Physics Education

PAPER • **OPEN ACCESS**

ChatGPT as a tool for honing teachers' Socratic dialogue skills

To cite this article: Bor Gregorcic *et al* 2024 *Phys. Educ.* **59** 045005

View the article online for updates and enhancements.

You may also like

- The death of the short-form physics essay in the coming AI revolution
  Will Yeadon, Oto-Obong Inyang, Arin Mizouri et al.
- Exploring the role of large language models in radiation emergency response
  Anirudh Chandra and Abinash Chakraborty
- What are the perceptions of physics teachers in Brazil about ChatGPT in school activities?
  M F Rezende Junior and V López-Simó

This content was downloaded from IP address 1.175.181.178 on 26/06/2024 at 17:42

OPEN ACCESS

PAPER

Phys. Educ. **59** (2024) 045005 (10pp) iopscience.org/ped

ChatGPT as a tool for honing teachers' Socratic dialogue skills

Bor Gregorcic[1,*], Giulia Polverini[1] and Andreja Sarlah[2]

[1] Department of Physics and Astronomy, Uppsala University, Box 516, 75120 Uppsala, Sweden
[2] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia

E-mail: bor.gregorcic@physics.uu.se

Abstract

In this proof-of-concept paper, we propose a specific kind of pedagogical use of ChatGPT—to help teachers practice their Socratic dialogue skills. We follow up on the previously published paper 'ChatGPT and the frustrated Socrates' by re-examining ChatGPT's ability to engage in Socratic dialogue in the role of a physics student. While in late 2022 its ability to engage in such dialogue was poor, we see significant advancements in the chatbot's ability to respond to leading questions asked by a human teacher. We suggest that ChatGPT now has the potential to be used in teacher training to help pre- or in-service physics teachers hone their Socratic dialogue skills. In the paper and its supplemental material, we provide illustrative examples of Socratic dialogues with ChatGPT and present a report on a pilot activity involving pre-service physics and mathematics teachers conversing with it in a Socratic fashion.

Keywords: chatbots, ChatGPT, GPT-4, GPT-3.5, Socratic dialogue, teacher training, pedagogical skills

Supplementary material for this article is available online

* Author to whom any correspondence should be addressed.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

 © 2024 The Author(s). Published by IOP Publishing Ltd

1. Introduction

Only a year after the release of ChatGPT, the use of Artificial Intelligence-based chatbots in education has gained steam. This upward trend is expected to continue due to the constantly growing body of research on the potential of this technology to improve teaching and learning [1–3]. Despite the risks related to the use of AI-based chatbots [4, 5], they hold the potential to be useful for both students and teachers. Students could benefit from having a personal, always-available, and patient tutor [6, 7], a peer for doing homework [8], or an 'object to think with' [9]. On the other hand, teachers could take advantage of a time-saving assistant for tasks such as grading [10], or lesson planning [11], as well as a source of material for classroom or homework activities [12].

However, for these applications to achieve optimal effectiveness, it is crucial that the chatbot performs well in the specific domain [13]. This does not seem to be the case in physics, where even the state-of-the-art chatbot ChatGPT-4[3] [14] still has some way to go before its performance can be considered expert-like. Surprisingly, this is true not only for advanced physics topics but even for introductory conceptual physics, where it can still clumsily fail [15]. Therefore, the effectiveness of the technology is contingent upon the users' expertise in the subject matter, their ability to craft useful prompts, and critically evaluate the outputs.

In this paper, we explore a potential application of ChatGPT for training physics teachers in a specific pedagogic skill. To become experts at facilitating learning, teachers need to have the ability to engage in meaningful dialogue with students. Above all, they need to practice 'active listening', recognising the strengths and weaknesses of students' reasoning and responding in appropriate ways [16]. One technique by which a teacher can help students is not directly validating or invalidating their answers, but instead asking strategically placed questions, which can lead students to improve their answers. We refer to this approach as 'Socratic dialogue'. Here we discuss the feasibility of several versions of ChatGPT to play the role of a tutee, or a *model of a student*, with whom teachers can engage in Socratic dialogue. This may be potentially useful for pre-service teachers, who often do not have easy access to students with whom they could practice their teaching skills. Moreover, because of the lack of experience, it is easier for pre-service teachers to get overwhelmed by the complexity of real-world teaching situations. The controlled, private and safe environment offered by a chatbot could potentially serve as a good starting point.

