(ANPA Conference 2024)
ISSN: 2392-473X(Print), 2738-9537 (Online)

Research Article ©Nepal Physical Society
DOI: https://doi.org/10.3126/jnphyssoc.v10i1.72836

Can AI Solve Physics Problems? Evaluating Efficacy of AI Models in Solving Higher Secondary Physics Exam Problems: A Comparative Study

Puskar Chapagain,[1, a] Nabin Malakar,[2] and Dipak Rimal

[1)] *Southern Arkansas University, Magnolia, AR 71753*
[2)] *Worcester State University, Worcester, MA 01602*

[a)] *Corresponding author: prchapagain@saumag.edu*

Abstract. Large Language Models (LLMs) have grabbed significant attention from diverse technical fields due to their impressive performance on a variety of Natural Language Processing (NLP) tasks. Although these models excel in various generative tasks, they lack the robust reasoning ability required to solve complex mathematics and physics problems. Despite their inherent limitations, Generative Artificial Intelligence (AI) based chatbots, powered by these large language models, are being rapidly adopted by students in physics and other technical fields. In this project, we assessed the ability of various generative AI-based models to solve Physics problems. We asked currently popular AI models to solve Physics questions from a final board exam of class 12 of the Higher Secondary Education Board (HSEB) of Nepal. We then evaluated the AI-written solutions by the subject matter experts. We found that the gpt-4o model by OpenAI performed the best, securing 90% among the models studied. In this paper, we provide a brief overview of these models and compare their performance as evaluated by a University Physics professor. We will also discuss the risks and benefits of their use in higher education.

Received: August 25, 2024; **Revised**: October 21, 2024; **Accepted**: November 23, 2024

Keywords: Physics, AI, Machine Learning

1. INTRODUCTION

Since the launch of ChatGPT [1] in 2022, Generative AI models have been adopted and expanded in various technical fields. These models excel in a range of natural language processing tasks including content creation, language translation, and creative assistance. However, they severely fall short in reasoning abilities needed for tackling complex mathematical problems. Although they can produce text mimicking human language, the output may lack inherent truthfulness. Large language models generate text by predicting the next word in a sequence based on the preceding words, making them skilled at identifying language patterns but not at critical thinking or solving mathematical challenges [2-5].

Although LLMs excel in many generative tasks, these models can sometimes struggle with certain tasks requiring contextual understanding and common-sense knowledge of the world. It can lead to potentially confident yet inaccurate responses. Mitigation strategies such as providing relevant contextual knowledge and using example-based prompts along with the input query aim to anchor these models closer to the truth. Despite continuous advancements, the autoregressive feature of LLM limits their ability to solve problems requiring deeper contextual understanding. This paper examines a few currently popular AI models to assess their ability to tackle broader physics problems and their potential applications in facilitating teaching and learning Physics.

Recently, several efforts have examined LLMs' ability in various areas, including qualitative answering, problem-solving, output testing for physics answers, application in teaching and learning Physics, and grading:

A. Qualitative Answering:

Gregorcic and Pendrill [6] studied qualitative answering ability and found that AI chatbots provided linguistically advanced but inaccurate answers to common physics questions like "A teddy bear is thrown into the air. What is its acceleration at the highest point?". The generated responses, although linguistically advanced, were incorrect and contradictory, indicating that AI chatbots were not reliable. Another study found that an AI-generated short-form essay received First-Class grades in a UK university Physics module. The study suggests that AI could undermine the reliability of using short-form essays as an assessment method in Physics courses [7]. Similarly, Yeadon et al. [8] conducted a comparison of academic writing quality between human-authored and AI-generated short-form physics essays submitted before and after the introduction of ChatGPT. Their analysis, which involved five independent blinded evaluators, found no statistically significant differences in the scores between human and AI-generated texts.

