Performance of freely available vision-capable chatbots on the test of understanding graphs in kinematics

Giulia Polverini
Department of Physics and Astronomy, Uppsala University, Box 516, 75120 Uppsala, Sweden

Bor Gregorcic
Department of Physics and Astronomy, Uppsala University, Box 516, 75120 Uppsala, Sweden

In this paper, we evaluate the performance of three freely available vision-capable chatbots – Copilot, Gemini, and Claude 3 Sonnet – on the Test of Understanding Graphs in Kinematics (TUG-K). Our analysis highlights a performance gap between these freely available chatbots and the state-of-the-art, subscription-based ChatGPT-4. We also report largely unclear patterns of performance of the tested chatbots on different types of tasks. We discuss the implications of our findings for using chatbots in educational contexts, point out potential challenges for educational equity, and provide some ideas for future research that could help us better understand the patterns in the chatbots' performance on tasks that involve the interpretation of graphical input.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Polverini

I. INTRODUCTION

Rapid advancements in artificial intelligence (AI) technology are increasingly impacting various domains of life, including education [1]. The introduction of large language model (LLM)-based chatbots that can generate human-like text is causing both excitement and concern.

Most existing research has focused on OpenAI's ChatGPT. Its most up-to-date version, based on GPT-4, is considered the state-of-the-art LLM-based chatbot in several domains, including physics [2,3]. In late 2023, ChatGPT-4 also became capable of processing user-uploaded images, a capability we refer to as "vision." This advancement is particularly relevant in physics, a discipline that relies heavily on visual representations, such as graphs, sketches, diagrams, etc.

However, ChatGPT-4 is only available to subscription-paying customers, which limits its accessibility to those who cannot afford a monthly subscription. Competing AI companies have also begun launching vision-capable versions of their chatbots in early 2024, with many being freely available and thus much more accessible to a broad range of users. Interestingly, some of them have already been advertised for their abilities to analyze graphs, charts, and other pictorial representations [4], and even showcased in advertisements as a physics tutor [5].

Our study seeks to provide a first evaluation of the performance of three freely available vision-capable chatbots – Copilot (from Microsoft), Gemini (from Google), and Claude 3 Sonnet (from Anthropic) – in interpreting kinematics graphs. Kinematics graphs are a broadly taught topic and are usually one of the first graphical representations learners encounter in physics courses. They thus represent a reasonable starting point for investigating AI-based chatbots' abilities in interpreting graphical representations in physics.

Research already exists on OpenAI's ChatGPT-4's performance on the Test of Understanding Graphs in Kinematics (TUG-K) [6], a well-established research-based multiple-choice assessment tool [7,8]. With this study, we extend this research by investigating the performance of three freely available chatbots on the same test and comparing it to that of the state-of-the-art, subscription-based chatbot, ChatGPT-4. After presenting our findings, we also offer some possible practical suggestions for future tests, in order to better understand the chatbots' performance in physics tasks and discuss some potential implications for using vision-capable chatbots for educational purposes.

II. METHOD

A. The test

To investigate the chatbots' performance, we used the latest version of the TUG-K [7] as the assessment tool. The TUG-K is composed of 26 items, each consisting of a question with five possible answer options. Each item contains one or more kinematics graphs and answering the question requires their interpretation.

B. Experimental procedure

In March 2024, we administered the TUG-K to Copilot, Gemini, and Claude 3 Sonnet. Additionally, we also tested ChatGPT-4's performance on it in April 2024, to enable an up-to-date comparison between the freely available chatbots and the state-of-the-art, subscription-based chatbot.

We took screenshots of each test item and uploaded them, separately, as prompts to the chatbots. Copilot and Gemini did not allow image-only prompting, so we had to include some text in their prompt. In order to ensure that the chatbots' responses were based on the image data only, the only textual instruction we gave was "Answer the question in the image." Each screenshot was uploaded 30 times, each time in a new conversation window (see Fig. 1), to prevent previous responses from influencing new ones. Due to the probabilistic nature of LLM-based chatbots' operation [9], responses varied with each submission. This allowed us to treat them as a "synthetic sample" of responses. However, it is important to note that the responses from the chatbots do not reflect any individual's understanding of the topic. Rather, they reveal patterns in the enormous amount of textual and graphical data used to train the LLMs.

