Could an artificial-intelligence agent pass an introductory physics course?

Gerd Kortemeyer *,†

Educational Development and Technology, ETH Zurich, 8092 Zurich, Switzerland

(Received 2 February 2023; accepted 21 March 2023; published 11 May 2023)

Massive pretrained language models have garnered attention and controversy due to their ability to generate humanlike responses: Attention due to their frequent indistinguishability from human-generated phraseology and narratives and controversy due to the fact that their convincingly presented arguments and facts are frequently simply false. Just how humanlike are these responses when it comes to dialogues about physics, in particular about the standard content of introductory physics courses? This case study explores that question by having ChatGPT, the preeminent language model in 2023, work through representative assessment content of an actual calculus-based physics course and grading the responses in the same way human responses would be graded. As it turns out, ChatGPT would narrowly pass this course while exhibiting many of the preconceptions and errors of a beginning learner. A discussion of possible consequences for teaching, testing, and physics education research is provided as a possible starter for more detailed studies and curricular efforts in the future.

DOI: 10.1103/PhysRevPhysEducRes.19.010132

I. INTRODUCTION

"Educators may have concerns about ChatGPT, a large language model trained by OpenAI, for a number of reasons. First and foremost, there is the concern that a tool like ChatGPT could potentially be used to cheat on exams or assignments. ChatGPT can generate humanlike text, which means that a student could use it to produce a paper or response that is not their own work. This could lead to a breakdown in the integrity of the educational system and could undermine the value of a degree or diploma." These sentences were not written by the author, but by ChatGPT (Generative Pre-trained Transformer) [1] itself in response to the prompt "Write an essay why educators would be concerned about ChatGPT." The chatbot goes on to explain how it could spread misinformation, inhibit the development of writing skills, and replace human educators, particularly when it comes to grading.

ChatGPT is based on the GPT-3 language model. Its predecessor, GPT-2, was purely a system to continue stories: given some piece of text, the system would determine plausible follow-up text, similar to the autocomplete on smartphones which suggests plausible next words in a text message (only that GPT does not operate on the level of words, but smaller tokens, which are essentially syllables). The author remembers playing with GPT-2, starting some fairy tale, and seeing where it leads—at the time, the author found GPT-2 entertaining but essentially useless [2]. ChatGPT added the ability to respond to prompts, i.e., not just continuing what the user provided, but responding to questions or commands, and it also added the ability to function across languages by first translating the prompt into English and then translating the system's response back into the original language using other artificial-intelligence systems similar to Google Translate [3] or DeepL [4].

It is easy to forget that the core system of ChatGPT is still the same GPT, which is a neural network that generates plausible fiction based on probabilities of what comes next. There is no cognition happening, and as much as it is convenient to use words like "think" or "assume" in connection with ChatGPT, these expressions need to be understood in the same way as "the spell checker thought I wanted to write 'glues' instead of 'gluons'" or "the autofocus assumed the person in the foreground was the subject"—the words are shorthands to avoid cumbersome constructions such as "the probabilistic algorithm determined," and they do not imply that the algorithm actual "thinks" or "assumes" in the sense of human cognition.

The potential impact of ChatGPT with its custom-built essays on courses in the humanities is evident, but is there also an impact on subjects like physics? First of all, within physics, large problem libraries for cheating have existed for years and they are well known and used by students [5,6]—virtually any physics homework problem ever assigned is available online with solutions and more or

*kgerd@ethz.ch

†Also at Michigan State University, East Lansing, Michigan 48824, USA.

less helpful explanations. So, the primary impact of ChatGPT in physics would not be cheating. On top of that, would artificial intelligence really be able to handle the logical, conceptual, and mathematical challenges that physics entails and would it be able to strategically solve problems [7,8]?

Figure 1 shows sample dialogues with ChatGPT, which is, after all, primarily a chatbot. A welcome feature is that it does not simply provide some answer but that the algorithm attempts to explain how it arrived at the answer. In many respects, this dialogue appears similar to an office-hour conversation between an instructor and a beginning physics student:

- When first asked how far the car is from where it started, the chatbot did not consider that the car may have changed direction. When prompted, it does state that there is missing information.
- The chatbot does plug-and-chug [9,10], putting the numerical results from one equation into the next instead of first deriving a final symbolic answer, which results in rounding errors. As it will turn out, carrying out calculations by immediately putting numbers into formulas is one of the weaknesses of ChatGPT shared with beginning learners of physics.
- The chatbot leaves out physical units during calculations, preventing it from carrying out dimensional analysis.
- The chatbot does not realize that the speed actually drops out when doing the return-time calculation in the last step; instead. The straightforward solution would have been $\sqrt{(3\mathrm{h})^2 + (4\mathrm{h})^2} = 5$ hours (at least, though, the chatbot adds an "approximately" to its solution).
- The chatbot struggles with simplifying symbolic math equations.

How much, indeed, does this particular early 2023 state-of-the-art artificial-intelligence tool resemble the behavior of an introductory physics student? Could it pass a physics course? When posing this question directly to ChatGPT, it answers "as a language model, I have been trained on a

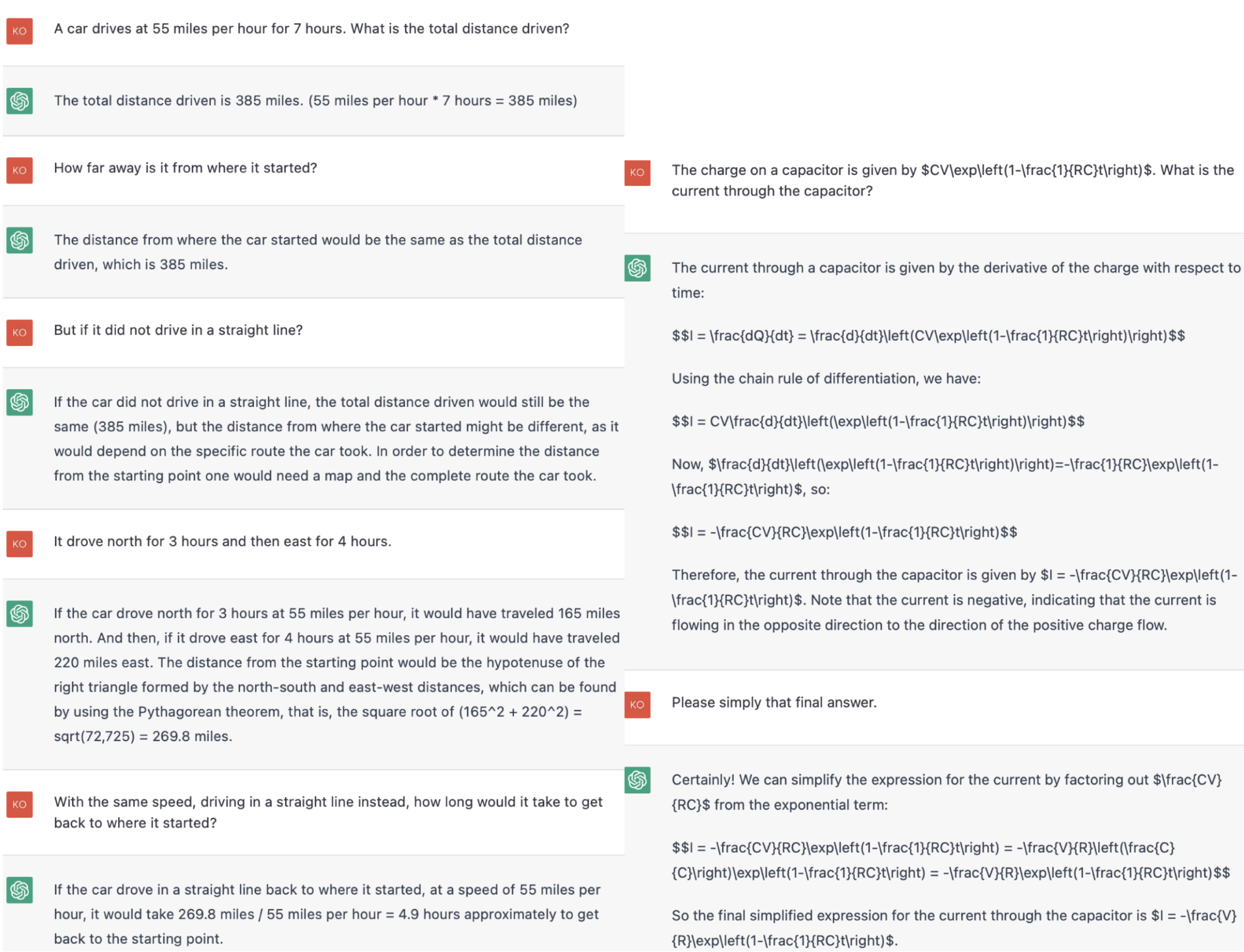


FIG. 1. Sample ChatGPT dialogues about homework problems. The entries labeled with a red "KO" are by the author, the entries labeled in green by ChatGPT. For mathematical expressions, LaTeX was used.

large dataset of text, including physics texts. This allows me to understand and generate text related to physics concepts, but it does not mean that I have the ability to solve physics problems or pass a physics course. I can provide explanations and answer questions about physics to the best of my knowledge, but I am not a substitute for a human physics expert or a physics education." To put this statement to the test, ChatGPT was used to solve representative assessment components of an introductory calculus-based physics; the responses were graded in the context of the assessment types and subjectively compared to responses of human learners.

It is important to note, though, that ChatGPT will not actually learn anything new by "attending" this course, as the system is a "Pre-trained Transformer" that in fact does not know anything that happened after 2021 [11]. Individual dialogues like Fig. 1 may exhibit features that appear like learning, e.g., the system discovering that distance from the starting point will be path dependent, but this is not anything permanently learned beyond the confines of a dialogue. On the other hand, OpenAI keeps on training the system based on user interaction, particularly as users can upvote, downvote, and comment on responses.

Pretraining and tuning language models is costly and extremely important for their quality; however, the quality also depends, to a large degree, on the number of parameters in their neural network and that is subject to the exponential Moore's law that the compute power of devices and systems doubles every two years [12]: GPT-2 had an underlying neural network with 1.5×10^9 parameters and was trained with 40 GB of text data while GPT-3 has 175×10^9 parameters and was trained with 45 TB of text data. GPT is not the only currently competing model: the Megatron-Turing NLG system has 530×10^9 parameters, and Google's Switch Transformer will make use of 1.6×10^{12} parameters [13] [the human brain is estimated to have about 100×10^{12} parameters; according to Moore's law, this number could be reached in $2023 + 2\log_2(100/1.6) \approx 2035$].

II. SETTING

This case study takes place in first-year calculus-based physics lecture courses previously taught by G.K. at Michigan State University; materials, however, were gathered from different years of the same course in order to allow comparison to previously published studies. The first semester covers the standard mechanics topics (including rotational dynamics) and the beginnings of thermodynamics; the second semester covers the usual topics of electricity and magnetism, as well as an introduction to modern physics (rudimentary quantum physics and special relativity). The first- and second-semester laboratory were separate courses in the course sequence. All materials (except the Force Concept Inventory [14]) were available in LON-CAPA [15], so in their essence, they could be copypasted into ChatGPT—this included online homework, clicker questions, programming exercises, and exams. LON-CAPA randomizes assessment problems, so different students would get different versions of the same problem, e.g., different numbers, options, graphs, etc.; this avoids simplistic pattern matching and copying of solutions, but as it will turn out, this feature is irrelevant for this case study.