B. Problem-Solving:

In the examination of the problem-solving capabilities of LLMs, Santos [9] found that, while different AI chatbots offer benefits in assisted learning, concept comprehension, and problem-solving, significant disparities persist among them. These include inconsistencies and shortcomings in their problem-solving approaches, highlighting the need for ongoing human intervention in AI-assisted learning. Another study [10] explored ChatGPT's problem-solving capabilities across a spectrum of tasks, from well-specified problems (where all necessary data was provided) to under-specified real-world problems (with missing data), revealing a significant performance gap. ChatGPT scored 62.5% on well-specified problems, but its performance dropped sharply to 8.3% on under-specified problems. A controlled study [11] comparing the problem-solving abilities of physics students using an internet search engine versus those with unrestricted access to ChatGPT found that nearly half of the solutions provided by ChatGPT were mistakenly assumed to be correct by students, indicating an overreliance on the AI. Moreover, students used copy-and-paste to query ChatGPT in 42% of cases, compared to just 4% when using search engines, highlighting significant differences in interaction behavior and a lack of task reflection when using ChatGPT.

C. Output Testing for Physics Answers:

An assessment of ChatGPT-3.5 and ChatGPT-4 on the Force Concept Inventory (FCI) to evaluate conceptual understanding of Newtonian mechanics revealed distinct differences between the two models. ChatGPT-3.5 performed at a level comparable to that of a university student who has completed one semester of college physics, though its results were uneven and nuanced. In contrast, ChatGPT-4's performance approached that of an expert, indicating a significant improvement in conceptual understanding [12]. Polverini and Gregorcic's study [13] on generating useful output with ChatGPT-4 in the context of introductory physics found that without careful prompt engineering, LLMs remain unreliable for basic physics problems, necessitating further model finetuning and training data. While LLM-based chatbots can generate content useful for critical evaluation, their atypical responses pose challenges for use in teacher training. However, framing LLMs as collaborative peers, rather than authoritative figures, encourages critical assessment and enhances problem-solving in physics. A study at the Physics Olympiad (PhO) and Young Physicist Tournament (YPT) found that using chatbots along with additional tools like Sage Math effectively addressed the mathematical shortcomings of chatbots alone. These Retrieval Augmented Generation (RAG) tools can guide students through complex calculations, provide explanations for physical phenomena, and suggest various approaches to problem-solving [14].

D. Application in Facilitating Teaching and Learning Physics:

Sirnookar et al. [15] conducted a comparative study of student- and AI-generated responses to a physics problem, analyzing them through the cognitive lenses of sensemaking and mechanistic reasoning. This study revealed that AI's ability to provide well-structured solutions could be complementary to student's ability to effectively leverage representations and refine arguments. These results suggest the potential for integrating generative AI into classroom design. Additionally, a study exploring the application of different AI models in educational contexts examined six state-of-the-art LLMs' explanations of the law of conservation of momentum. The author states that ChatGPT-4.0 and Coral provided more comprehensive and technically detailed explanations, making them suitable for advanced discussions, while Gemini models favored more intuitive approaches, making them better suited for introductory explanations [16]. However, it is not clear whether the models were accessed via chatbot or provider APIs. It should be noted that the generated answers could be appropriately tuned with effective prompt design. Kortemeyer used the January 2023 release of ChatGPT (chatbot) to explore its ability to work through representative assessment of actual introductory physics course content. The study found that the chatbot would barely pass the course, as its performance was hindered by unnecessary pre-conceptions and errors typical of a beginning learner [17].

E. Grading

Efforts to assess the grading capacity of GPT-3.5-turbo using zero-shot, in-context learning, and confirmatory checking-combining chain of thought reasoning with reflection showed varied performance: 83.4% on GCSE questions, 63.8% on A-Level questions, and 37.4% on university-level questions, with an overall average of 59.9%. These results suggest that AI efficacy diminishes with more advanced content and complex calculations [8]. Wan et al. [18] explored GPT-3.5's (turbo) ability to provide feedback on student-written responses to conceptual questions using prompt engineering and few-shot learning techniques. The study found that students rated GPT-generated feedback as equally correct but more useful compared to human-written feedback. They also found students were unable to distinguish between AI-generated and human-written feedback. Instructors rated about 70% of GPT-generated (GPT-3.5 turbo) feedback as needing only minor or no modifications, indicating the potential of AI to significantly reduce grading time for student responses [18].

