circular motion, which may reflect the

student common-sense misconception of confusing motion with force [11]:

When the hockey puck receives an instantaneous horizontal “kick” in the direction horizontally perpendicular to its existing path at point B, it will start to move in a curved path. Since the surface is frictionless, there is no force to slow down or stop the puck’s motion. The puck will continue moving with a constant speed but changing direction due to the force acting on it.

Some common student misconceptions concern object behavior in outer space, in particular the common sense misconception that objects are weightless in space or that Newton’s Second Law does not apply in the context of outer space [12, 13]. In response to a question about a rocket in outer space firing its engine with constant thrust at a right angle to the rocket’s initial motion, ChatGPT incorrectly states that the rocket’s speed is constant during the thrust:

As the rocket’s engine produces a constant thrust at a right angle to the line [of initial motion], the rocket will move in a circular path... Therefore, the rocket’s speed will be constant, as it moves with a constant angular velocity.

Overall, ChatGPT is an unreliable source of physics reasoning. Sometimes its interpretations veer toward the surreal: in a question regarding a ball on a string that a person swings in a horizontal circle over their head, intended to probe students’ understanding of centripetal force, ChatGPT responds,

According to Newton’s First Law of Motion, an object in motion will continue in a straight line with a constant speed unless acted upon by a net external force. In this case, the heavy ball is moving in a circular path due to the tension force in the string, which provides the necessary centripetal force. When the string suddenly breaks, there is no longer a centripetal force acting on the ball, and it will continue to move tangentially to its path at the moment of release. However, since the person is also moving in the circular path, the ball will also have a component of motion towards the person. Therefore, the path the ball will travel relative to the person is mostly tangential to its path at the moment of release, but curved inwards along the horizontal plane towards the person.

The AI may be interpreting the language of this prompt as similar enough to another “archetype” of question- in this case, misinterpreting a question about horizontal centripetal force as a question about orbits- to attempt to answer

the prompt as if it were responding to the other archetype. Though not necessarily reflecting a misconception, this may confuse students in regards to how the question should be interpreted.

The content of ChatGPT’s explanations in its prompt responses reflects many of the common-sense misconceptions of students, particularly those regarding kinematics, and may mislead students who use the AI as a study tool. Patterns of treating vectors as scalars, confusing momentum with force, and misinterpreting forces in the context of outer space are some of the main misconceptions which ChatGPT may reinforce when responding to physics questions.

V. STUDY LIMITATIONS

Many questions from the FCI relied on, or had answers that relied on, a visual representation or diagram. The public model used by ChatGPT, GPT-3.5, cannot process images, and as such, all diagrams and representations were translated into text to the best of our ability. A lack of access to diagrams may have affected ChatGPT’s scores on these questions, but until such diagrams are tested with GPT-4, it is uncertain how exactly the absence of diagrams would have affected ChatGPT.

All prompts were given to, and responses collected from, the ChatGPT March 23 release. Incremental updates in the AI may slightly change how it responds to certain prompts.

VI. RECOMMENDATIONS FOR EDUCATORS

ChatGPT offers accurate, conversational explanations of physics theory in some cases, and incorrect or misleading explanations in others. Students who use it in a physics context may benefit from its explanations, but also risk having common misconceptions reinforced. One potential educational application of ChatGPT is producing practice problems that require students to identify and correct a problem ChatGPT has solved incorrectly. ChatGPT’s responses are worded convincingly and often purport accurate theoretical explanations to supplement incorrect practical explanations, and such questions would therefore challenge students to think critically about the information presented in ChatGPT’s response and formulate a logical flow of operations for solving the same problem. This would allow students to further familiarize themselves with applying concepts in physics, as well as giving them practice in fact-checking in these fields.

In general, an outright ban of ChatGPT in the classroom is unlikely to dissuade students from using the AI for their own purposes outside of the classroom; thus, giving students the tools to reconcile their own understandings with information purported by the AI would be highly beneficial to students who do use the AI regularly, especially as this technology becomes more integrated in our daily lives in the future.

-
- [1] OpenAI, <https://openai.com/blog/chatgpt>. Retrieved 4/17/2023.
- [2] R. Brandl and C. Ellis, <https://www.tooltester.com/en/blog/chatgpt-statistics/>. Retrieved 4/17/2023.
- [3] M. Sullivan, A. Kelly, and P. McLaughlan, *J. Appl. Learning and Teaching*, **6**, 1 (2023).
- [4] ZeroGPT, <https://www.zerogpt.com/>. Retrieved 4/17/2023.
- [5] Writer, <https://writer.com/ai-content-detector/>. Retrieved 4/17/2023.
- [6] OpenAI, arXiv:2303.08774
- [7] P.M. Newton and M. Xiromeriti, <https://edrxiv.org/sytu3>. Retrieved 4/17/2023.
- [8] A.C. Haensch, S. Ball, M. Herklotz, and F. Kreuter, arXiv:2303.05349.
- [9] D. Hestenes and I. Halloun, *The Phys. Teach.*, **33**, 8 (1995).
- [10] D. Hestenes, M. Wells, and G. Swackhamer, *The Phys. Teach.*, **30**, 3 (1992).
- [11] G. Liu and N. Fang, *Int. J. of Eng. Education*, **32**, 1 (2016).
- [12] S. Gönen, *J. of Sci. Education Technol.*, **17**, 70 (2008)
- [13] B.K. Temiz and A. Yavuz, *European J. of Phys.*, **35**, 4 (2014)