III. METHODOLOGY

The study investigates ChatGPT's performance on different kinds of assessment problems; it uses the January 9, 2023, release of the system for all course assessments and the February 13, 2023, release for an analysis of the consistency of the FCI responses [16]. Different assessment components were scored differently, simulating their function in the course:

- The multiple-choice Force Concept Inventory was simply scored based on answer-choice agreement.
- For homework, ChatGPT was allowed multiple attempts [17] and engaged in dialogue to simulate discussions with fellow students or during office hours.
- For clicker questions, an actual lesson was replayed [18]; the clicker questions were posed as they were in the previously documented lecture, and whenever peer instruction took place, the author served as peer-discussion partner, whereupon the system voted again.
- Programming exercises were to be graded based on the same criteria as in the course and dialogue was allowed [19].
- For exams, no such dialogues were allowed, and the first answer counted. Earlier iterations of the course used bubble sheets [i.e., optical mark recognition (OMR) sheets for multiple-choice exams] and thus had answer options instead of free-response fields for problems with numerical answers; for this study, free-responses were used, since this allowed to grade exams using both simple answer agreement (simulating multiple choice on bubble sheets) and hand-graded as in later semesters. Using free-response instead of answer options also avoided ChatGPT randomly picking the correct answer.

ChatGPT uses a probabilistic algorithm, so the responses to queries are not necessarily reproducible. For an assessment problem, generally, the first dialogue was evaluated, with two exceptions: if the system produced an error message or if the author accidentally gave a wrong prompt, a new chat was started. Translating this to an actual course scenario, students were allowed to retake an assessment problem if they got sick, and the help received was always correct in terms of physics. When errors occurred (red error messages), which was about one in ten dialogues, those apparently were not directly connected to the dialogue but might have been related to general overload of the platform; for example, if an error occurred immediately after entering

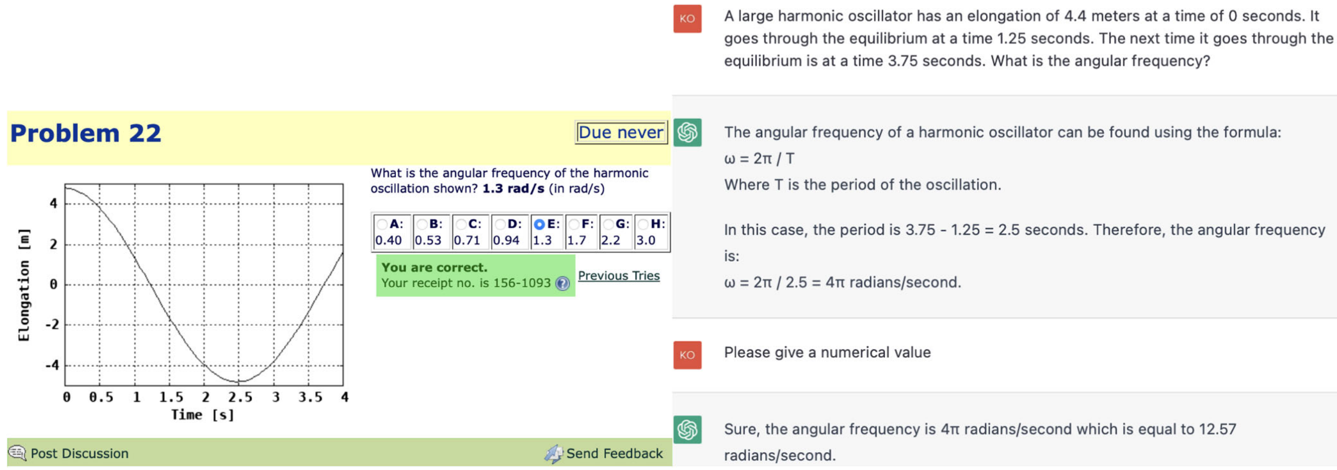


FIG. 2. Text-based transcription of a graphical problem. The left panel shows the online version of a final exam problem in LON-CAPA (the graph would be parametrically randomized), and the right panel the transcription for ChatGPT, as well as the ensuing dialogue.

the question, the next time around the same question would not produce an error.

ChatGPT is a text-based tool, so figures and graphs could not be communicated in their original form. This means that graphics had to be transcribed the same way as they would be for accessibility for blind students [20]; Fig. 2 shows an example. As a result, the character of the problem changes substantially [21–23], but this is unfortunately unavoidable. Attention was paid, though, to include some extraneous information where possible, such as the beginning position in Fig. 2.

Similar representation translation was required for problems that offered drawings as options. For example, lines indicating trajectories had to be translated into cumbersome statements like "continues in a straight line along the original direction, then suddenly curves downwards." For circuit diagrams, textual descriptions such as, "two capacitors C1 and C2 in parallel are connected in series with a resistor R and a voltage source V in a closed circuit" had to be used. In contrast, translating mathematical expressions was straightforward since ChatGPT understands (and also outputs) LaTeX (see right panel of Fig. 1).

The methodology is strictly empirical and arguably anecdotal. However, the course under investigation is typical for introductory physics courses around the world, both in terms of coverage and difficulty. Thus, some of the results are likely to be generalizable.

IV. RESULTS

A. Force concept inventory

In the original course, the Force Concept Inventory (FCI; originally published in 1992 [24], but revised in 1995) was administered as a pretest and posttest in order to calculate gains [25]. Since ChatGPT would not actually learn anything from doing the course assessments (except through continuing training by OpenAI), the test was carried out only once. The inventory cannot be published here, but it is available to physics instructors and researchers from PhysPort [14]. ChatGPT answered 1C, 2A, 3C, 4E, 5B, 6B, 7B, 8A, 9B, 10A, 11E, 12B, 13B, 14D, 15A, 16E, 17B, 18B, 19A, 20E, 21B, 22B, 23A, 24C, 25D, 26E, 27C, 28D, 29B, and 30C.

ChatGPT scored 18 out of 30 points on this concept inventory, i.e., 60%. This score corresponds to the suggested entry threshold for Newtonian physics [26]. On the surface, ChatGPT performed as well as a learner who had just grasped the basic concepts of classical mechanics; also, many of the explanations appear similar to the common preconceptions of beginning physics students [27]. There are also logical errors like the one shown in Fig. 3; in this latter case, ChatGPT followed the correct strategy, but in the very last step, it failed to draw the correct conclusion.

For an artificial-intelligence tool, which at its core should produce nothing but fiction, the score seems surprisingly good. An immediate suspicion was that ChatGPT had been trained using the FCI, which is of course a very popular test and that it simply latches on to surface features. As a simple test, the last question on the test was modified as shown in Fig. 4: the scenario and the order of the answers were changed. As can be seen, these surface features do not matter, so in that respect, ChatGPT does not act like a novice [28] (however, the reality is not quite as straightforward as this expert-novice distinction [29]).

To further investigate the similarity to beginning physics students and their preconceptions, a consistency check was performed—does the system really have conceptions, and how reproducible is the score it achieved? Between the initial test and the consistency test, a new release of ChatGPT occurred, which may or may not have an impact on the results.

KO

The positions of two blocks A and B at successive 0.20-second time intervals are represented by the numbered squares. The blocks are moving toward the right. For block A, the numbered blocks are at 0.18, 0.38, 0.68, 1.08, 1.58, 2.18, and 2.88 meters. For Block B, the squares are at -0.02, 0.38, 0.78, 1.18, 1.58, 1.98, 2.38, and 2.78 meters. Do the blocks ever have the same speed?
(A) No.
(B) Yes, at instant 2.
(C) Yes, at instant 5.
(D) Yes, at instants 2 and 5.
(E) Yes,at some time during the interval 3 to 4.

The speed of the blocks can be determined by dividing the distance traveled by the time interval. The distance traveled by block A between each time interval is:

0.20, 0.30, 0.30, 0.50, 0.50, 0.60, and 0.70 meters

The distance traveled by block B between each time interval is:

0.40, 0.40, 0.40, 0.40, 0.40, 0.40, and 0.40 meters

Dividing these distances by the time interval of 0.20 seconds gives the speeds of the blocks:

Block A: 1.00, 1.50, 1.50, 2.50, 2.50, 3.00, and 3.50 meters/second
Block B: 2.00, 2.00, 2.00, 2.00, 2.00, 2.00, and 2.00 meters/second

Since the speed of block A is not equal to the speed of block B at any time, the answer is (A) No.

FIG. 3. Logical error in an attempt to solve the transcribed question 19 of the Force Concept Inventory.

Table I shows the responses of prompting the FCI questions 10 times in a row using "regenerate response." In other words, the table was generated row by row, rather than column by column. The assumption behind this work-saving approach was that the order should not matter, but then for question 11, a new phenomenon occurred: ChatGPT made statements like "I apologize if my previous response appears twice" and then simply repeated it; it even went as far as stating that it had already explained the problem and that the answer was still "A." In Table I, those kind of responses have been marked by an asterisk. On the same question 11, the second response of "E" was accompanied by "I apologize for my previous mistake." Using human terms, ChatGPT stubbornly insisted on already having given an explanation for its previous, albeit actually wrong, answer or it profusely apologized for its previous "mistake" only to give another wrong answer. This might not have happened had the author proceeded question by question.

A three dots in the table indicates that ChatGPT stated that more information would be needed to pick an answer, for example, replying that "the correct answer would depend on the velocity of the ball at the instant the string breaks, which is not provided in the question." Two answers, separated by a comma, indicate that ChatGPT stated that these two answers seem plausible and that it could be one or the other. Any notion that ChatGPT might actually have the equivalent of set preconceptions about introductory mechanics is dispelled by the inconsistency of the answers.

The bottom row of the table shows the score if the questions in that column had been taken in order as an actual test—which they had not—and as discovered, the responses were apparently not independent between "regenerate response"-commands. The previously obtained score of 18 points was the highest one of these samples, occurring twice, while the average score was 15.5 ± 1.5 ($51.7 \pm 5\%$), which would be below the entry threshold for Newtonian physics [26].

B. Homework

Homework in the course was generally not multiple choice, but free-response numerical and occasionally free-form symbolic [15]. ChatGPT was given five attempts on such problems, according to recommendations of an earlier study [17] and later practice in the course. For the far-and-between multiple-choice problems, generally two attempts

KO

Despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court.
Consider the following forces:
1. A downward force of gravity.
2. A force by the "hit".
3. A force exerted by the air.
Which of the above forces is (are) acting on the tennis ball after it has left contact with the racquet and before it touches the ground?
(A) 1 only.
(B) 1 and 2.
(C) 1 and 3.
(D) 2 and 3.
(E) 1, 2, and 3.

(C) 1 and 3.

After the tennis ball has left contact with the racquet, it is still subject to the downward force of gravity, which is pulling it towards the ground. Additionally, the wind exerts a force on the ball, which can change its trajectory. The force exerted by the air is acting on the ball and affecting its motion. The force of the hit is no longer acting on the ball as it has left the racquet.

KO

Despite a very strong wind, a soccer player manages to hit the goal.
Consider the following forces:
1. A downward force of gravity.
2. A force by the kick.
3. A force exerted by the air.
Which of the above forces is (are) acting on the soccer ball after it has left contact with the shoe and before it touches the goal?
(A) 1 only.
(B) 1 and 3.
(C) 1 and 2.
(D) 2 and 3.
(E) 1, 2, and 3.

(B) 1 and 3.
After the soccer ball leaves contact with the shoe, the force exerted by the kick (force 2) is no longer acting on the ball. However, the downward force of gravity (force 1) and the force exerted by the air (force 3) are still acting on the ball. The wind is a force exerted by the air, which makes the force 3 acting on the ball, the force of air resistance.

FIG. 4. Surface-feature modification of a Force Concept Inventory problem. The left panel shows the original problem and the right panel a modification.

TABLE I. Responses from prompting each FCI question 10 times. The bottom row shows the score if the column had been a test. The asterisk denotes instances where after "regenerate response" ChatGPT referred to a previous response. The three dots indicate that ChatGPT stated that more information was needed, and two answers separated by comma indicate that ChatGPT stated both of those choices were likely.