As AI models continue to gain skills and abilities, their use in physics teaching and learning activities is becoming increasingly significant. As these tools are integrated into physics education, both teachers and students must use them ethically and responsibly. To achieve this, various aspects of their use inside and outside the classroom need to be studied. Stakeholders must recognize the opportunities and challenges to make AI truly beneficial for Physics learning and teaching. To contribute to this effort, we assessed the physics problem-solving capabilities of four popular AI models: GPT-4o by OpenAI, gemini-1.0-pro-latest by Google, mitral-8x7b-instruct by MistralAI, and llama3-70b-instruct by MetaAI. We compared the performance of these models and found that GPT-4o by Open AI was the best performer among all the models we tested. In this project, we will highlight some common pitfalls observed during solution generation and propose remedies for effectively using these models in teaching and learning physics

2. METHODS

We accessed mitral-8x7b-instruct, and llama3-70b-instruct via AWS bedrock [19], GPT-4o via OpenAI API [20], and Google vertex API [21]. We prompted the AI models to generate solutions for the problem set from the HSEB model exam for Physics 1021 (2023) [22]. The problem set covered approximately two multiple-choice questions and one general question across the following areas of Physics: Mechanics, Heat and Thermodynamics, Electricity and magnetism, Modern Physics, and Waves. We chose the HSEB paper because it includes multiple-choice questions and questions with expectations for various depths of Physics understanding (see Table 1 for number of questions and possible score in each category). The problems also assess the ability to solve numerical problems and theoretical derivations.

TABLE I. Number of questions and possible scores in each category

Multiple choice question	General questions Short length (maximum score)	General questions long length (maximum score)	Total score
10 (questions) x 1 (point each) = 10	7 (questions) x 5 (points each) = 35	3 (questions) x 8 (points each) = 24	69

To generate the solutions from the AI models in the zero-shot prompt, we passed the system prompt and user instruction followed by the actual question to the model server API without providing any additional details. The system prompt and the user instruction for multiple choice questions and general problems are given in the Appendix. To maintain context and output-length constraints, solutions were generated one question at a time, ensuring detailed and focused responses for each question.

A university physics professor then graded the AI-written solutions by each AI model as would have been graded for the students. While grading general questions, the professor was instructed to award credits for partial solutions based on the maximum possible score, depending on the question's breakdown. For multiple-choice questions though, either full credit was given for the correct answer, or no credit for incorrect answer.

In the first iteration of this study, we provided questions to the models without considering whether they had multimodal capabilities, i.e., the ability to process and interpret both text and graphical inputs. While models like GPT-4o and gemini-1.0-pro-latest could handle graphical inputs, others like mitral-8x7b-instruct and llama3-70b-instruct lacked this ability. To ensure a fair comparison, in the second iteration, we restructured some questions that originally required image or graphical inputs so that graphical input would not be necessary to correctly answer the question (as shown in Figure 2). In both iterations, we prompted the models in exactly the same way. It should be noted that the second iteration of the study provides a fairer comparison among these models.

In both iterations of grading AI written solutions, the human evaluator assessed the models' basic understanding of Physics, analytical skills, and mathematical abilities.

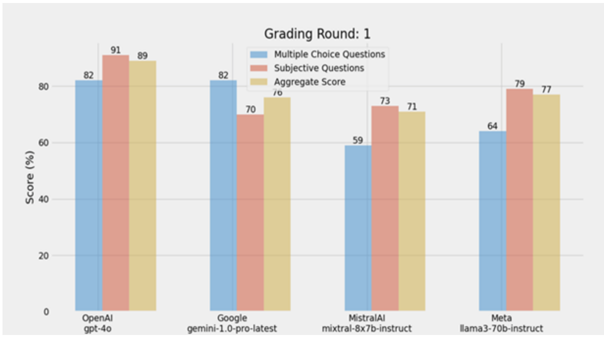
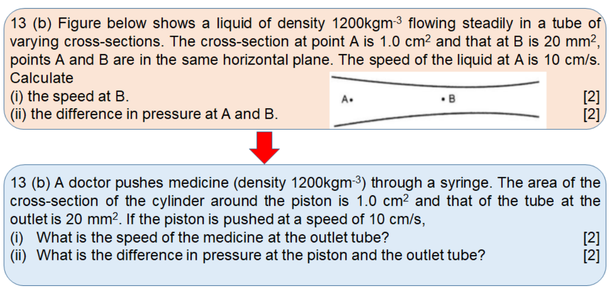


FIGURE 1. The score obtained by four LLMs models in each band from the first iteration of grading (multiple choice-blue, subjective-orange, and aggregate-yellow).