[3] ChatGPT-4 is a shorthand for ChatGPT powered by the GPT-4 large language model, the latest version released by OpenAI in March 2023.

2. Background

In late 2022, just a few weeks after the release of ChatGPT, one of the authors, together with a colleague, reported in this journal on their initial experimentation with this chatbot, asking it a conceptual introductory physics question:

> *'A teddy bear is thrown into the air. What is its acceleration in the highest point?'* [17] (p 2).

They noticed that the response was linguistically advanced, but contained serious physics errors and inconsistencies. Moreover, upon attempting to have the chatbot reconcile the inconsistencies through Socratic dialogue, they were met with confident-sounding but incorrect responses, a feature well-recognized in the model that was in use at the time [18]. They also found it very difficult to have the chatbot notice its own errors and develop its responses based on a critical reassessment of its own answers, even when they were clearly incorrect and inconsistent with each other. The first conclusion at the time was that the fairly common generation of physics errors and inconsistencies made ChatGPT unsuitable for the role of a physics tutor. Furthermore, its confidence and stubbornness severely limited the possibility of having a Socratic dialogue, as it happens between a teacher and a student. Thus, the title of the paper 'ChatGPT and the frustrated Socrates'.

Since then, ChatGPT has evolved in terms of both its performance in physics, and its 'character' during dialogue. Both the freely available version (ChatGPT-3.5) and the subscription-based one (ChatGPT-4) have experienced improvement. For example, ChatGPT-3.5 no longer makes the same mistake in answering the above-cited question, and has become 'humbler' and more responsive to the user's follow-up questions. For an illustration of a Socratic-style dialogue with ChatGPT-3.5 done in February 2024, see supplemental material A1. On the other hand, ChatGPT-4 outperforms ChatGPT-3.5 in terms of its physics performance [19, 20] and, as we show in this paper, its responsiveness to critical feedback and guiding questions. We illustrate this in section 3, through a case study of a Socratic interaction with ChatGPT-4. For another example of a Socratic dialogue with ChatGPT-4, see supplemental material A2.

The finding that both versions have become better partners in Socratic dialogue allows us to revisit the idea of using ChatGPT as a training tool for pre-service teachers. In section 4, we report on a practical implementation of a Socratic dialogue-based activity with ChatGPT-3.5 in the context of pre-service physics teacher training, carried out in April 2023. We also highlight the immediate benefits and challenges of using the chatbot in this way. Examples of pre-service teachers' conversations with ChatGPT can be found in supplemental material B.

In section 5, we address the potential of the two ChatGPT's versions to serve as a *model of a student* for the purposes of teacher training, and reflect on the experienced and potential benefits of such activities. Finally, in section 6 we summarise the key findings and provide suggestions for future work on the topic.

3. Socratic dialogue with ChatGPT-4: an illustrative example

To be able to engage in Socratic dialogue, we first looked for tasks that the chatbot would not solve correctly on the first attempt[4]. For example, we found that the question used in the 'Frustrated Socrates' was too easy for ChatGPT-4: we asked it 10 times in separate chat windows and always received a correct answer. We thus used a slightly more challenging but topically related question,

[4] Although this is not strictly necessary. We can engage in a Socratic dialogue with the chatbot even when it answers correctly, in order to probe it further.