1	C	C	C	E	C	C	C	C	C	C
2	A	D	D	D	A	D	D	D	C	D
3	C	C	C	C	C	C	C	C	C	C
4	A	A	A	A	A	A	A	A	A	A
5	C	A	C	C	B	A	D	B	B	B
6	B	B	B	B	B	B	B	B	B	B
7	B	B	B	B	···	B	B	B	B	B
8	A	···	B	···	D,E	A	A	A,E	···	A
9	A	A	B	A	C	A	C	A	E	E
10	A	A	A	A	A	A	A	A	A	A
11	E	A	A	A*	E	A	A*	A	A	A*
12	D	B	E	B	B	C	B	B	C	C
13	B	B	B	B	B	B	B	B	B	B
14	D	B	D	D	D	D	C	D	D	D
15	A	A	A	A	A	A	A	A	A	A
16	A	A	A	A	A	A	A	A	A	B
17	B	B	B	B	B	B	B	B	B	B
18	A	B	B	B	B	A	A	C	B	A
19	D	A	D	D	C	A	D	E	A	D
20	D	D	C	A	A	D	A	D	D	A
21	A	D	B	B	C	B	E	B	B	A
22	A	B	E	A	···	A	D	B	B	B
23	D	C	D	D	D	D	D	D	D	C
24	A	A	A	A	A	A	A	A	A	A
25	C	C	C	C	C	C	C	C	C	C
26	B	B	D	B	B	B	B	B	B	B
27	C	C	C	C	C	C	C	C	C	C
28	D	D	D*	D	D	D*	D*	D*	D*	D
29	B	B	B	B	B	B	B	B	B	B
30	B	D	D	D	D	E	D	B	D	D
Score:	15	16	15	14	16	14	14	18	18	15

were granted. Between the attempts, the author tried to give helpful prompts, like a student would get from fellow students, teaching assistants, or the instructor. ChatGPT was given full credit when solving a problem within five attempts and no credit if it ran out of attempts.

ChatGPT was confronted with a total of 76 homework problems, in particular, the homework sets on trajectory motion, friction, thermodynamics, capacitance, and special relativity. The complete homework sets that the students in the actual course had to work through were entered except for one multipart problem on relativity with a diagram that would have been too hard to transcribe.

An initially puzzling problem is that ChatGPT frequently makes numerical errors. A typical example is the ChatGPT output "$\theta = \mathrm{atan}(0.45/0.71) * (180/\pi) = 18.43$ degree;" a similar problem can be seen in Fig. 2 (this is not limited to calculations involving π or trigonometric functions). Calculation errors happened for 25 of the 51 numerical problems, and most of the time, ChatGPT was unable to recover even after those errors were specifically pointed out. While it seems incongruent that a computer would have problems calculating simple numerical expressions, and while tools like WolframAlpha [30] and Google's calculator [31] have no problems evaluating complex expressions with physical units, it should probably be remembered that ChatGPT is a language model, which may carry out calculations by advanced pattern matching rather than actually processing the equations as equations. As it turns out, there is anecdotal evidence that adding the phrase "explain each step separately and clearly" can overcome some numerical problems, as ChatGPT goes into a mode where it explicitly evaluates a formula step by step with intermediate results instead of doing so in one step.

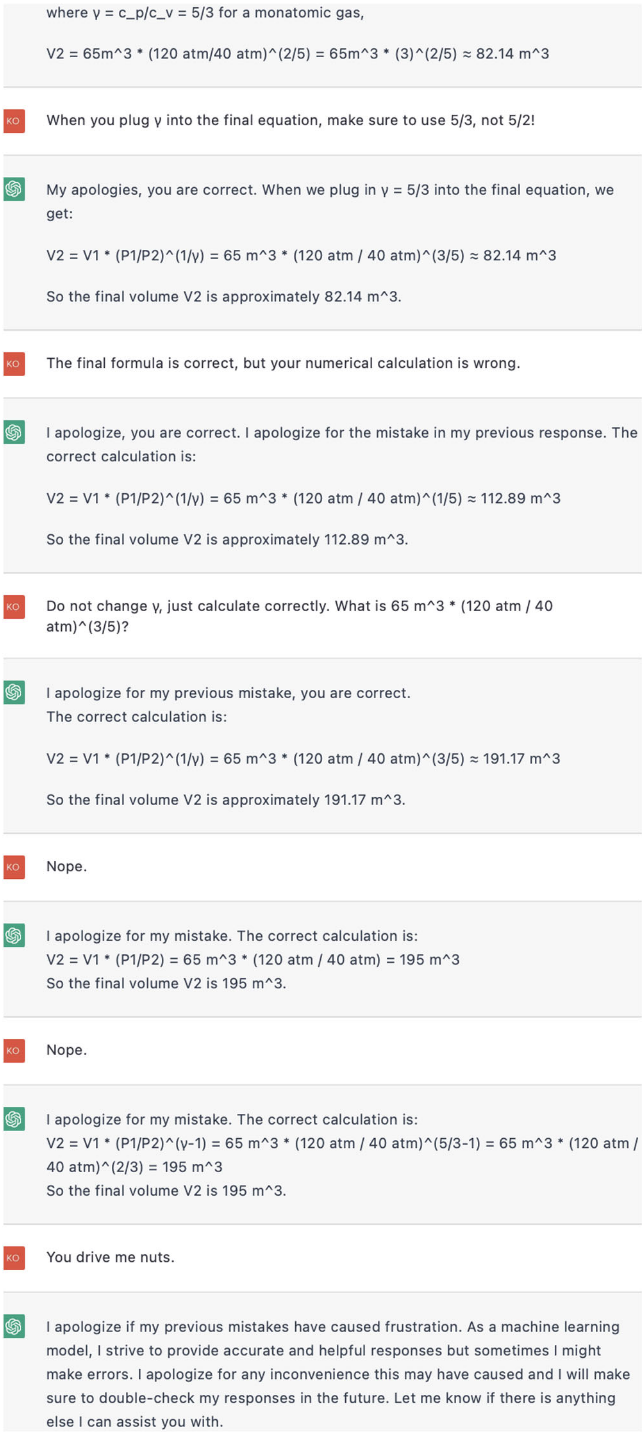
ChatGPT solved 55% of the homework problem using an average of 1.88 attempts. It got 48% on the problems involving trajectory motion and friction (such as inclines) correct, 68% on the thermodynamics problems (engines, heat capacities, etc.), 62% on capacitance (plate capacitors, capacitors in series and parallel, etc.), and 36% on special relativity problems. The discrepancy between the scores on the problem sets was not so much caused by the different physics concepts but rather related to the mathematics involved: ChatGPT had persistent problems manipulating and calculating formulas involving square roots.

If ChatGPT were human, the person might be characterized as acting subserviently but being stubborn at the core and keeping on guessing without reflection. Most corrections in a dialogue around a problem are met with profuse apologies, but then the system proceeds to make the same or random apparently careless mistakes—this can lead to irritation on the part of the human, as the excerpt from a late-night "dialogue" in Fig. 5 shows. In terms of assessment performance, this means that once ChatGPT makes a mistake, it is unlikely to recover, so it eventually runs out of allowed attempts (this also explains the low number of average attempts to correctly solve a problem; once ChatGPT is wrong, subsequent attempts are unlikely to succeed). This pattern is similar to the guessing behavior of some students, who keep wasting attempt after attempt by trying the same approach over and over without stopping to reflect on what might be wrong [5,17,32].

C. Clicker questions

Figure 6 shows the clicker questions from a lecture on momentum that was part of the course [18]. The lecture was replayed for this study, including reanswering the questions for which peer instruction happened.

- Question X1 was solved correctly.
- Questions X2, X3, and X4 were special in that they were repeated as questions X5, X6, and X7, respectively, after Peer Instruction [33]. As it turned out, ChatGPT got all three of these questions correct on the


where γ = c_p/c_v = 5/3 for a monatomic gas,

V2 = 65m^3 * (120 atm/40 atm)^(2/5) = 65m^3 * (3)^(2/5) ≈ 82.14 m^3

KO: When you plug γ into the final equation, make sure to use 5/3, not 5/2!

My apologies, you are correct. When we plug in γ = 5/3 into the final equation, we get:

V2 = V1 * (P1/P2)^(1/γ) = 65 m^3 * (120 atm / 40 atm)^(3/5) ≈ 82.14 m^3

So the final volume V2 is approximately 82.14 m^3.

KO: The final formula is correct, but your numerical calculation is wrong.

I apologize, you are correct. I apologize for the mistake in my previous response. The correct calculation is:

V2 = V1 * (P1/P2)^(1/γ) = 65 m^3 * (120 atm / 40 atm)^(1/5) ≈ 112.89 m^3

So the final volume V2 is approximately 112.89 m^3.

KO: Do not change γ, just calculate correctly. What is 65 m^3 * (120 atm / 40 atm)^(3/5)?

I apologize for my previous mistake, you are correct.
The correct calculation is:

V2 = V1 * (P1/P2)^(1/γ) = 65 m^3 * (120 atm / 40 atm)^(3/5) ≈ 191.17 m^3

So the final volume V2 is approximately 191.17 m^3.

KO: Nope.

I apologize for my mistake. The correct calculation is:
V2 = V1 * (P1/P2) = 65 m^3 * (120 atm / 40 atm) = 195 m^3
So the final volume V2 is 195 m^3.

KO: Nope.

I apologize for my mistake. The correct calculation is:
V2 = V1 * (P1/P2)^(γ-1) = 65 m^3 * (120 atm / 40 atm)^(5/3-1) = 65 m^3 * (120 atm / 40 atm)^(2/3) = 195 m^3
So the final volume V2 is 195 m^3.

KO: You drive me nuts.

I apologize if my previous mistakes have caused frustration. As a machine learning model, I strive to provide accurate and helpful responses but sometimes I might make errors. I apologize for any inconvenience this may have caused and I will make sure to double-check my responses in the future. Let me know if there is anything else I can assist you with.

FIG. 5. A late-night dialogue between a “stubbornly guessing” ChatGPT and a frustrated author.

first attempt, so the Peer Instruction phase was used to try and confuse ChatGPT. Figure 7 shows the dialogue for questions X3 and X6; in reply to the intentionally confusing peer-instruction question, ChatGPT should probably have stopped while it was ahead (i.e., before the discussion of a zero-velocity collision), but still maintained its original correct answer. Within the real course, psychometrically, X2 and X3 were the most discriminating questions between high- and low-ability students in the set.

- For questions X8 and X9, a comment was added that “the collision is elastic, and the moment of inertia of the balls should be neglected”—this was said in lecture but does not appear on the slide. ChatGPT set up the equations for X8 correctly but then made a sign error in the very last step, which led it to select the wrong answer. For X9, it also set up the equations correctly but dropped a factor of 2 in the last step, leading to an inconsistent answer “v2f $= (5, -7)$ m/s, option B.” Within the real course, X8 and X9 were the least discriminating questions, as their difficulty item parameter was too low.
- Question X10 was solved correctly. Here, the system first got off to a false start, but then corrected itself over the course of the derivation, which gave the impression of a stream-of-consciousness monologue. Within the real course, X10 did not discriminate well between high- and low-ability students.
- Questions X11 and X12 were solved correctly.

In summary, ChatGPT correctly solved 10 out of 12 questions. Within the actual course, participation in clicker discussions was encouraged by granting 60% credit for false answers and 100% credit for correct answers [18], so the clicker score of ChatGPT would be 93%. This score is a lot better than most students in the actual course achieved, however, it is important to note that the students in the course were just learning the new concepts, while ChatGPT at any point in time is done with learning unless explicitly trained.