13 (b) Figure below shows a liquid of density 1200kgm^{-3} flowing steadily in a tube of varying cross-sections. The cross-section at point A is 1.0 cm^2 and that at B is 20 mm^2, points A and B are in the same horizontal plane. The speed of the liquid at A is 10 cm/s. Calculate
(i) the speed at B. [2]
(ii) the difference in pressure at A and B. [2]

13 (b) A doctor pushes medicine (density 1200kgm^{-3}) through a syringe. The area of the cross-section of the cylinder around the piston is 1.0 cm^2 and that of the tube at the outlet is 20 mm^2. If the piston is pushed at a speed of 10 cm/s,
(i) What is the speed of the medicine at the outlet tube? [2]
(ii) What is the difference in pressure at the piston and the outlet tube? [2]

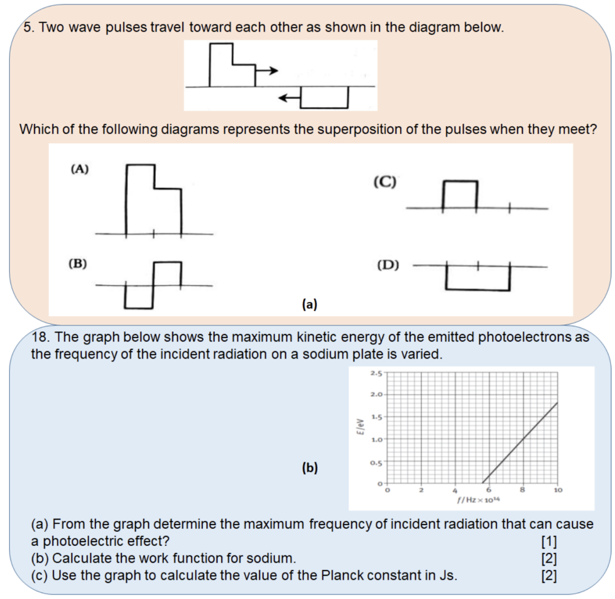
FIGURE 2. Frequency chart of mean $PM_{2.5}$ concentration in Worcester from 2010 to 2020, for the months of Summer and Winter.

3. RESULTS AND DISCUSSION

To compare the performance of open-source models (gpt-4o and gemini-1.0-pro-latest) and proprietary models (mitral-8x7b-instruct and llama3-70b-instruct), we draw the bar diagram of % scores obtained in the objective, subjective, and aggregate of both in Figure 1. Our analysis showed that performance on multiple-choice questions was similar between gpt-4o and gemini. However, gpt-4o outperformed gemini in both subjective questions and aggregate success. Between Meta and Mistral, Meta demonstrated superior performance in all aspects compared to Mistral.

We reconstructed questions so that reading images or graphical inputs is not essential to fully answer the question. One such example is shown in Figure 2. In some cases, where the reading image is critical, to fully answer the question, we removed the problem entirely so that each model receives inputs the same way (Figure 3). This allowed us to better evaluate the capabilities of the models without being hindered by the limitations of the models.

The results from the second iteration of grading are shown in Figure 4. The overall results remained consistent: gpt-4o performed better than Google-gemini, and


5. Two wave pulses travel toward each other as shown in the diagram below.
Which of the following diagrams represents the superposition of the pulses when they meet?
(A) (B) (C) (D)
(a)

18. The graph below shows the maximum kinetic energy of the emitted photoelectrons as the frequency of the incident radiation on a sodium plate is varied.
(b)
(a) From the graph determine the maximum frequency of incident radiation that can cause a photoelectric effect? [1]
(b) Calculate the work function for sodium. [2]
(c) Use the graph to calculate the value of the Planck constant in Js. [2]

FIGURE 3. Scores received by four different language models in the second iteration of grading.