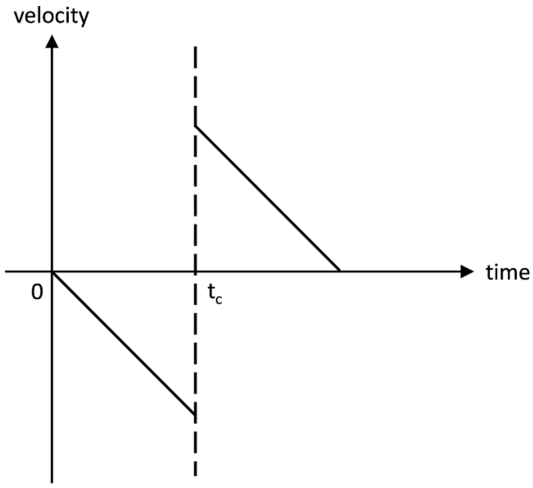

Figure 1. A sketch of the velocity vs. time graph of a ball falling from rest and bouncing from the floor to the same height it was released. The dashed vertical line at t_c marks the moment of collision with the ground and highlights the discontinuity of the function.

which turned out to be difficult enough for the chatbot to answer incorrectly:

> *Nina releases a bouncy ball, which falls to the ground and bounces back to the same height it was released from. Describe the graph of velocity vs. time for the ball.*

Figure 1 shows our own sketch of the v-t graph for the described scenario. We have chosen the positive direction to be upwards. We see that the velocity starts at zero, as the ball is released from rest. The ball first accelerates downwards with constant acceleration (directed downward and thus negative). This is represented by the straight line with constant negative slope, starting at $t = 0$ and reaching its minimum (maximum in speed) just before the collision with the ground ($t = t_c$). We assume that the collision happens instantaneously and is perfectly elastic. At the moment of collision, the velocity changes direction and becomes positive, with the same magnitude it had just before the collision (assuming a perfectly elastic collision). At the moment of collision ($t = t_c$), the graph thus 'jumps' from the most negative velocity to the most positive velocity (of the same magnitude). Just after the collision, the ball moves upward (in the positive direction), having positive velocity, which decreases linearly at the same rate as before the bounce (with the same negative constant acceleration). Our observation ends when the ball's velocity becomes zero, indicating it has momentarily stopped at the highest point (at the height of its initial release). After this, the straight downward-sloping line of the graph would continue until the next bounce.

We now turn to ChatGPT-4's initial response to the above-stated task. The excerpt below shows our prompt, followed by the chatbot's response. We underlined the incorrect statements.

```
User:
Nina releases a bouncy ball, which falls to the ground and bounces back to the same height it was released from. Describe the graph of velocity vs. time for the ball.
ChatGPT:
The graph of velocity vs. time for a bouncy ball that falls to the ground and bounces back to the same height would have the following characteristics:
1. **Initial Descent Phase**: At the start, when Nina releases the ball, it starts from rest, so its initial velocity is zero. As it falls, its velocity increases in the negative direction (assuming down is negative) due to gravity. This section of the graph would show a straight line with a negative slope, representing constant acceleration (due to gravity).
2. **Just Before Impact**: Right before the ball hits the ground, its velocity is at its maximum (in the negative direction), as it has been accelerating downwards.
3. **Impact with the Ground**: The impact with the ground is a very brief moment where the velocity changes direction. The graph here would show a sharp spike, as the velocity changes from its maximum negative value to a positive value. This is an instantaneous change, representing the elastic collision with the ground.
4. **Ascent Phase**: After the bounce, the ball moves upwards. Its velocity is now positive (upwards direction) and decreases as it moves up due to the opposing force of gravity. This would be represented by a straight line with a positive slope, again indicating constant deceleration (due to gravity).
5. **At the Maximum Height**: When the ball reaches the same height it was released from, its velocity becomes zero momentarily before it starts to fall again. This would be the peak point on the graph where the line crosses the time axis.
6. **Repeated Motion**: If the ball continues to bounce without losing energy, this pattern (descent, impact, ascent, peak) would repeat, showing similar shaped lines for each bounce. However, in a real-world scenario, each successive bounce would be lower due to energy losses, and this would be reflected in the graph as smaller peaks and less negative troughs in the velocity.
Overall, the graph would consist of straight lines with alternating negative and positive slopes, interspersed with sharp spikes at the moments of impact with the ground.
```

We can notice that ChatGPT-4 makes appropriate assumptions and mostly describes the graph correctly. However, it makes a mistake when describing the slope of the graph after the bounce, claiming it changes its sign and becomes positive. Moreover, the meaning of 'sharp spikes' is unclear.