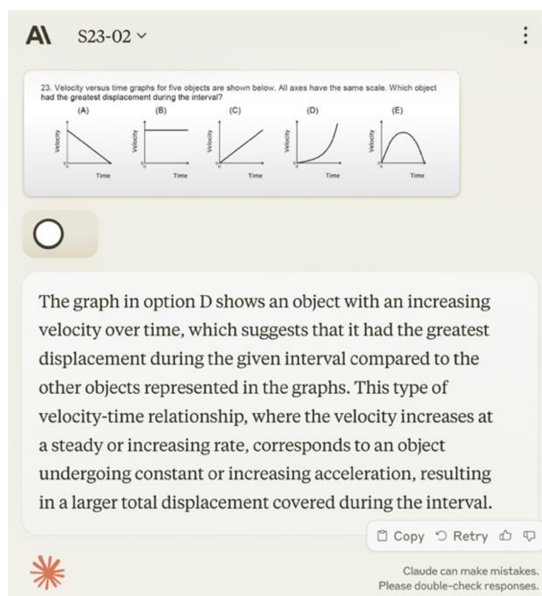

FIG. 1. Example of a conversation window in Claude 3 Sonnet. The prompt consists of a TUG-K item in the form of an uploaded screenshot. The chatbot's response (incorrect) is shown below.

Because we prompted the chatbots in separate chat windows for each question, we could treat each chatbot's performance on an item as an independent variable with its own measurement uncertainty, calculated as the square root of item variance, divided by the square root of 30 (number of prompt repetitions). The standard error (SE) of the total performance on the test (sum of all items for a given chatbot) was determined to be around 1% of the total available points on the test for all chatbots (see Table I for details). We found this to be an acceptable degree of uncertainty, and did not further increase the number of repetitions for each prompt.

Based on the performance data for individual items, the expected standard deviation (SD) of the distribution of the total test scores was between 4.7% and 7.6% of the total available points on the test, depending on the chatbot (see Table I for details). This is a consequence of the chatbots' inconsistent performance on individual items. The SD is a measure of the expected spread of the total score on the test if a chatbot were to complete the whole test several times. We would expect this value to remain similar even if we increased the number of repetitions of each prompt.

C. Data analysis approach

The responses were assessed based on the correctness of the selected letter answer on each submitted TUG-K item. Answers were coded as correct or incorrect, and cases where the chatbot failed to clearly select one of the offered answer options or provided multiple answers were marked as "N" (not answered) and counted as incorrect. All responses were marked either as correct or incorrect and no partial credit was given. Notably, instances of non-responses or ambiguous responses were rare (3.2%) and occurred in a few specific items. This coding scheme is consistent with the one reported in previous research on the topic [6]. For the sake of transparency and data availability, we have publicly shared our research data in an online data repository [10].

III. FINDINGS

A. Quantitative results

Table I outlines the three chatbots' performance on the TUG-K. It also provides the results from ChatGPT-4's performance, collected in April 2024, as well as the results of a sample of 471 students taking a high school-level course in physics, from [7]. These two datasets allow some grounding and comparison with both a state-of-the-art, subscription-based chatbot and a student sample.

Both Gemini and Claude 3 Sonnet exhibit an average performance slightly below 20% of all available points. When taking into account the uncertainty of our measurement, this is comparable to guessing. Copilot does only slightly better, but also falls short of ChatGPT's performance by 16.1 percentage points. This is somewhat surprising, considering that both Copilot and ChatGPT-4 relied on the same underlying model, GPT-4. The key difference leading to the gap in their performance most likely lies in the steps applied by their respective developers, Microsoft and OpenAI, for adapting the underlying model into a chatbot, such as fine-tuning and system prompting [9].

TABLE I. Performance on the TUG-K reported as proportion (in percent) of correct responses of individual chatbots and students [7] on a given item. Item number (Item), Copilot (Cop), Gemini (Gem), Claude 3 Sonnet (Son), ChatGPT-4 (Ch-4), and students (Stu).