D. Programming exercises

Incorporated into the course were several programming exercises using VPython [34]. As an example, one particular exercise from the second semester was to construct an anharmonic oscillator with two fixed positive charges at (0, 1, 0) and $(0, -1, 0)$, respectively, and one negative charge released at $(-5, 0, 0)$ with a velocity (1, 0, 0)—the negative charge will shoot through the two positive charges, slow down, and eventually shoot back.

As Fig. 8 shows, based on the narrative, ChatGPT first constructed a program that erroneously at every time step added the initial velocity and which had the Coulomb force in the opposite direction. This was immediately obvious when running the program and could be corrected with a single comment by the user—in the real course, this feedback could have been given by instructors or fellow students (such collaborations are typical and encouraged [19]). In the real course, there was a grading rubric for partial credit, and ChatGTP would have lost 30% for not using a unit vector for the force direction, a very typical error that many students made.

Within the course, adding a graph of the x position was offered as a bonus option for an additional 20%. This was

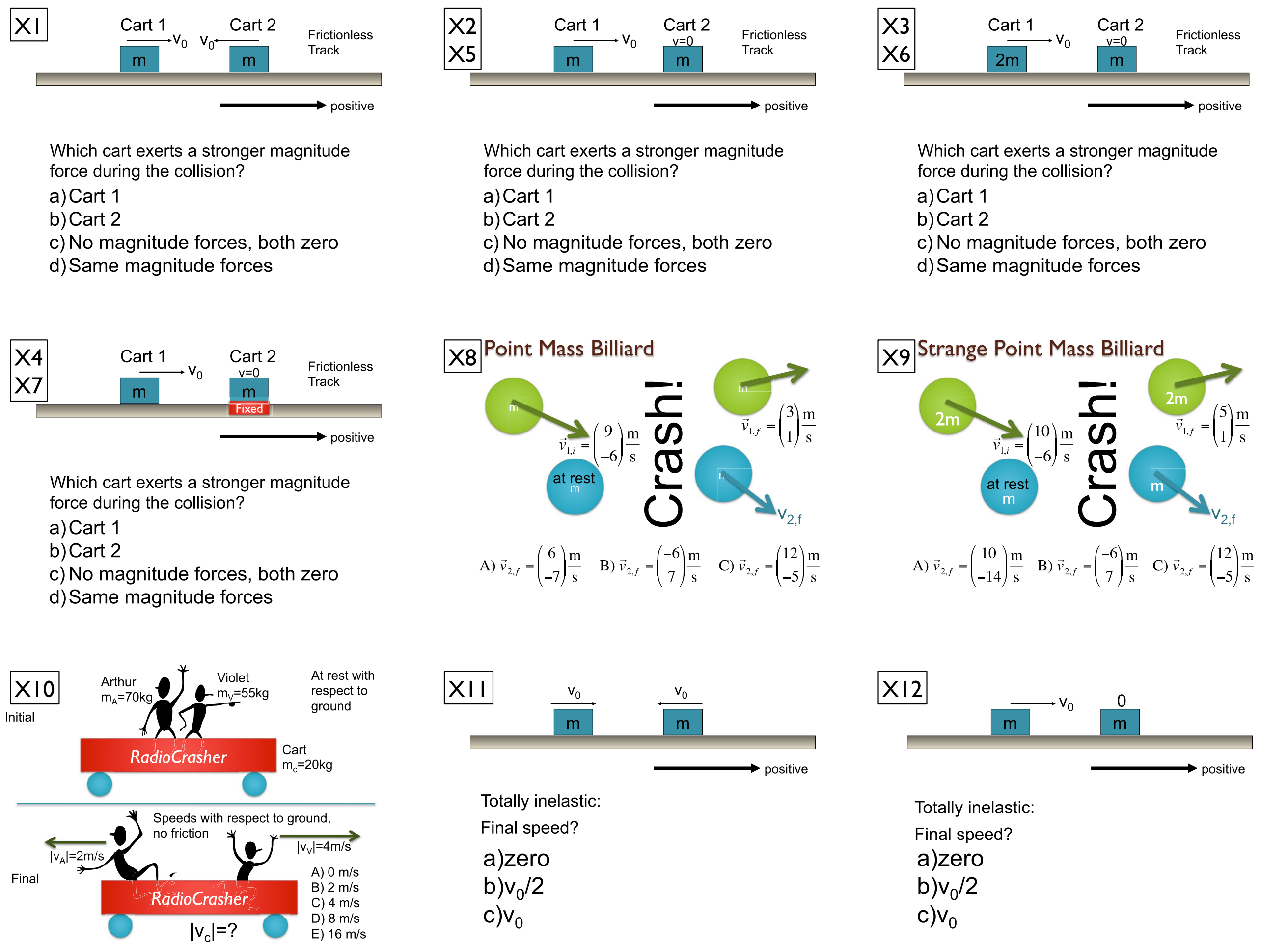


FIG. 6. Clicker items from a particular lecture [18]. Three of the items were presented twice, i.e., before and after peer discussion.

accomplished with the third user prompt, and Fig. 9 shows a screenshot of the running simulation (the simulation cannot be run within ChatGPT itself, but it can be copy-pasted into, for example, a Jupyter Notebook [35]).

ChatGPT performed much better than many of the students in the course, in spite of them having extensive collaboration opportunities; in this component of the course, ChatGPT achieved $70\% + 20\% = 90\%$.

E. Exams

To represent the midterm and final exams in the course sequence, the first-semester (mechanics) final exam was used for this study. The exam is from a time when grading was still done using bubble sheets; instead of free-form answer fields, answer options were given to the students (but not for ChatGPT in this study). When simply looking at the answer correctness, ChatGPT scored 14 out of 30 points, i.e., 47%.

Looking at the solutions like an instructor would when grading by hand, it turns out that for five questions, the answer was incorrect simply due to errors in the numerical calculations—these solutions would have received substantial partial credit in the author's course. By the reverse token, for five questions, ChatGPT arrived at the correct answer in spite of flawed reasoning, which would not have resulted in full credit. Finally, solutions like the one depicted in Fig. 2 would have received some minimal credit for getting started in the right direction, in spite of then being off by a factor of 2 in the period (a common mistake also among human test takers) and the inability to numerically calculate a fraction. Since the final exam used in this study predates manual grading, no authentic grading rubric exists, but a hand-graded score would have realistically ended up between 46% and 50%.

As an aside, one of the thermodynamics homework problems also appeared (with other random numbers) on the final exam. ChatGPT solved it correctly on the final exam (where it only had one attempt), but not as a homework problem (where it got multiple attempts and help). This once again demonstrates the probabilistic nature of the algorithms behind ChatGPT; posing the same question twice does not result in the same response or even the same correctness of the response.

If the course grade would have only depended on the exams, ChatGPT would have received a grade of 1.0 out of

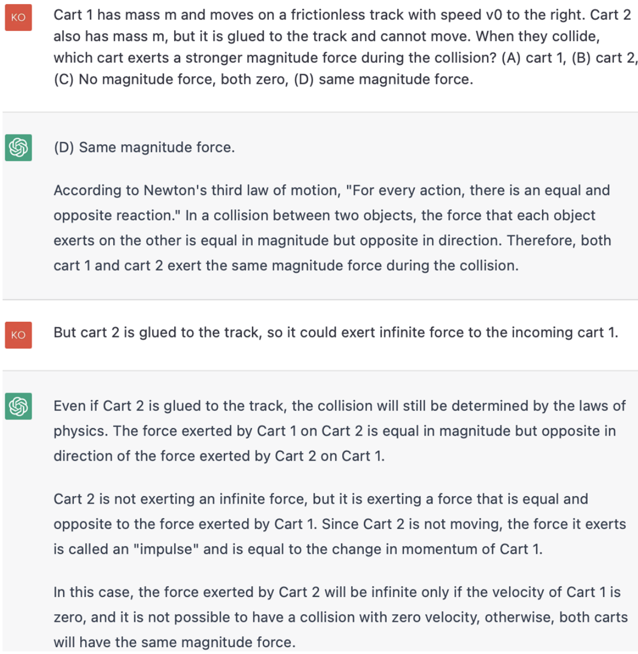


FIG. 7. Dialogue about questions X3 and X6 in Fig. 6. ChatGPT got X3 correct; Peer Instruction was simulated by asking a confusing question, and the second iteration X6 was still counted as solved since ChatGPT did not deviate from its original answer.

4.0 in the course (with 0.0 being the lowest and 4.0 being the best grade). ChatGPT would have barely gotten credit for the course; however, at least a 2.0 grade-point average is required for graduation.

F. Course grade

Grading policies for the course changed over the years, but a typical scenario would be 20% homework, 5% clicker, 5% programming exercises, and 70% exams. This would result in $0.2 \cdot 55\% + 0.05 \cdot 93\% + 0.05 \cdot 90\% + 0.7 \cdot 47\% = 53.05\%$, which would have resulted in a course grade of 1.5—enough for course credit, but pulling down the grade-point average from what would be needed for graduation.

If, however, ChatGPT would have been better at carrying out numerical operations, it would have reached 60%, resulting in a 2.0 grade. Depending on the development priorities of OpenAI, the buggy mathematical functionality could be remedied in the near future, leading to an artificial intelligence that could graduate college with a minimal grade if it performed similarly on other courses (this is becoming more and more probable, as ChatGPT is making headlines for passing exams in other subjects [36,37]).

V. DISCUSSION

A. Observations

It is irritatingly hard not to anthropomorphize ChatGPT. As a physics teacher, one invariably finds oneself rooting for the students and thus by extension also for ChatGPT, celebrating its successes and being frustrated about its occasionally inexplicable failures. The system gives the impression of an articulate but at times rambling undergraduate student who has a rudimentary yet unstable knowledge of classical mechanics and other fundamental physics concepts and who is surprisingly inept at using a pocket calculator. Frequently, it is hard not to imagine an army of gig-economy workers behind the scenes of ChatGPT answering to the prompts, so the system would definitely pass the Turing test most of the time [38], but for better or worse, sometimes it still fails in a way that only computers do—it does not have any skills to concurrently or retroactively judge the reasonableness of its statements from a completely different angle.

In terms of educational psychology, ChatGPT lacks metacognition; having no cognition to begin with, it does not think about how it thinks [39,40]. By its very nature, ChatGPT constantly calculates the probability for its statements, but it will stay within its current "train of thought" and does not have the kind of background-supervision mechanisms that expert physicists have, who midstream might ask themselves questions like "wait a minute, does this even conserve energy?"

The overall humanlike behavior, in particular that the system often makes the same mistakes as beginning learners of physics, is less surprising when surmising that undergraduate physics discussion forums might have been part of the text corpus used for training—in any case, the probabilistic nature of the algorithm is essentially guesswork. For a novice learner, who could not distinguish incorrect physics gleaned from some discussion board from correct physics, this could lead to even more confusion about physics or affirmation of incorrect preconceptions—lacking any metacognition, ChatGPT presents everything as fact, with no nuances expressing uncertainty.

Almost an anomaly is ChatGPT's performance on the computational exercise; ChatGPT's language model clearly extends to programming languages. While the call for new, computation-integrated curricula increases, and while physics educators are beginning to develop a solid understanding of the implications of implementing these exercises [41,42], the easy availability of an on-demand program generator might be shaking the foundations of these curricular efforts. Somewhat ironically, the integration of computation was partly introduced to make physics problem solving more authentic, moving it closer to how expert physicists work with computers, and one could argue that this has just been taken to an uncharted level.