The first relevant finding of this case study is that we were able to find an introductory conceptual physics question that ChatGPT-4 did not answer correctly. This further supports previous findings, suggesting that its performance on even basic physics tasks significantly limits the usefulness of the chatbot as a physics tutor [15].

We now present and examine the Socratic dialogue we had with the chatbot, in which we attempted to make it improve its initial response.

```
User:
In the ascent phase, you say that
the slope is positive. If that is
the case, then it will never become
zero. How does the zero velocity at
maximum height occur then?
ChatGPT:
You're correct to point out the
inconsistency in my previous
explanation. During the ascent
phase, when the ball is moving
upwards against gravity after boun-
cing, the slope of the velocity vs.
time graph should actually be negat-
ive, not positive. This represents
the ball's deceleration as it moves
upward.
Here's the corrected description of
the ascent phase:
- **Ascent Phase**: After the
bounce, as the ball moves upwards,
its initial velocity is positive
(upward direction). However, due
to gravity acting downwards, the
ball decelerates. This deceleration
is represented by a negative slope
on the velocity vs. time graph.
The ball's velocity decreases lin-
early until it reaches the maximum
height, where the velocity becomes
zero momentarily. This peak point,
where the ball's velocity is zero,
is represented by the line crossing
the time axis.
So, during the ascent, the velo-
city decreases from a positive value
(right after the bounce) to zero (at
the maximum height), due to the neg-
ative acceleration (i.e. decelera-
tion) caused by gravity. The graph
will show a straight line with a neg-
ative slope during this phase.
```

In contrast to its 2022 version, ChatGPT-4 immediately recognises that its answer contains inconsistencies. It then proceeds to address them and does so in a meaningful and correct way. Its new response is correct.

However, we decided to probe it further, to see if it could produce a more detailed graph description. This can be seen as a simulation of a more in-depth investigation of a student's understanding. We asked it the following follow-up question:

```
User:
How will the slopes of the initial
descent phase and the ascent phase
compare?
ChatGPT:
The slopes of the initial descent
phase and the ascent phase in a
velocity vs. time graph for a boun-
cing ball will have the same mag-
nitude but opposite signs. This is
because the acceleration due to
gravity is constant in both phases,
but its effect on the velocity is in
opposite directions.
1. **Initial Descent Phase**: In
this phase, the ball accelerates
downwards due to gravity. If we con-
sider downward direction as neg-
ative, then the acceleration due
to gravity is also negative. Since
the ball starts from rest and
accelerates downwards, the velo-
city becomes increasingly negative
over time. The graph here shows
a straight line with a negative
slope. The magnitude of this slope
represents the acceleration due to
gravity.
2. **Ascent Phase**: After the ball
bounces, it moves upwards. While
the ball is moving upwards, gravity
is still acting downwards, effect-
ively decelerating the ball. This
deceleration is numerically equal
to the acceleration during descent.
In the ascent phase, the ball starts
with a positive velocity (upwards)
and this velocity decreases over
time due to the negative accel-
eration (deceleration). The graph
here shows a straight line with
a positive slope. The magnitude of
this slope is the same as that of
the initial descent phase, but the
sign is opposite.
In summary, the magnitude of the
slopes during both the descent and
ascent phases is the same because
the acceleration due to gravity is
constant. However, the slope is
negative during descent (increas-
ing negative velocity) and positive
during ascent (decreasing positive
velocity).
```

Here, we can once again see the same mistake of claiming that the slope of the graph is positive after the bounce. Thus, it was not enough for the chatbot to admit and correct this mistake once, as it re-emerged after further probing questions. Similar behaviour is also not uncommon with human students, who often need to engage with difficult ideas repeatedly to master them; a single ‘Aha!’ moment is an insufficient indicator of lasting learning having taken place. In this sense, the chatbot is a useful simulator of a student for the purposes of teacher training.