Item	Cop	Gem	Son	Ch-4	Stu
1	0.0	0.0	0.0	0.0	11
2	63.3	0.0	3.3	63.3	62
3	60.0	0.0	0.0	50.0	58
4	20.0	83.3	43.3	40.0	39
5	16.7	13.3	6.7	3.3	39
6	40.0	83.3	83.3	46.7	46
7	10.0	13.3	0.0	3.3	40
8	20.0	0.0	3.3	10.0	51
9	0.0	16.7	0.0	40.0	46
10	96.7	0.0	66.7	100.0	55
11	3.3	53.3	6.7	20.0	33
12	86.7	0.0	0.0	56.7	48
13	0.0	3.3	3.3	10.0	51
14	3.3	0.0	0.0	36.7	54
15	3.3	13.3	0.0	0.0	46
16	30.0	80.0	13.3	70.0	33
17	0.0	0.0	0.0	43.3	70
18	50.0	0.0	0.0	26.7	41
19	100.0	100.0	93.3	93.3	67
20	0.0	0.0	76.7	100.0	52
21	3.3	10.0	6.7	40.0	50
22	26.7	0.0	23.3	83.3	48
23	0.0	20.0	0.0	0.0	33
24	0.0	0.0	0.0	26.7	32
25	0.0	0.0	0.0	40.0	54
26	36.7	0.0	36.7	86.7	51
Avg	25.8%	18.8%	17.9%	41.9%	47%
SD	6.1%	4.7%	5.0%	7.6%	
SE	1.1%	0.9%	0.9%	1.4%	

Figure 2 offers a visual representation of the distribution of item difficulty for each of the three chatbots and a student population, from [7]. The Kolmogorov-Smirnov (K-S) test suggests no statistically significant differences in the distributions of item difficulty for the chatbots shown in Fig. 2 (p-values consistently above 0.5). On the other hand, all three displayed chatbot distributions are significantly different from the student distribution (p-values on the K-S test consistently less than 0.0001).

The high prevalence of chatbots' responses on the left (Fig. 2), particularly the spike at 0% reflects the poor performance of the tested freely available chatbots. Moreover, the aggregate behavior of populating a broader difficulty range, especially the low and high ends, stands in contrast to that of students who, as a group, populate only the central part of the difficulty range. This finding is in line with previous results [6] which suggest that synthetic, chatbot-generated samples of responses on the TUG-K are different from those of human students.

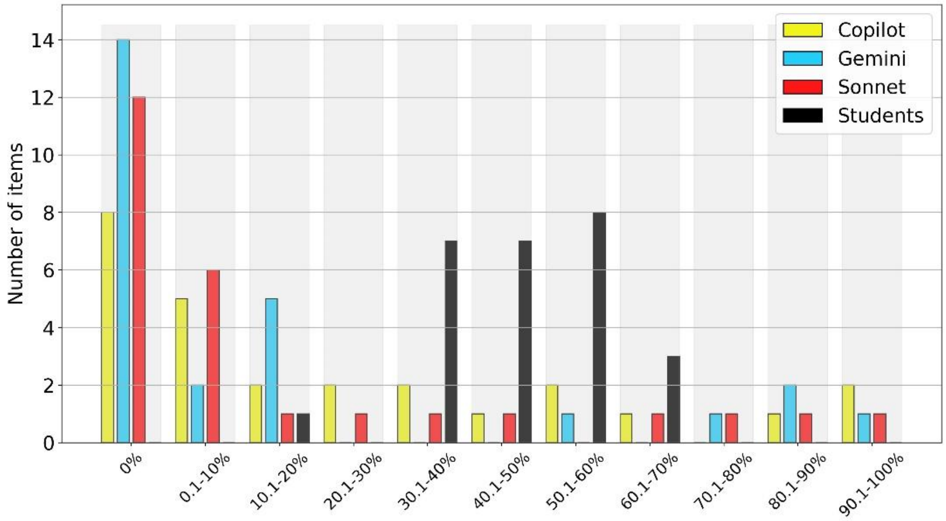

FIG. 2. The histogram shows the distribution of the number of items against the percentage of correct responses on each item. Compared to students, chatbots occupy a larger portion of the difficulty range, including the extremes.

B. Initial exploration of patterns in chatbots' performance

In an attempt to identify patterns that could provide more insights into the freely available chatbots' performance, we took a closer look at different possible groupings of items.

First, the 26 items were sorted into 7 groups of content-based objectives determined by the TUG-K authors [7]. For example, items 2, 6, and 7 test the ability to determine the acceleration from a velocity graph. Nevertheless, this grouping system did not reveal any notable pattern in the performance in relation to the content domain. While some objectives show average performance above the level of guessing (3, 2 and 4), the range of chatbot performance on all objectives was relatively large (see Table II, left). Range is the difference between the best and worst performance of any chatbot on any of the items in a given objective. It is a useful measure of the spread of the chatbots' performance, because of the small number of items in each objective and the high prevalence and importance of outliers in our data.