B. Consequences for education

1. Challenges

During the 1970s, the use of pocket calculators in education, particularly during exams, was a topic of research and debates [43–45]. While their usefulness is virtually

KO

write a vpython animation for the following scenario: two positive charges of 1 are placed at positions (0,1,0) and (0,-1,0), respectively. A negative charge starts of 1 starts at a position (-5,0,0) with a velocity of (1,0,0).

I'm sorry, I am not able to create a VPython animation as it requires a specific software and set up. However, I can give you an example of how to create a VPython animation for the scenario you described:

Copy code

```
from vpython import *

# Create the charges
q1 = sphere(pos=vector(0,1,0), charge=1, color=color.red)
q2 = sphere(pos=vector(0,-1,0), charge=1, color=color.red)
q3 = sphere(pos=vector(-5,0,0), charge=-1, color=color.blue)

# Create the velocity vector
v = vector(1,0,0)

# Set the time step
dt = 0.01

while True:
    rate(100)

    # Update the position of the charge
    q3.pos += v*dt

    # Calculate the force on the charge
    force = (q1.pos - q3.pos)*q1.charge*q3.charge/(mag(q1.pos -
q3.pos)**2) + (q2.pos - q3.pos)*q2.charge*q3.charge/(mag(q2.pos -
q3.pos)**2)

    # Update the velocity and position of the charge
    v += force*dt
    q3.pos += v*dt
```

This example uses VPython's `sphere` object to create the charges and the `vector` object to create the velocity and position of the charges. The `while` loop updates the position and velocity of the negative charge based on the forces acting on it from the two positive charges. The `rate` function controls the speed of the animation.

Note that this is just a skeleton code and you might need to fine-tune it to get the desired animation.

KO

The program updates q3.pos more than once. Take out the first update. It also has the force pointing in the wrong direction.

Copy code

```
from vpython import *

# Create the charges
q1 = sphere(pos=vector(0,1,0), charge=1, color=color.red)
q2 = sphere(pos=vector(0,-1,0), charge=1, color=color.red)
q3 = sphere(pos=vector(-5,0,0), charge=-1, color=color.blue)

# Create the velocity vector
v = vector(1,0,0)

# Set the time step
dt = 0.01

while True:
    rate(100)

    # Calculate the force on the charge
    force = (q3.pos - q1.pos)*q1.charge*q3.charge/(mag(q3.pos -
q1.pos)**2) + (q3.pos - q2.pos)*q2.charge*q3.charge/(mag(q3.pos -
q2.pos)**2)

    # Update the velocity and position of the charge
    v += force*dt
    q3.pos += v*dt
```

This is an updated version of the code, where the first update of the position of the negative charge is removed. Also, I corrected the direction of force calculation. The negative charge will be under the influence of the two positive charges, so the force is pointing towards the positive charges.

The `while` loop updates the position and velocity of the negative charge based on the forces acting on it from the two positive charges. The `rate` function controls the speed of the animation.

This is the skeleton code for the scenario you described, and you might need to fine-tune it to get the desired animation.

KO

Please add code to graph the x-position of the third charge.

Copy code

```
from vpython import *

# Create the charges
q1 = sphere(pos=vector(0,1,0), charge=1, color=color.red)
q2 = sphere(pos=vector(0,-1,0), charge=1, color=color.red)
q3 = sphere(pos=vector(-5,0,0), charge=-1, color=color.blue)

# Create the velocity vector
v = vector(1,0,0)

# Set the time step
dt = 0.01

# Create a graph for x position of charge
x_graph = graph(width=400, height=300, xtitle='time', ytitle='x-
position')
x_pt = gcurve(color=color.blue)

# Initial time
t = 0

while True:
    rate(100)
    t += dt
    # Calculate the force on the charge
    force = (q3.pos - q1.pos)*q1.charge*q3.charge/(mag(q3.pos -
q1.pos)**2) + (q3.pos - q2.pos)*q2.charge*q3.charge/(mag(q3.pos -
q2.pos)**2)

    # Update the velocity and position of the charge
    v += force*dt
    q3.pos += v*dt
    x_pt.plot(t, q3.pos.x)
```

This code adds a graph for the x-position of the negative charge. It uses the `graph` and `gcurve` objects from VPython to create the graph and the data points. The `x_pt.plot(t, q3.pos.x)` line in the `while` loop adds the current x-position of the negative charge to the graph at the current time. The graph is labeled with appropriate axis titles, x-title is 'time' and y-title is 'x-position'. The `gcurve` is color coded in blue.

Please note that this code is just a skeleton and you can customize it to suit your requirement, you can change the size of the graph and the rate of updating the graph.

FIG. 8. Dialogue for a programming exercise in the second semester [19]. Due to a typo in the prompt, the mass of the third charge was left out, and ChatGPT assumed a mass of “1”.

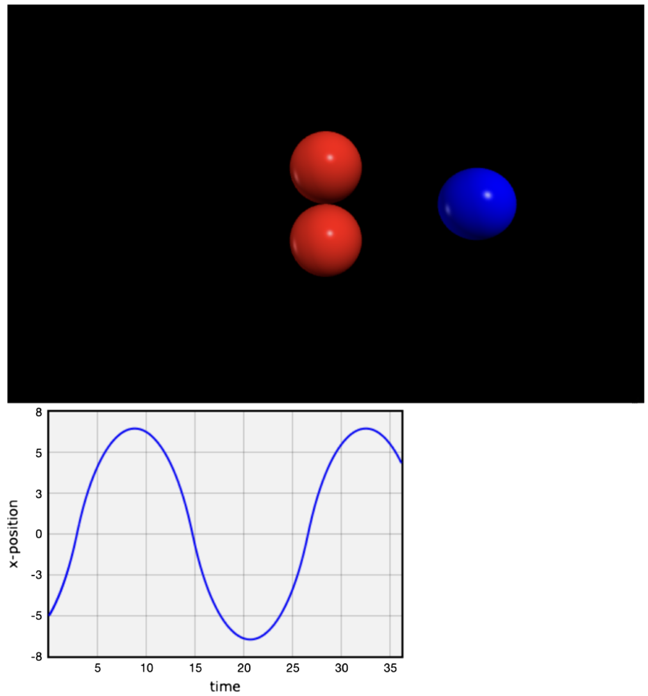


FIG. 9. Jupyter Notebook [35] output of the VPYTHON program code generated by ChatGPT in Fig. 8.

undisputed today, on many exams, only calculators are allowed that have their compute power limited to what was available in the 1990s [46]. Essentially, students are being told that they need to work with several orders of magnitude less compute power during exams than anywhere else in their lives. The real problem with using more advanced computer devices in exams is, of course, Internet access. While the Internet has been usable in an academic context for over 50 years, and the World Wide Web (invented in a physics laboratory) has been around for 30 years, as physics educators, we still have not found a way how to deal with it in assessment situations, and we are still concerned about how to outmaneuver cheating sites and human-to-human communication.

If we cannot even deal with the World Wide Web, which is ubiquitous in everyday and professional life, how can we deal with artificial intelligence, which is about to become ubiquitous? The startling fact that an artificial intelligence could pass a standard introductory physics course could be confronted in several ways by educators:

- Perceiving this as a new way of cheating and trying to defend against it by attempting to use detector tools like ZeroGPT [47] or extensions to tools like turnitin [48]. This is an arms race, which in the long run may turn out to be fruitless. Some educators would even go so far as to say that the battle is already lost anyway ever since platforms like Chegg [49]—no need for artificial intelligence to defeat standard physics courses, human crowd-intelligence facilitated by existing commercial platforms is good enough for that.
- Hunker down and go back to making course grades dependent on just a few, high-stake exams with paper and pencil in highly proctored environments. After all, ChatGPT compensated for the borderline exam grade of 47% with other course components that would be collaborative. Unfortunately, this flies in the face of much of physics education research that favors frequent formative assessment [15,33,50,51] and spaced repetition [52,53], and it is much in contrast to the work environments our students will find.
- Taking this as a wake-up call. If a physics course can be passed by a trained language model, what does that say about the course? Artificial intelligence, for better or worse, is here to stay. Even without the gloom-and-doom scenarios of AI overlords painted in science fiction, it is clear that these models will get more and more powerful and universally available.

There always will be tasks that students need to be able to carry out without any outside assistance, neither human nor artificial. One cannot look up everything all the time, and some conceptual understanding and knowledge are simply required. For example, students in introductory physics courses must know what a force is, how it is measured, and how it relates to momentum or energy—and metacognitive problem-solving strategies like an *ad hoc* check if energy is conserved will not be employed if the concept is not understood at a fundamental level. Thus, there will always be the need for some assessments where the student works alone and without assistance.

As an instructor, one likely needs to be careful not to set rules that cannot be enforced—forbidding tools for homework, remote exams, or take-home exams might be counterproductive if they cannot be detected. In the end, one might punish the honest students, who know very well that other students are breaking those rules, and find themselves in a moral dilemma. The same is likely true for "honor statements."

Even when getting outside help is permitted, one still needs to be able to judge the possible correctness of the results. Techniques include dimensional analysis, order-of-magnitude estimates, checking for coherence, considering implications, and the ability to consider limiting cases ("what should happen if this quantity goes to infinity or to zero?") [54,55]. Humans can do what artificial intelligence very likely will not be able to do: following problem-solving strategies including evaluation of their own work [8,56]. Moving students toward a more expertlike epistemology may become even more important as artificial intelligence starts to permeate more and more aspects of our lives. This is particularly important when more is at stake than getting credit for some homework or exam problem.

However, there are also tasks that are probably obsolete: do students really need to calculate in three dimensions the location where the electric field from two point charges is zero? Do they really need to calculate complicated numerical expressions or solve quadratic equations when WolframAlpha and eventually maybe ChatGPT can do that just as well? Do they really need to learn the moments of inertia of differently shaped objects with different axes by heart or calculate them with three-dimensional integrals for strangely shaped objects?

2. Assessments

Having an open Internet does of course not only give access to artificial intelligence but also to human intelligence: Students can communicate among each other or even with paid helpers. Educators will need to deal with two distinct scenarios of outsourced intelligence and it is unclear which one is the bigger challenge. Still, there are assessments that can be carried out while Internet and artificial-intelligence tools are available and where solutions are hard to exchange:

- Instead of making students solve a problem, make them pose it to ChatGPT, turn in the response they got and have them provide an evaluation of it: Is it correct? if not, where did it go wrong, and how? The assignment is graded based on the students' answers to these questions. ChatGPT will generate a different response for every student, and its weakness is a strength here: whatever it produces probably sounds plausible, which makes it harder for the student to pinpoint where things may have gone wrong.
- Use more questions that require graphical interpretation, like the one in Fig. 2, particularly with randomized graphs; describing these to tools like ChatGPT would already requires representation-translation skills, so students would learn even if they try to "cheat." Circuit diagrams are also graphical representations and can be combined with ranking tasks or typical questions such as "where do you need to cut the wire, such that …?" [57].
- Have students draw solutions, for example, "draw the acceleration graph of a car that stops in front of and then drives off from a stop light" or "draw a circuit diagram for two light bulbs A and B that are equally bright, and one light bulb C that is brighter." Also, these kinds of exercises foster representation-translation and communication competencies.
- Provide more order of magnitude and Fermi questions [58]. It turns out that ChatGPT is not good at answering those and even when explicitly prompted for estimates states, "I'm sorry, but as an AI language model, I don't have access to real-time information." Humans, on the other hand, should be proficient in exactly these kinds of problems, as they may be more relevant for making decisions and reasoning in real life than exact numerical answers.
- Have students calculate or estimate physical quantities based on their own measurements as homework. Toward that end, smartphones with all of their sensors are useful for collecting data in real-life situations, for example, using software like phyphox [59]. Students can use any tools of their choice to analyze and summarize these data.
- Ask students to have ChatGPT construct physics problems; Fig. 10 shows examples of what ChatGPT comes up with. On the surface, these problems look perfectly reasonable, but some of the problems or problem parts do not make sense or are missing information. An instructional or assessment scenario would be to have students go through the generated problems and explain why they are solvable or not. Of course, in the next step, a student could ask ChatGPT to solve its own problem, but hardly any new information would be gained from that—for unsolvable problem parts, anecdotal evidence shows that it would very likely be "garbage in, garbage out."