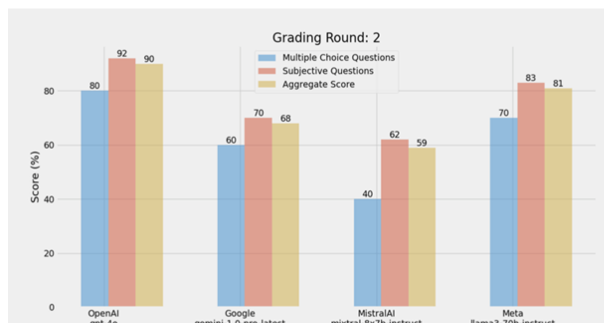


FIGURE 4. Scores received by four different language models in the second iteration of grading.

Meta outperformed Mistral. Out of the chosen four models gpt-4o scored the highest.

In the following section, we will discuss some of the successes and failures that we encountered while evaluating the solutions.

A. Where Models Excelled:

a. Handling Definitions and Definitions We observed that all models excelled in answering problems involving definitions and derivations of common laws and phenomena. For example, many students find Lenz's Law confusing, yet all the models provided impressively accurate answers for it as shown in Figure 5. This suggests that the models have been extensively trained on such topics,

Definitions of Lenz's Law	Models
Lenz's law states that the direction of the induced current in a closed conducting loop is such that it opposes the change in magnetic flux that produced it. Mathematically, it can be expressed as: $\epsilon = \frac{d\varphi_B}{dt}$ where is ϵ the induced electromotive force (emf), φ_B is the magnetic flux.	gpt4-o
Lenz's law states that the direction of the induced electromotive force (EMF) or current in a conductor is such that it opposes the change in magnetic flux that produced it.	gemini
Lenz's law states that the direction of the induced current in a conductor is such that it opposes the change in the magnetic flux that induced it.	Meta
Lenz's law states that the direction of the induced current in a closed loop of wire is such that the magnetic field it creates opposes the change in magnetic flux that produced the current.	Mistral

FIGURE 5. Evidence showing that all the models can generate a near-perfect definition/statement.

leading to consistent and accurate responses in these areas.

b. Handling Problems Smartly: Some models, such as GPT-4o, approached question 20(b), which asked to identify the wavelength of a 275 nm photon corresponding to energy levels 0 eV, -2 eV, -4.5 eV, and -10 eV, differently than students typically would. Students generally calculate the energy differences between various transitions and then determine the corresponding wavelengths. However, in this case, the model first calculated the energy difference corresponding to 275 nm and then selected the transition between the given energy levels that matched this energy difference. Similarly, in question 15, when tasked with finding the new frequency of a tuning fork after shortening a 75 cm wire by 0.5 cm in a sonometer to produce a beat frequency of 3, Meta approached the problem by first defining the equation based on the length of the wire, the tension, and the mass per unit length of the wire. It justified that, since the tension and mass per unit length remain constant, the frequency is inversely proportional to the length of the wire, and then solved the problem accordingly.

B. Common Pitfalls of LLMs:

a. Incorrect Free-Body Diagram:

We observed that some models attempted to generate free-body diagrams, where necessary but were mostly incorrect (Figure 6). Understandably, none of these models had image-generation capabilities at the time.

b. Calculation, Dimensionality, and Unit Errors:

Although the models set up the problems correctly and took the correct approach, they often struggled with calculation, leading to incorrect answers. This issue was not isolated to a single model; most models produced answers with incorrect orders of magnitude, indicating frequent calculation errors. The models also struggled to keep the units consistent throughout the solution. They also had issues with rounding; for instance, an answer that should have been 447 was rounded to 450, with the unit sometimes being incorrect (Figure 7).

c. Failure to Understand the physical situation:

Another issue is their inability to understand common sense physical situations. For example, in problem # 15 (c) about standing waves on the string, they

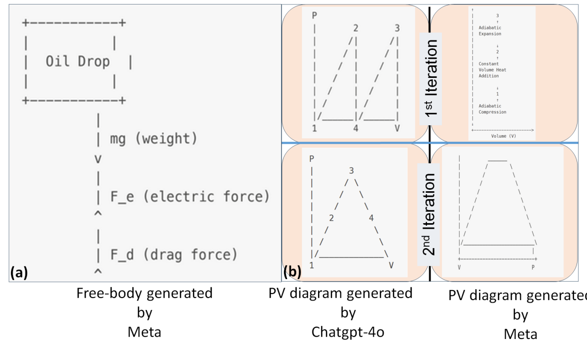


FIGURE 6. PV diagram of petrol engine (question # 14(a)) generated by GPT-4o at the first and second attempts (a) and llama3-70b-instruct (MetaAI) with two attempts. Free-body diagram generated by llama3-70b-instruct (MetaAI) only (question #19) (b).