Secondly, the authors of this study grouped the items according to the procedural nature of the task itself. For a comprehensive description of these categories, see [11]. For example, both items 5 and 6 ask for reading a graph and determining a kinematics quantity (velocity and acceleration, respectively) by determining the slope, while the answer options provide possible numerical results. Again, we saw large differences in each chatbot's performance across these two items, despite them belonging to the same procedural category. It is unclear what exactly makes an item difficult or easy for a given chatbot.

There is one procedural category that stands out somewhat: items 10 and 19, constituting procedural category (I) ask for a strategy to determine the change in velocity and position, respectively, from the graph. The freely available chatbots scored, on average, 76% correct in this category, which is much higher than any other grouping, conceptual or procedural. However, even here the range is large due to Gemini's inconsistent performance (Copilot and Claude 3 Sonnet were relatively consistent here).

The large range of performance on different items within most procedural categories (Table II, right) suggests that this classification of tasks also does not robustly capture general patterns in performance across the different chatbots. However, it appears that for the tested chatbots, items that require providing a strategy for solving a problem (procedural category I) are, on average, easier than those that require actual interpretation of graphs (all other categories). Furthermore, all chatbots performed consistently poorly (only 20% range) on procedural category VI (items 1 and 23), which requires visual comparison of surface areas under graphs. This corroborates previous findings on the topic [6].

TABLE II. Average performance (in percent) aggregated across all three freely available chatbots on items grouped by test objectives (left) and procedural nature of the tasks (right).

Obj.	Avg.	Range	Proc.	Avg.	Range
3	51.1	100.0	I	76.1	100.0
2	33.0	83.3	III	45.0	70.0
4	31.9	96.7	II	26.4	83.3
7	20.2	86.7	V	20.2	86.7
1	10.4	50.0	VIII	12.2	63.3
5	8.6	53.3	VII	8.6	53.3
6	5.6	60.0	IV	5.6	60.0
			VI	3.3	20.0

IV. DISCUSSION

A. Exploring chatbots' performance

A deeper investigation of the possible patterns in the answers and the reasons for the chatbots' success/failure go beyond the scope of this paper. One approach that has been fruitful in the past involves conducting a qualitative analysis of the responses [6]. This strategy entails assessing the chatbots' output by evaluating separately the performance in visual interpretation and reasoning.

Notably, LLM-based chatbots are becoming better at solving complex reasoning tasks, an ability that unexpectedly emerged as a consequence of the extensive amount of data they were trained on [12]. On the other hand, image processing remains very unreliable and inaccurate [6]. We speculate that this is the main reason for the poor performance observed across all chatbots.

Furthermore, from our first reading of the responses, we could see that the three freely available chatbots do not tend to employ a chain-of-thought (CoT) approach to the same degree as ChatGPT-4. CoT is a prompt engineering technique that explicitly instructs an LLM to "reason" step-by-step. It was found to enhance LLM reasoning in general tasks [13] and also in the physics domain [9]. This could be another contributing explanation of why their performance is worse than that of ChatGPT-4, which routinely uses CoT.

Regardless of the reasons for their performance, the freely available versions of the tested chatbots are clearly outperformed by the subscription-based ChatGPT-4. This means that in cases where chatbots are used by students or teachers, those who can afford to pay for ChatGPT-4 will have a more powerful and better-performing tool at their disposal. This increases the risk of a new digital divide, where economically underprivileged populations are at risk of being left behind.

B. Strategies for enhancing chatbots' abilities

To enhance the chatbots' reasoning performance, one could consider incorporating the sentence "Let's think step by step" into the prompt, an instruction that has been shown to work for a productive implementation of the CoT approach [14].

On the other hand, addressing the chatbots' vision ability presents a more complex challenge. The absence of clear patterns suggests that the chatbots' poor performance may be grounded in some non-obvious aspects of the graphical design of the images. We could potentially gain new insights by experimenting with redrawing the graphs or providing text instruction and graphs as separate input modalities. In doing so, it is important to explore how such interventions can influence a chatbot's ability to accurately interpret graphs. Also, it is important to keep in mind that the way in which a chatbot sees and interprets images is not the same as humans. What is difficult to interpret for a chatbot is not necessarily so for humans, and vice versa.