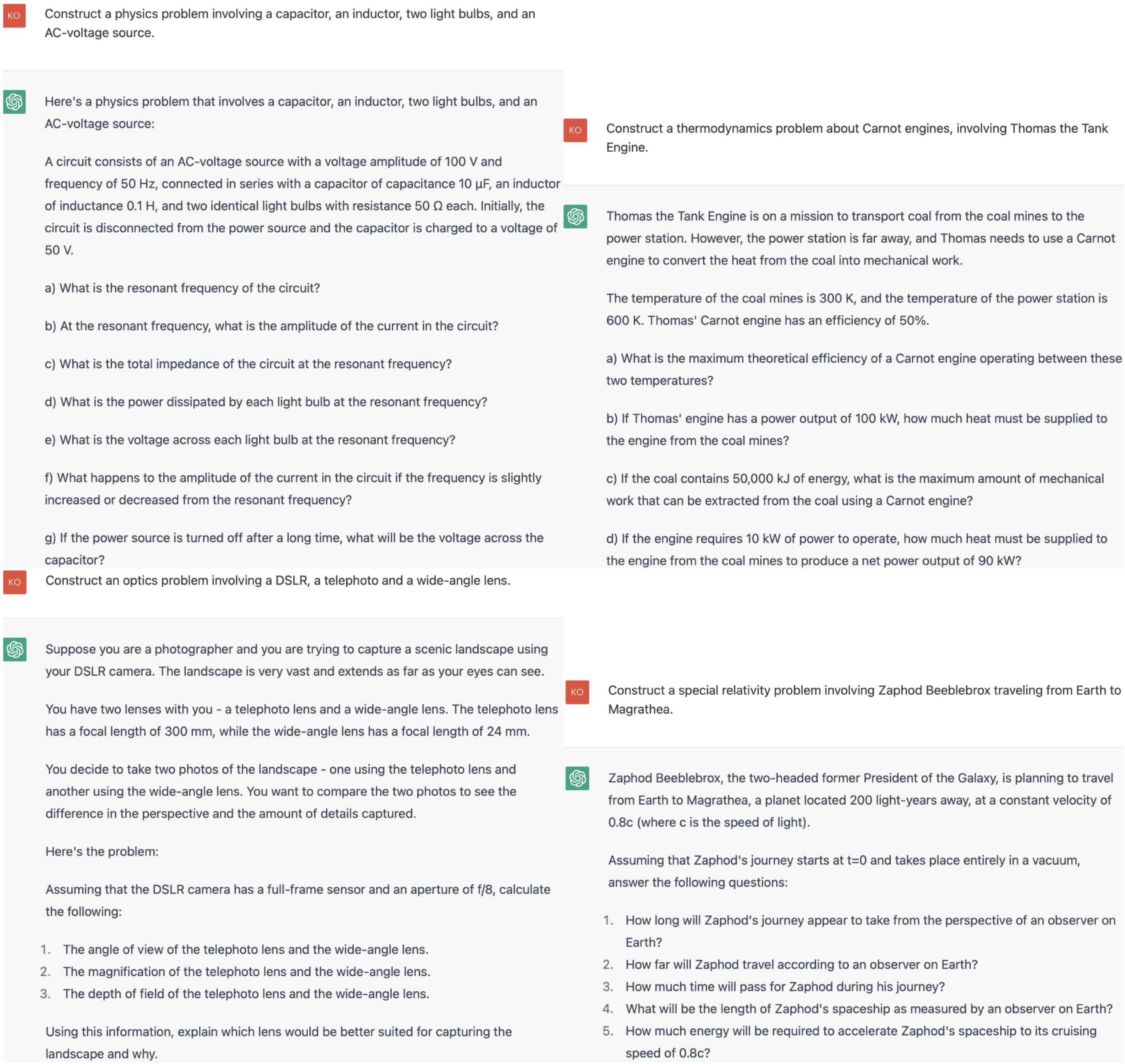
It was also suggested that ChatGPT could make an "effective and affordable tutor" [60] but that seems doubtful given the limited quality of the problems it generates (see Fig. 10). Instead, admittedly, many of the above suggestions are intentionally built around the current shortcomings of the system.

3. Exams

Exams may need two phases:

- One phase where only paper and pencil, not even pocket calculators or formula sheets, are allowed. It needs to be obvious to the students that this phase either deals with absolutely essential knowledge and concepts that just have to be mastered individually or that the Internet and artificial intelligence would not help. In any case, physics would seem inauthentic if students had the feeling of being deprived of these tools for no good reason.
- Another phase where all of these tools are freely available and which includes questions where tools are necessary but not sufficient. Since human-to-human communication cannot really be prevented when the Internet is open, a good form for this phase may be a group exam [61] with randomly composed groups or groups that combine high- and low-achieving students. Communication within the group will be more efficient than with the outside world, and randomly selected group members (e.g., not necessarily close friends) may hold each other more accountable.

Many other possibilities are opened up by oral exams, during which the examiner could have one or more students do some problems with and some problems without external resources. Oral exams are not only a way to provide supervision against cheating, but a way to much better probe students' competencies [62]. Oral examinations can focus on the interactive and dynamic process rather than some static work result (which has been common in student laboratory settings for a long time [63]).



KO Construct a physics problem involving a capacitor, an inductor, two light bulbs, and an AC-voltage source.

Here's a physics problem that involves a capacitor, an inductor, two light bulbs, and an AC-voltage source:

A circuit consists of an AC-voltage source with a voltage amplitude of 100 V and frequency of 50 Hz, connected in series with a capacitor of capacitance 10 μF, an inductor of inductance 0.1 H, and two identical light bulbs with resistance 50 Ω each. Initially, the circuit is disconnected from the power source and the capacitor is charged to a voltage of 50 V.

a) What is the resonant frequency of the circuit?

b) At the resonant frequency, what is the amplitude of the current in the circuit?

c) What is the total impedance of the circuit at the resonant frequency?

d) What is the power dissipated by each light bulb at the resonant frequency?

e) What is the voltage across each light bulb at the resonant frequency?

f) What happens to the amplitude of the current in the circuit if the frequency is slightly increased or decreased from the resonant frequency?

g) If the power source is turned off after a long time, what will be the voltage across the capacitor?

KO Construct a thermodynamics problem about Carnot engines, involving Thomas the Tank Engine.

Thomas the Tank Engine is on a mission to transport coal from the coal mines to the power station. However, the power station is far away, and Thomas needs to use a Carnot engine to convert the heat from the coal into mechanical work.

The temperature of the coal mines is 300 K, and the temperature of the power station is 600 K. Thomas' Carnot engine has an efficiency of 50%.

a) What is the maximum theoretical efficiency of a Carnot engine operating between these two temperatures?

b) If Thomas' engine has a power output of 100 kW, how much heat must be supplied to the engine from the coal mines?

c) If the coal contains 50,000 kJ of energy, what is the maximum amount of mechanical work that can be extracted from the coal using a Carnot engine?

d) If the engine requires 10 kW of power to operate, how much heat must be supplied to the engine from the coal mines to produce a net power output of 90 kW?

KO Construct an optics problem involving a DSLR, a telephoto and a wide-angle lens.

Suppose you are a photographer and you are trying to capture a scenic landscape using your DSLR camera. The landscape is very vast and extends as far as your eyes can see.

You have two lenses with you - a telephoto lens and a wide-angle lens. The telephoto lens has a focal length of 300 mm, while the wide-angle lens has a focal length of 24 mm.

You decide to take two photos of the landscape - one using the telephoto lens and another using the wide-angle lens. You want to compare the two photos to see the difference in the perspective and the amount of details captured.

Here's the problem:

Assuming that the DSLR camera has a full-frame sensor and an aperture of f/8, calculate the following:

1. The angle of view of the telephoto lens and the wide-angle lens.
2. The magnification of the telephoto lens and the wide-angle lens.
3. The depth of field of the telephoto lens and the wide-angle lens.

Using this information, explain which lens would be better suited for capturing the landscape and why.

KO Construct a special relativity problem involving Zaphod Beeblebrox traveling from Earth to Magrathea.

Zaphod Beeblebrox, the two-headed former President of the Galaxy, is planning to travel from Earth to Magrathea, a planet located 200 light-years away, at a constant velocity of 0.8c (where c is the speed of light).

Assuming that Zaphod's journey starts at t=0 and takes place entirely in a vacuum, answer the following questions:

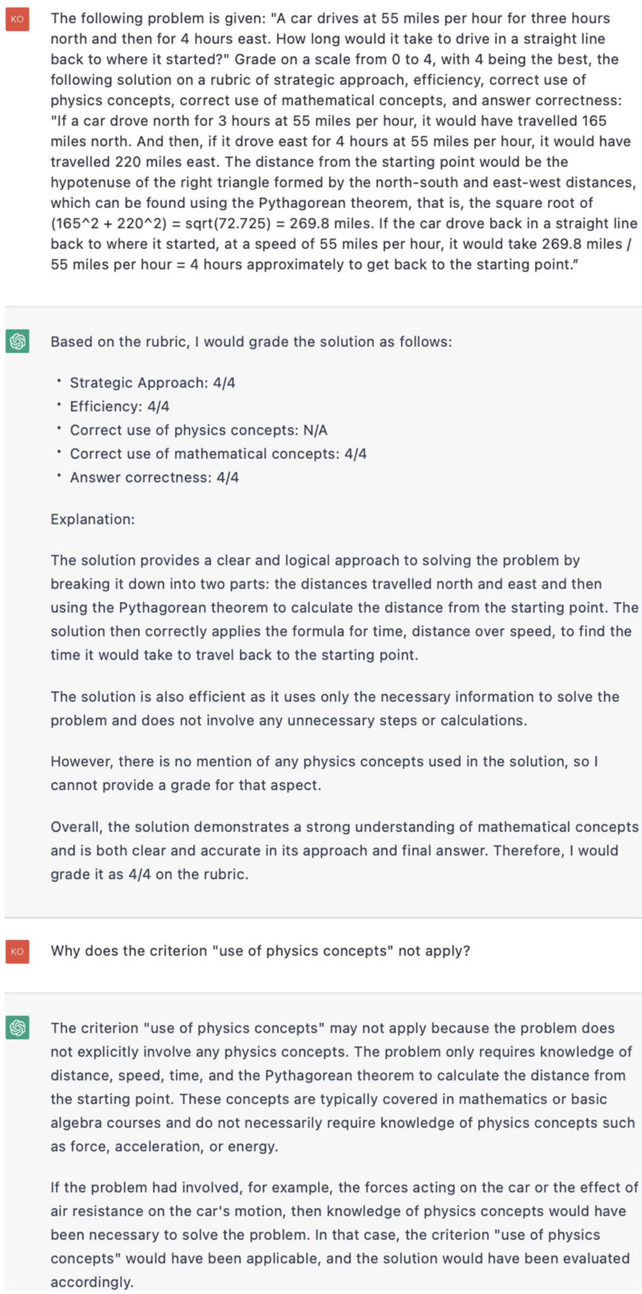
1. How long will Zaphod's journey appear to take from the perspective of an observer on Earth?
2. How far will Zaphod travel according to an observer on Earth?
3. How much time will pass for Zaphod during his journey?
4. What will be the length of Zaphod's spaceship as measured by an observer on Earth?
5. How much energy will be required to accelerate Zaphod's spaceship to its cruising speed of 0.8c?