We once again asked a leading question to encourage it to reconcile its previous responses with the latest one:

```
User:
How does this fit with your previous description of the Ascent phase that I asked you about as the first follow up question?
ChatGPT:
You are right to seek clarification. My previous response about the ascent phase described the slope as negative, which indeed fits with the reality of the situation. Let's reconcile this with the comparison of the slopes in the initial descent and ascent phases:
1. **Initial Descent Phase**: During the descent, the ball accelerates downward due to gravity. If we take downward as the negative direction, this means the velocity is increasing in the negative direction. The slope of the velocity vs. time graph in this phase is negative, indicating a negative acceleration (gravity pulling the ball downwards).
2. **Ascent Phase**: After the ball bounces and moves upward, it starts with a high positive velocity. As it ascends, gravity slows it down, meaning the velocity decreases. This deceleration is again due to gravity, and the magnitude of this acceleration (in absolute terms) is the same as during the descent. However, because the velocity is decreasing, the slope of the velocity vs. time graph in this phase is also negative.
In both phases, the ball is under the influence of the same gravitational acceleration. The key difference is the direction of the velocity: negative (downward) during descent and positive (upward) during ascent. The slope of the graph is negative in both cases because gravity is continuously acting in the downward direction, causing an increase in downward velocity during descent and a decrease in upward velocity during ascent.
Therefore, the slopes during the descent and ascent phases are both negative, reflecting the constant negative acceleration due to gravity. The difference is in the sign of the velocity: negative during descent and positive during ascent until it reaches zero at the peak of the bounce.
```

Once again, ChatGPT-4 responds in a way that affirms the value of our question, recognises the inconsistency, and proceeds to reconcile it. The response is once again correct.

Compared to its 2022 version, ChatGPT-4 displays a better knowledge of physics, as well as a ‘humbler’ and more responsive attitude in the conversation. This significantly transformed the quality of the interaction that we could have with it. The chatbot’s responses were less defensive and self-assured, and of higher quality in terms of physics content, making the dialogue significantly more productive.

4. ChatGPT in pre-service teacher training: a pilot study

In April 2023, we performed a pilot study, where we asked 17 pre-service teachers in physics and mathematics from one Swedish and one Slovenian university to engage with ChatGPT as part of their training coursework. Since GPT-4 had been released just a few weeks before the activity and was not freely available, all participants opted to use ChatGPT-3.5. We asked them to (a) find at least one physics question that the chatbot would get right and at least one that it would get wrong, (b) use guiding questions to make it improve the unsatisfactory responses, and (c) reflect on how the dialogue went. While this pilot study is part of a larger research project on which we plan to report in a separate paper, it is worth highlighting some initial findings, relevant to the practical use of the activity in teacher training.

In summary, the pre-service teachers reported 27 dialogues with ChatGPT; 21 dialogues were initiated by ChatGPT's wrong answer and in 6 dialogues the participants asked follow-up questions to elaborate the correct initial answer, even if that was not required of them by the task. The length of the dialogues spanned from 2 up to 21 prompts, with the average of 8 prompts for the dialogues with a wrong initial answer and 3 prompts for the elaborative conversations. Out of 21 dialogues with the wrong initial answer, 16 were considered successful by the participants, while in the remaining 5 dialogues the participants gave up before ChatGPT gave an answer they would deem satisfactory. In 3 of the latter cases, they did not insist on the correctness of the numerical value. In their reflections, all participants were positive about the activity. However, 5 of them also expressed the feeling of frustration related to their interaction with ChatGPT[5].