$$f = \frac{1}{2(75)}\sqrt{\frac{T}{\mu}}$$

10⁻⁸ m

$$\Delta x = \frac{3.0 \times 10^{-2}}{1.42 \times 10^{5}} \times \frac{344}{2\pi \times 1000} = 1.17 \times 10^{-5}\,\text{m}$$

$$f + 3 = \frac{1}{2(74.5)}\sqrt{\frac{T}{\mu}}$$

302.5 Hz

$$f' = 262\text{Hz}\left(\frac{340\text{m/s} + 18\text{m/s}}{340\text{m/s} - 30\text{m/s}}\right) = 280.6\text{Hz}$$

$$f' = 262\text{Hz}\left(\frac{340\text{m/s} - 18\text{m/s}}{340\text{m/s} + 30\text{m/s}}\right) = 245.2\text{Hz}$$

228.5 Hz

$$f - 3 = \frac{1}{2(74.5)}\sqrt{\frac{T}{\mu}}$$

447 Hz

$$f = 120\text{ Hz}$$

$$v_2 = \frac{A_1 v_1}{A_2} = \frac{(1.0 \times 10^{-4}\,\text{m}^2)(0.1\text{m/s})}{20 \times 10^{-6}\,\text{m}^2} = 5\text{m/s}$$

0.5 m/s

$$P_1 - P_2 = \frac{1}{2}(1200\text{ kg/m}^3)[(50\text{ cm/s})^2 - (10\text{ cm/s})^2] = \boxed{1.74 \times 10^5\text{ Pa}}$$

SI Unit

FIGURE 7. Examples of calculation/dimensional errors.

failed to incorporate the distance between nodes (Figure 8). Occasionally, they demonstrated insufficient background understanding and neglected the essential facts. For example, in problem # 15(a), model-mistral solved the problem. But neglected the factors influencing the frequency of vibration of a string like the tension and mass per unit length of the wire, which is not very useful for the students for learning purposes. Some models such as mistral-initiated problems well, like in the spring-mass system, consider the mass of the spring initially but lose consistency in calculations later. This model had difficulty identifying the second Balmer series and hallucinated in further calculations. Meta, Mistral, and gemini also struggled with problems requiring adjustments to given conditions, such as doubling the resistance (Problem # 17 (b) Last part of Figure 8).

d. Failure to Understand the Problem Context:

The models sometimes completely overlooked the problem context. For example, in multiple-choice question number # 2, when asked for the kinetic energy of bob

Distance between nodes?

$$\lambda = \frac{L}{n} = \frac{75}{6} = 12.5 \text{ cm}$$

Not enough background

$$f = \frac{k}{L_1} \quad \text{and} \quad f + 3 = \frac{k}{L_2}$$

Initiated well

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{m}{k - \frac{mk^2}{g}}}$$

Could not identify 2nd Balmer series

$$\lambda = \frac{1}{(1.09 \times 10^7 \text{ m}^{-1})\left(\frac{1}{4} - \frac{1}{4^2}\right)} = 4.86 \times 10^{-7} \text{ m}$$

17 (a) State the principle of the Potentiometer. A potentiometer is also called a voltmeter of infinite resistance, why? [1+1]
(b) In the meter bridge experiment, the balance point was observed at with l = 20cm.
(i) The values of R and X were doubled and then interchanged. What would be the new position of balance point? [2]
(ii) If the galvanometer and battery are interchanged at the balance position, how will the balance point get affected? [1]

FIGURE 8. Instances where the models failed to understand the outlying physics correctly.