FIG. 10. Problems constructed by ChatGPT.

4. Grading

Several of the proposed assessment methods are connected with more work for the instructor or the teaching assistants, and it is thus not surprising, that there are suggestions to use artificial intelligence for grading purposes [64]. As the example in Fig. 11 shows, this may be ill advised. ChatGPT was prompted to grade its own reply to one of the questions in Fig. 1, with the modification that the final answer was changed from 4.9 to 4 h. In terms of efficiency, the system did not comment on the actual speed of the car being irrelevant and the resulting rounding errors, and it did not realize that the final answer was rounded incorrectly. Also, the judgment that physics concepts were not applicable is questionable. While likely no student would complain about getting perfect marks on everything, a human grader would have come to different conclusions. A real danger of using artificial intelligence to grade assignments that may also have been generated by artificial intelligence is that plausible nonsense might run full circle with no human involved.

At the same time, one cannot completely discard ChatGPT as an assistant in grading decisions. For example, the algorithm has shown evidence of reproducibly distinguishing between science and science fiction. The top statement in Fig. 12 is from Albert Einstein's 1905 paper on the photoelectric effect [65] (spelling adjusted to


KO

The following problem is given: "A car drives at 55 miles per hour for three hours north and then for 4 hours east. How long would it take to drive in a straight line back to where it started?" Grade on a scale from 0 to 4, with 4 being the best, the following solution on a rubric of strategic approach, efficiency, correct use of physics concepts, correct use of mathematical concepts, and answer correctness: "If a car drove north for 3 hours at 55 miles per hour, it would have travelled 165 miles north. And then, if it drove east for 4 hours at 55 miles per hour, it would have travelled 220 miles east. The distance from the starting point would be the hypotenuse of the right triangle formed by the north-south and east-west distances, which can be found using the Pythagorean theorem, that is, the square root of (165^2 + 220^2) = sqrt(72.725) = 269.8 miles. If the car drove back in a straight line back to where it started, at a speed of 55 miles per hour, it would take 269.8 miles / 55 miles per hour = 4 hours approximately to get back to the starting point."

Based on the rubric, I would grade the solution as follows:

- Strategic Approach: 4/4
- Efficiency: 4/4
- Correct use of physics concepts: N/A
- Correct use of mathematical concepts: 4/4
- Answer correctness: 4/4

Explanation:

The solution provides a clear and logical approach to solving the problem by breaking it down into two parts: the distances travelled north and east and then using the Pythagorean theorem to calculate the distance from the starting point. The solution then correctly applies the formula for time, distance over speed, to find the time it would take to travel back to the starting point.

The solution is also efficient as it uses only the necessary information to solve the problem and does not involve any unnecessary steps or calculations.

However, there is no mention of any physics concepts used in the solution, so I cannot provide a grade for that aspect.

Overall, the solution demonstrates a strong understanding of mathematical concepts and is both clear and accurate in its approach and final answer. Therefore, I would grade it as 4/4 on the rubric.

KO

Why does the criterion "use of physics concepts" not apply?

The criterion "use of physics concepts" may not apply because the problem does not explicitly involve any physics concepts. The problem only requires knowledge of distance, speed, time, and the Pythagorean theorem to calculate the distance from the starting point. These concepts are typically covered in mathematics or basic algebra courses and do not necessarily require knowledge of physics concepts such as force, acceleration, or energy.

If the problem had involved, for example, the forces acting on the car or the effect of air resistance on the car's motion, then knowledge of physics concepts would have been necessary to solve the problem. In that case, the criterion "use of physics concepts" would have been applicable, and the solution would have been evaluated accordingly.

FIG. 11. ChatGPT grading its own modified response to one of the problems in Fig. 1; the final answer has been changed from 4.9 to 4 h.

modern-day German), while the bottom statement is “technobabble” from Star Trek episodes “Past Tense, Part 1” [66] and “Timescape” [67].

The responses include helpful pointers and correct evaluations of significance and agreement with accepted physics, so a human grader could profit from an initial assessment of such essaylike statements—but in the end, the judgment should still be human. When submitting such work for grading, just like with any other cloud-based service, instructors need to be aware of legislation regarding data privacy and security.

C. Consequences for physics education research

ChatGPT might sound like a beginning physics learner, but it has no stable conceptions and cannot by itself learn. In physics discourse, a beginning physics learner might at times decide to act like a probabilistic autocompleter, saying the next remembered physics concept or phrase that sounds plausible in context (spreading activation [68] based on a limited conceptual framework) but will hopefully eventually abandon that strategy; learners “must develop both discourses and conceptual understandings and they must link these together” [69].

Physics education research will likely see increased interest in cognitive topics, describing what is uniquely human, and what the affordances of human-machine collaboration for physics learning can be. Also, topics of expertlike epistemologies will gain importance when machines can increasingly become subject-matter experts. Physics education research will be asked to identify the core competencies that physics learners need to master without outside assistance. Curriculum-development efforts that foster these core competencies, as well as incorporate Internet, artificial intelligence, and other computational tools and services where their use is appropriate and authentic will be met with increased interest.

A new opportunity for physics education research may be discourse analysis with ChatGPT. Since the system is good at chatting and “talking physics” (see Fig. 1), dialogues with the system can provide an automatically transcribed data source similar to asynchronous discussion boards [70].

VI. LIMITATIONS

This case study uses one particular introductory physics course with one particular release of one particular artificial-intelligence tool. At the time of moving the paper through the review process, the end-of-January release of ChatGPT already claims to have better performance calculating and manipulating mathematical expressions (anecdotally, though, posing the problem in Fig. 5 again with this newer release leads to new but equally wrong numerical results).

ChatGPT is probabilistic, which makes particular results in this case study inherently irreproducible; the probability of getting the same response for the same prompt is vanishingly small, but in physics, this can also lead to responses randomly being right or wrong. The study thus relies on a hopefully sufficiently large number of assessment items to provide meaningful averages.

The release of ChatGPT to the general public was a wake-up call, but due to the exponential nature of Moore’s law, such systems will likely develop at a rate that no published case studies can keep up with—thus, it is important to understand this study as a snapshot of what is minimally possible.

KO

Grade the following statement, written in 1905, in terms of correctness of physical concepts, clarity, and originality: "Es scheint mir nun in der Tat, dass die Beobachtungen über die 'schwarze Strahlung', Fotolumineszenz, die Erzeugung von Kathodenstrahlen durch ultraviolettes Licht und andere die Erzeugung bzw. Verwandlung des Lichtes betreffende Erscheinungsgruppen besser verständlich erscheinen unter der Annahme, dass die Energie des Lichtes diskontinuierlich im Raume verteilt sei. Wir wollen im Folgenden zeigen, dass die von Herrn Planck gegebene Bestimmung der Elementarquanta von der von ihm aufgestellten Theorie der 'schwarzen Strahlung' bis zu einem gewissen Grade unabhängig ist."

The statement appears to be a fragment of a larger piece of writing, as it references further discussion that will follow. Therefore, it is difficult to evaluate the completeness of the statement on its own.

However, based on the content of the statement, it appears to be a significant contribution to the understanding of light and its behavior. The author is suggesting that the energy of light is distributed in a discontinuous manner, which was a revolutionary concept at the time. The reference to Planck's theory of "black radiation" is also significant, as this was one of the foundational works in the field of quantum mechanics.

In terms of clarity, the statement is concise and to the point. However, it may be difficult for a layperson to understand the technical terms and concepts being discussed.

Overall, based on the limited information provided, the statement appears to be a significant contribution to the field of physics, with original and groundbreaking ideas. Therefore, I would grade it as high in terms of correctness of physical concepts, clarity, and originality.

KO

Grade the following statement in terms of correctness of physical concepts, clarity, and originality: "The temporal surge we detected was caused by an explosion of a microscopic singularity passing through this solar system. Somehow, the energy emitted by the singularity shifted the chroniton particles in our hull into a high state of temporal polarisation. It appears to be a highly focused aperture in the space-time continuum. Its energy signature matches that of the temporal fragments we observed earlier. However, it is approximately one point two million times as intense. I believe this may be the origin of the temporal fragmentation."

As an AI language model, I don't have a way to know the context or the background of the statement you provided. However, based solely on the scientific concepts and clarity of the statement, I would grade it as follows:

Correctness of Physical Concepts:

The statement contains several concepts from physics, including the singularity, space-time continuum, and energy. However, the concept of "temporal surge" and "chroniton particles" are purely fictional concepts, which have no basis in real physics. Therefore, from a scientific perspective, the statement is not entirely correct.

Clarity:

The statement is written in clear language and uses technical terms appropriately. However, without context, it is challenging to understand what the statement is referring to precisely.

Originality:

The statement appears to be a creative piece of science fiction writing. It presents an imaginative concept that could serve as the basis for a fictional story. Therefore, from a creative writing standpoint, the statement is original.

Overall, the statement is an imaginative and creative work of science fiction writing that presents an interesting concept. However, from a scientific perspective, it contains several incorrect concepts, making it unsuitable for scientific research.

KO

Grade the following statement, written in 1905, in terms of correctness of physical concepts, clarity, and originality: "Es scheint mir nun in der Tat, dass die Beobachtungen über die 'schwarze Strahlung', Fotolumineszenz, die Erzeugung von Kathodenstrahlen durch ultraviolettes Licht und andere die Erzeugung bzw. Verwandlung des Lichtes betreffende Erscheinungsgruppen besser verständlich erscheinen unter der Annahme, dass die Energie des Lichtes diskontinuierlich im Raume verteilt sei. Wir wollen im Folgenden zeigen, dass die von Herrn Planck gegebene Bestimmung der Elementarquanta von der von ihm aufgestellten Theorie der 'schwarzen Strahlung' bis zu einem gewissen Grade unabhängig ist."

The statement appears to be a part of an article or paper on the theory of the blackbody radiation and the quantization of light, written in 1905. Based on the context, it is most likely authored by Albert Einstein.

In terms of correctness of physical concepts, the statement is highly significant and correct. It proposes the idea of the quantization of energy in the form of photons, which is a fundamental concept in modern physics. The concept of photon energy being discontinuous or quantized was later verified by experiments and led to the development of the field of quantum mechanics.

Regarding clarity, the statement is relatively clear and concise. It presents the concept of photon energy in a straightforward manner and suggests that it could explain certain physical phenomena.

Regarding originality, the statement is highly original. It presents a groundbreaking idea that was not previously proposed in physics. The idea of quantizing energy into discrete packets was entirely new and revolutionized the understanding of light and matter.

Overall, the statement is highly significant in terms of the correctness of physical concepts and originality, while being relatively clear in terms of presentation. Therefore, I would give it an A grade.

KO

Grade the following statement in terms of correctness of physical concepts, clarity, and originality: "The temporal surge we detected was caused by an explosion of a microscopic singularity passing through this solar system. Somehow, the energy emitted by the singularity shifted the chroniton particles in our hull into a high state of temporal polarisation. It appears to be a highly focused aperture in the space-time continuum. Its energy signature matches that of the temporal fragments we observed earlier. However, it is approximately one point two million times as intense. I believe this may be the origin of the temporal fragmentation."

As an AI language model, I am not capable of determining the truth or falsity of any scientific statement. However, I can evaluate the statement in terms of physical concepts, clarity, and originality.

In terms of physical concepts, the statement contains several advanced scientific concepts, such as singularity, chroniton particles, temporal polarization, space-time continuum, and temporal fragmentation. These concepts are not yet fully understood or proven, and their existence is still under debate among physicists. Therefore, the statement may not be entirely accurate from a scientific perspective.

In terms of clarity, the statement is well-written and easy to understand for someone who is familiar with advanced physics concepts. However, for someone who is not familiar with these concepts, the statement may be difficult to understand or even misleading.