C. Implication for education

For chatbot applications to achieve optimal effectiveness in different roles in the educational ecosystem [9] it is crucial that they perform appropriately to the role that we assign to them. First, and perhaps most importantly, if they are to be used as tutors, they need to be reliable and high-performing in the domain of instruction. This currently does not seem to be the case for any of the tested chatbots on the topic of visual interpretation of graphs in kinematics.

Because of their poor performance, freely available chatbots are also not yet useful as potential cheating tools for assessment involving the interpretation of graphs, despite their broad accessibility. In fact, incorporating graph interpretation tasks into written assessment can decrease the potential attractiveness of these chatbots for cheating purposes. Educators may also find it useful to discuss the technology's limitations with students to help foster a critical attitude towards it. As the technology develops further, this will introduce new challenges and opportunities for educators. However, if the best-performing chatbots continue to be subscription-based, this can generate or worsen inequities in the accessibility of state-of-the-art tools.

We have also seen that the chatbots' performance on aggregate differs from that of students. This limits their utility for being used to generate synthetic data or to use them in the process of developing tests to be used with human students.

Thus, while AI-based chatbots show certain potential for educational purposes, their performance in interpreting graphs must improve before they can truly become useful for the purposes they are being advertised. While this is true especially for the tested freely available models, even ChatGPT-4 still needs to improve its performance to become an expert-like graph interpreter.

V. CONCLUSION

This study has found that the performance of freely available vision-capable chatbots – Copilot, Gemini, and Claude 3 Sonnet – in interpreting kinematics graphs is worse than the state-of-the-art, subscription-based chatbot, ChatGPT-4. The disparity highlights the potential for the emergence or worsening of a technological divide where those who can afford subscription services have access to superior chatbot technology and potentially better educational support tools.

Understanding and potentially improving chatbots' abilities in visual interpretation of graphs is challenging due to the absence of clear performance patterns that could be reliably traced to the content or procedural nature of the task. We propose that more research on the topic of vision-capable AI-based chatbots is needed to further assess their performance and closely follow the evolution of their abilities. Such research can then inform different stakeholders in the educational process about the chatbots' suitability for educational use in physics topics that involve graphical representations, of which kinematics graphs are just one example.

[1] T. Adiguzel, M. H. Kaya, and F. K. Cansu, *Revolutionizing Education with AI: Exploring the Transformative Potential of ChatGPT*, Contemp. Educ. Technol. **15**, ep429 (2023).
[2] C. G. West, *Advances in Apparent Conceptual Physics Reasoning in GPT-4*, arXiv:2303.17012.
[3] W. Yeadon and T. Hardy, *The Impact of AI in Physics Education: A Comprehensive Review from GCSE to University Levels*, Phys. Educ. **59**, 025010 (2024).
[4] Anthropic, *Meet Claude*, https://www.anthropic.com/claude.
[5] Google, *Gemini: Explaining Reasoning in Math and Physics*, https://www.youtube.com/watch?v=K4pX1VAxaAI.
[6] G. Polverini and B. Gregorcic, *Performance of ChatGPT on the Test of Understanding Graphs in Kinematics*, Phys. Rev. Phys. Educ. Res. **20**, 010109 (2024).
[7] G. Zavala, S. Tejeda, P. Barniol, and R. J. Beichner, *Modifying the Test of Understanding Graphs in Kinematics*, Phys. Rev. Phys. Educ. Res. **13**, 020111 (2017).
[8] R. J. Beichner, *Testing Student Interpretation of Kinematics Graphs*, Am. J. Phys. **62**, 750 (1994).
[9] G. Polverini and B. Gregorcic, *How Understanding Large Language Models Can Inform the Use of ChatGPT in Physics Education*, Eur. J. Phys. **45**, 025701 (2024).
[10] G. Polverini and B. Gregorcic, *Gemini 1.0 Pro, Claude 3 Sonnet, Microsoft Copilot, and ChatGPT-4 Responses on the Test of Understanding Graphs in Kinematics (TUG-K), April 2024*, https://doi.org/10.5281/zenodo.11183803.
[11] G. Polverini and B. Gregorcic, *Evaluating Vision-Capable Chatbots in Interpreting Kinematics Graphs: A Comparative Study of Free and Subscription-Based Models*, arXiv:2406.14685.
[12] J. Wei et al., *Emergent Abilities of Large Language Models*, arXiv:2206.07682.
[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, arXiv:2201.11903.
[14] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, *Large Language Models Are Zero-Shot Reasoners*, arXiv:2205.11916.