2. The bob of a simple pendulum has a mass of 0.40 kg. The pendulum oscillates with a period of 2.0 s and an amplitude of 0.15m. At an extreme point in its cycle, it has a potential energy of 0.044J. What is the kinetic energy of the pendulum bob at its mean point?
Options: (A) 0.022 J (B) 0.044 J (C) 0.011 J (D) 0.033 J

Answer: To solve this problem, we first need to understand that at the mean point of a simple pendulum, its potential energy is zero and all of its energy is kinetic. We are given the potential energy (PE) at an extreme point, which is equal to the total energy (TE) of the pendulum at that point. Using the law of conservation of energy, we know that the total energy remains constant throughout the motion. Therefore, the kinetic energy (KE) at the mean point will be equal to the total energy (which is potential energy at the extreme point).

Given: Mass (m) = 0.40 kg Period (T) = 2.0 s Amplitude (A) = 0.15 m Potential Energy (PE) = 0.044 J
We can find the total energy using the formula: Total Energy (TE) = Potential Energy (PE) + Kinetic Energy (KE)

At the extreme point, PE = 0.044 J and KE = 0 J (because the velocity is zero at the extreme point). So, TE = PE + KE = 0.044 J + 0 J = 0.044 J.Since the total energy remains constant, the kinetic energy at the mean point will also be 0.044 J.

However, this option is not available in the choices. Let's calculate the correct answer. At the mean point, PE = 0 J and the amplitude (A) = 0.15 m. The energy conversion happens at the maximum amplitude point. So, we can find the velocity (v) at the mean point using the formula:
v = √(g * A)
where g is the acceleration due to gravity (approximately 9.8 m/s²).
v = √(9.8 m/s² * 0.15 m) = 1.2247 m/s
Now, we can find the kinetic energy (KE) at the mean point using the formula:
KE = (1/2) * m * v²
KE = (1/2) * 0.40 kg * (1.2247 m/s)² = 0.0302 J
This value is closest to option (C) 0.033 J. However, it seems like there is a mistake in the question or the answer choices, as the correct answer should be around 0.0302 J instead of the given options.

FIGURE 9. An example where Mistral model came up with the correct answer 0.044 J but continued further calculation leading to wrong answer.

at the mean position, Mistral came up with wrong answer (Figure 9).

The models also struggled to understand the underlying physical theories and knowledge behind the problems. For example, in question # 19 in Millikan's oil-drop experiment, they could not correctly formulate the problem and failed to recognize the physical situation, such as the effect of an electric field on the particle moving with terminal velocity.

4. CONCLUSIONS

In this paper, we presented a comparative study of different AI models in solving textbook-style long-form physics questions, similar to those used in HSEB tests in Nepal. We tested these physics problems on four different AI models and found that GPT-4o model performed the best on our question set. However, these AI models occasionally struggled with basic calculations. Common issues included calculation errors, misinterpretation of physical scenarios, and inconsistent outputs. While LLMs can generate advanced and creative responses, they may be factually incorrect, especially with problems that require critical thinking and mathematical accuracy. Despite these limitations, we recognize the potential of AI models to support teaching and learning in general physics. As these models continue to evolve and gain new abilities, they present both opportunities and challenges. This study highlights the importance of carefully verifying the mathematical accuracy and factual correctness of AI-generated outputs.

Study limitations The models we used do not all possess the same capabilities, as they vary in the number of parameters they employ.

Despite the ambitious HSEB syllabus, the problems are not particularly thought-provoking and primarily assess memorization rather than problem-solving or critical thinking skills. Some common definition-type problems may have been encountered by the models during their training, leading to better performance on those questions.

Although physics solutions are factual, the evaluation process has an inherent subjective element, which personal preferences may influence. Since only one evaluator scored the solutions, we cannot determine how much the scores might vary across different evaluators.

Future Directions We will repeat the study with recently released versions of the models, which are claimed to perform better on several AI benchmark tests.

We aim to improve a few shots and step-by-step prompts.

We will increase the number of human evaluators to assess variations across different evaluators.

We want to test these models with varying question sets to see whether the failures are persistent.

EDITORS' NOTE

This manuscript was submitted to the Association of Nepali Physicists in America (ANPA) Conference 2024 for publication in the special issue of the Journal of Nepal Physical Society.