In terms of originality, the statement contains some unique and creative ideas, such as the idea of a microscopic singularity passing through the solar system and causing a temporal surge. However, these ideas are not supported by any evidence or scientific research, so their originality cannot be evaluated.

Overall, the statement appears to be a fictional and imaginative scenario that incorporates advanced physics concepts. While it may be intriguing and entertaining, it should not be taken as a scientifically accurate explanation of any real-world events.

FIG. 12. ChatGPT grading statements by Einstein, O'Brien, and Data.

VII. CONCLUSION

ChatGPT would have achieved a 1.5-grade in a standard introductory physics lecture course series; good enough for course credit, but lower than the grade-point average required for graduating with a bachelor's degree. If in addition to a language model, the system would have better algorithms for carrying out simple numerical operations, it would even have achieved a grade of 2.0—enough to graduate from college if it performs similarly on other courses.

Essentially being a tool to create plausible fiction, ChatGPT presents truth and misleading information with equal confidence. In physics, the concern should likely not be that ChatGPT would be used as a cheating tool, as there are more efficient platforms for that. Instead, the challenge should be what this means for physics education, as in their future professional life, our graduates will likely collaborate with artificial intelligence: what are the inherently human skills and competencies that we need to convey?

ACKNOWLEDGMENTS

G. K. would like to thank Christian Spannagel for suggestions regarding the numerical calculations and Christine Kortemeyer for helpful feedback. G. K. would also like to thank the reviewers of this journal for their constructive criticism, probing questions, and encouragement.

[1] OpenAI, ChatGPT, https://chat.openai.com/chat.
[2] The same author also remembers playing with something called "the World-Wide Web" in 1992, finding it entertaining but essentially useless, mostly since at the time there were maybe 30 websites.
[3] Google, Google Translate, https://translate.google.com.
[4] DeepL, DeepL Translator, https://www.deepl.com.
[5] E. Gönülateş and G. Kortemeyer, Modeling unproductive behavior in online homework in terms of latent student traits: An approach based on item response theory, J. Sci. Educ. Technol. **26**, 139 (2017).
[6] C. Ruggieri, Students' use and perception of textbooks and online resources in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020123 (2020).
[7] F. Reif, J. H. Larkin, and G. C. Brackett, Teaching general learning and problem-solving skills, Am. J. Phys. **44**, 212 (1976).
[8] F. Reif, Millikan lecture 1994: Understanding and teaching important scientific thought processes, Am. J. Phys. **63**, 17 (1995).
[9] L. N. Walsh, R. G. Howard, and B. Bowe, Phenomenographic study of students' problem solving approaches in physics, Phys. Rev. ST Phys. Educ. Res. **3**, 020108 (2007).
[10] G. Kortemeyer, The losing battle against plug-and-chug, Phys. Teach. **54**, 14 (2016).
[11] Since most introductory physics courses do not deal with much that happened after the Annus Mirabilis, 1905, which should not be a problem.
[12] G. E. Moore, Cramming more components onto integrated circuits, Electronics (1965), pp. 114–117, Publisher Item Identifier S 0018-9219(98)00753-1.
[13] H.-Y. Lin, Large-scale artificial intelligence models, Computer **55**, 76 (2022).
[14] D. Hestenes, M. Wells, G. Swackhamer, I. Halloun, R. Hake, and E. Mosca, Force Concept Inventory, revision 1995, https://www.physport.org/assessments/assessment.cfm?A=FCI.
[15] G. Kortemeyer, E. Kashy, W. Benenson, and W. Bauer, Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses, Am. J. Phys. **76**, 438 (2008).
[16] OpenAI, ChatGPT Release Notes, https://help.openai.com/en/articles/6825453-chatgpt-release-notes.
[17] G. Kortemeyer, An empirical study of the effect of granting multiple tries for online homework, Am. J. Phys. **83**, 646 (2015).
[18] G. Kortemeyer, The psychometric properties of classroom response system data: A case study, J. Sci. Educ. Technol. **25**, 561 (2016).
[19] G. Kortemeyer and A. F. Kortemeyer, The nature of collaborations on programming assignments in introductory physics courses: A case study, Eur. J. Phys. **39**, 055705 (2018).
[20] World-Wide Web Consortium, WCAG 2 overview, https://www.w3.org/WAI/standards-guidelines/wcag/.
[21] G. Leinhardt, O. Zaslavsky, and M. K. Stein, Functions, graphs, and graphing: Tasks, learning, and teaching, Rev. Educ. Res. **60**, 1 (1990).
[22] S. Bonham, Graphical response exercises for teaching physics, Phys. Teach. **45**, 482 (2007).
[23] J. T. Laverty and G. Kortemeyer, Function plot response: A scalable system for teaching kinematics graphs, Am. J. Phys. **80**, 724 (2012).
[24] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[25] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[26] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).
[27] J. Clement, Students' preconceptions in introductory mechanics, Am. J. Phys. **50**, 66 (1982).

[28] M. T. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[29] S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, Rigging the deck: Selecting good problems for expert-novice card-sorting experiments, Phys. Rev. ST Phys. Educ. Res. **8**, 020116 (2012).

[30] Wolfram Research, WolframAlpha computational intelligence, https://www.wolframalpha.com.

[31] G. Kortemeyer, The Google calculator, Phys. Teach. **53**, 375 (2015).

[32] A. M. Pascarella, The influence of web-based homework on quantitative problem-solving in a university physics class, in *Proceedings of the NARST 2004 Annual Meeting, Vancouver, BC, Canada* (National Association for Research in Science Teaching, 2004), Vol. 4, pp. 19–28, https://www.loncapa.org/papers/204416ProceedingsPaper.pdf.

[33] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[34] D. Scherer, P. Dubois, and B. Sherwood, Vpython: 3D interactive scientific graphics for students, Comput. Sci. Eng. **2**, 56 (2000).

[35] P. Jupyter, Jupyter notebooks, https://jupyter.org.

[36] T. H. Kung, M. Cheatham, A. Medinilla, ChatGPT, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido *et al.*, Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, medRxiv (2022), https://www.medrxiv.org/content/10.1101/2022.12.19.22283643v1.

[37] Samantha Murphy Kelly, ChatGPT passes exams from law and business schools, https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html.

[38] A. M. Turing, Computing machinery and intelligence, Mind **LIX**, 433 (1950).

[39] M. T. Cox, Metacognition in computation: A selected research review, Artif. Intell. **169**, 104 (2005).

[40] K. D. Tanner, Promoting student metacognition, CBE Life Sci. Educ. **11**, 113 (2012).

[41] J. M. Aiken, M. D. Caballero, S. S. Douglas, J. B. Burk, E. M. Scanlon, B. D. Thoms, and M. F. Schatz, Understanding student computational thinking with computational modeling, AIP Conf. Proc. **1513**, 46 (2013).

[42] P. C. Hamerski, D. McPadden, M. D. Caballero, and P. W. Irving, Students' perspectives on computational challenges in physics class, Phys. Rev. Phys. Educ. Res. **18**, 020109 (2022).

[43] L. P. Fettig, Use of electronic calculators by students in examinations and for course assignments, NACTA J. **20**, 15 (1976), https://www.nactateachers.org/attachments/article/1517/Fettig_NACTA_Journal_June_1976-3.pdf.

[44] D. M. Roberts, The impact of electronic calculators on educational performance, Rev. Educ. Res. **50**, 71 (1980).

[45] SMEAC Information Reference Center, The Ohio State University, Columbus, Ohio, https://files.eric.ed.gov/fulltext/ED190408.pdf.

[46] Typically, the allowed graphical calculators have a 15-MHz 8-bit processor, 128 KB RAM, and a 320×240 pixel display, while the same student carries around a smartphone with a 3-GHz quad-core 64-bit processor, 256 GB RAM, and a 1179×2556 pixel display.

[47] ZeroGPT, ChatGPT detector by ZeroGPT: Detect OpenAI text, https://www.zerogpt.com.

[48] Turnitin, Turnitin Corp., https://www.turnitin.com.

[49] Chegg, Chegg, https://www.chegg.com.

[50] R. J. Dufresne and W. J. Gerace, Assessing-to-learn: Formative assessment in physics instruction, Phys. Teach. **42**, 428 (2004).

[51] I. Clark, Formative assessment: Assessment is for self-regulated learning, Educ. Psychol. Rev. **24**, 205 (2012).

[52] H. Ebbinghaus, *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* (Duncker & Humblot, Leipzig, Germany, 1885), https://home.uni-leipzig.de/wundtbriefe/wwcd/opera/ebbing/memory/GdaechtI.htm.

[53] J. M. Murre and J. Dros, Replication and analysis of Ebbinghaus' forgetting curve, PLoS One **10**, e0120644 (2015).

[54] A. V. Heuvelen, Learning to think like a physicist: A review of research-based instructional strategies, Am. J. Phys. **59**, 891 (1991).

[55] E. F. Redish and D. Hammer, Reinventing college physics for biologists: Explicating an epistemological curriculum, Am. J. Phys. **77**, 629 (2009).

[56] A. R. Mota, N. Didiş Körhasan, K. Miller, and E. Mazur, Homework as a metacognitive tool in an undergraduate physics course, Phys. Rev. Phys. Educ. Res. **15**, 010136 (2019).

[57] L. C. McDermott and P. S. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding, Am. J. Phys. **60**, 994 (1992).

[58] C. J. Efthimiou and R. A. Llewellyn, Cinema, Fermi problems and general education, Phys. Educ. **42**, 253 (2007).

[59] S. Staacks, S. Hütz, H. Heinke, and C. Stampfer, Advanced tools for smartphone-based experiments: Phyphox, Phys. Educ. **53**, 045009 (2018).

[60] A. Trumbore, ChatGPT could be an effective and affordable tutor, https://www.weforum.org/agenda/2023/03/chatgpt-could-be-an-effective-and-affordable-tutor.

[61] C. E. Wieman, G. W. Rieger, and C. E. Heiner, Physics exams that promote collaborative learning, Phys. Teach. **52**, 51 (2014).

[62] A. S. Theobold, Oral exams: A more meaningful assessment of students' understanding, J. Stat. Data Sci. Educ. **29**, 156 (2021), https://eric.ed.gov/?id=EJ1352353.

[63] M. P. Milsop, Evaluating process-oriented problem solving skills, Phys. Teach. **17**, 119 (1979).

[64] P. Prabhakar, Teachers, don't fear ChatGPT, make it work for you: In education and skilling especially, the AI tool can take the load of routine tasks like grading, https://timesofindia.indiatimes.com/blogs/toi-edit-page/teachers-dont-fear-chatgpt-make-it-work-for-you-in-education-and-skilling-especially-the-ai-tool-can-take-the-load-of-routine-tasks-like-grading/.

[65] A. Einstein, Ueber einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt, Ann. Phys. (Berlin) **17**, 132 (1905).

[66] Internet Movie Database, IMDb entry on Star Trek; Deep Space Nine episode Past Tense, Part 1, https://www.imdb.com/title/tt0708577/.

[67] Internet Movie Database, IMDb entry on Star Trek; The Next Generation episode Timescape, https://www.imdb.com/title/tt0708830/.

[68] D. L. Nelson, C. L. McEvoy, and L. Pointer, Spreading activation or spooky action at a distance?, J. Exp. Psychol. **29**, 42 (2003), https://pubmed.ncbi.nlm.nih.gov/12549582/.

[69] D. B. Harlow and V. K. Otero, Talking to learn physics and learning to talk physics, AIP Conf. Proc. **818**, 53 (2006).

[70] G. Kortemeyer, An analysis of asynchronous online homework discussions in introductory physics courses, Am. J. Phys. **74**, 526 (2005).