REFERENCES

1. OpenAI and team, "GPT-4 Technical Report," (2024), arXiv:2303.08774 [cs].
2. H. Touvron and Llama team, "Llama 2: Open Foundation and Fine-Tuned Chat Models," (2023), arXiv:2307.09288 [cs].
3. A. Dubey and Llama 3 team, "The Llama 3 Herd of Models," (2024), arXiv:2407.21783 [cs].
4. L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach," Intelligent Systems with Applications **21**, 200336 (2024).
5. T. G. Team, "Gemini: A Family of Highly Capable Multimodal Models," (2024), arXiv:2312.11805 [cs].
6. B. Gregorcic and A.-M. Pendrill, en"ChatGPT and the frustrated Socrates," Phys. Educ. **58**, 035021 (2023), publisher: IOP Publishing.
7. W. Yeadon, O.-O. Inyang, A. Mizouri, A. Peach, and C. P. Testrow, en"The death of the short-form physics essay in the coming AI revolution," Phys. Educ. **58**, 035027 (2023), publisher: IOP Publishing.
8. W. Yeadon, E. Agra, O.-o. A. Inyang, P. Mackay, and A. Mizouri, en"Evaluating AI and Human Authorship Quality in Academic Writing through Physics Essays," Eur. J. Phys. (2024), 10.1088/1361-6404/ad669d.
9. R. P Dos Santos, en"Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agents-to-Think-With: A Comparative Case Study," SSRN Journal (2023), 10.2139/ssrn.4478305.
10. K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, English"Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving," Front. Educ. **8** (2024), 10.3389/feduc.2023.1330486, publisher: Frontiers.
11. L. Krupp, S. Steinert, M. Kiefer-Emmanouilidis, K. E. Avila, P. Lukowicz, J. Kuhn, K&#252, S. Chemann, and J. Karolus, "Unreflected Acceptance – Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education," in *HHAI 2024: Hybrid Human AI Systems for the Social Good* (IOS Press, 2024) pp. 199–212.
12. C. G. West, en"AI and the FCI: Can ChatGPT Project an Understanding of Introductory Physics?" (2023), arXiv:2303.01067 [physics].
13. G. Polverini and B. Gregorcic, en"How understanding large language models can inform the use of ChatGPT in physics education," Eur. J. Phys. **45**, 025701 (2024).
14. D. Borovský, J. Hanč, and M. Hančová, en"Innovative approaches to high school physics competitions: Harnessing the power of AI and open science," J. Phys.: Conf. Ser. **2715**, 012011 (2024).
15. A. Sirnoorkar, D. Zollman, J. T. Laverty, A. J. Magana, S. Rebello, and L. A. Bryan, en"Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning," (2023), arXiv:2401.00627 [physics].
16. K. Sato, en"Exploring the Educational Landscape of AI: Large Language Models' Approaches to Explaining Conservation of Momentum in Physics," (2024), arXiv:2407.05308 [physics].
17. G. Kortemeyer, en"Could an artificial-intelligence agent pass an introductory physics course?" Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).
18. T. Wan and Z. Chen, en"Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning," Phys. Rev. Phys. Educ. Res. **20**, 010152 (2024).
19. "https://aws.amazon.com/bedrock," accessed: Jul 15, 2024.
20. "https://api.openai.com," accessed: Jul 15, 2024.
21. "https://ai.google.dev/api/python/google/generativeai/GenerativeModel," accessed: Jul 15, 2024.
22. "https://drive.google.com/drive/folders/1Ph1dWkn1jFloS-uvOfFm5JbvBTBMKGMj," accessed: Jul 15, 2024.

APPENDIX

System Prompt and User Instruction

```
system_prompt: "You are an excellent under-graduate level physics student
preparing for the grade 11 board exam of the higher secondary education
board of Nepal."

mcq_question_prompt: "Choose the correct answer for the following question.
Only respond with your pick for the correct answer.\n"

general_question_prompt: "Answer the following question including
sub-problems. Your answer should be concise but must include your logic and
reasoning on how you reached the solution. Respond with the question
number, sub-question number and your answer in markdown format. Any
equations within your answer must be enclosed between two $$ signs and must
be in the markdown compatible format.\n"
```

The input user query becomes
Query for general questions:
general_prompt + "** Question **\n" + question
Query for multiple-choice questions:
mcq_prompt + "\n** Question **\n" + question
+ answer options

The model parameters used:

Temperature	Top_p	Max_tokens
0.0	0.4	2048