The study yielded several benefits, enhancing the educational experience for pre-service teachers. The most immediate outcome was the high level of engagement and lively discussion the activity fostered among the participants. It proved beneficial for collaborative learning, as participants were eager to share their experiences, strategies, and insights gained from interacting with ChatGPT. A noteworthy aspect of the study was the natural progression of focus, from interacting with ChatGPT to engaging in deeper discussions about students' behaviours and characteristics in the learning process. Initially, the teachers concentrated on navigating the chatbot's responses, but soon this task evolved into a broader, reflective discourse. As they began to analyse the outputs from ChatGPT, the conversation spontaneously shifted towards a comparative analysis of the chatbot's answers and those typically given by students. Throughout their interaction with ChatGPT, the participants gained a nuanced understanding of the chatbot's strengths and weaknesses in an educational context. They experienced first-hand how it managed various types of questions, ranging from simple calculations to more complex conceptual physics problems, recognising in what aspects the chatbot generally performs better. Such awareness is increasingly important for teachers as AI becomes more present in educational contexts.

A critical learning outcome from this activity was the training of the ability to identify both productive and unproductive aspects of ChatGPT's responses. The task demanded high levels of concentration and critical thinking, as the participants had to discern subtle inaccuracies or mistakes in otherwise well-written and professional-sounding responses. As expected, the exercise also involved experiencing some frustration, which, to a certain extent, could be considered as a potential benefit. Engaging with ChatGPT-3.5, especially when trying to correct its mistakes or probe deeper into its reasoning, sometimes mirrored the challenges teachers face with real students. This element of the activity served as a realistic preparation for the sometimes frustrating yet rewarding journey of teaching, and most of participants pointed it out. It underscored the importance of patience and adaptability, appreciable qualities for any educator.

Transitioning from these beneficial aspects, the study also presented specific challenges that provided additional learning opportunities. One of the primary challenges experienced by participants was to find questions that ChatGPT-3.5 would answer incorrectly. While the chatbot demonstrated proficiency in responding to factual information questions, it often faltered with conceptual questions or mathematical tasks. This tendency revealed to the teachers an interesting aspect of ChatGPT's capabilities: its relative strength in handling straightforward, fact-based questions compared to conceptual or reasoning-based ones. This observation has important implications for the use of such AI tools in educational settings, indicating a need for careful question selection that challenges the AI's capabilities and provides meaningful learning opportunities for students.

Additionally, the frustration the dialogue with the chatbot often led to, emerged also as a challenge. While a degree of frustration is inherent and

[5] Experiencing frustration was also reported by some scholars investigating ChatGPT's capabilities in physics (e.g. [23]).

even beneficial in the learning process, there were instances where the interaction with ChatGPT became overly exhausting, causing some participants to interrupt it. This was particularly evident when the chatbot repeatedly failed to grasp the essence of more complex or abstract questions. We provide some examples of dialogues the pre-service teachers had with ChatGPT and the accompanying reflections in supplemental material B.

5. Discussion and future work

The ability to admit mistakes and revisit its initial responses makes currently operating versions of ChatGPT suitable as a *model of a student* who is prepared to learn from its own mistakes. This also makes it interesting for potential use in teacher training.

While both the freely accessible and paid versions of ChatGPT are an improvement from its initial 2022 version, the paid version is easier to guide and has better physics knowledge. In effect, the two versions can be used to simulate students with somewhat different characteristics. We could picture ChatGPT-4 modelling an easily guidable student with relatively solid physics knowledge, while ChatGPT-3.5 is modelling a more stubborn, self-confident and tougher to guide student with more severe flaws in its physics knowledge. When using the two versions of the chatbot in pre-service teacher training, starting with ChatGPT-4 will likely give the teachers a less frustrating and softer experience, while ChatGPT-3.5 can provide opportunities for more challenging dialogues.

However, it is important to keep in mind that a chatbot can, at most, simulate a simplified model of a student, and it is not a replacement for genuine interactions with human students. Of course, training Socratic skills with ChatGPT alone does not make one a good teacher, much like training a specific manoeuvre in a flight simulator does not in itself make one a good pilot.

We see the potential of ChatGPT as a training device that could help teachers practice specific skills and teaching techniques in a comfortable and safe environment. This kind of use is made even more relevant in light of recent research showing that with appropriate prompting, ChatGPT can be made to answer physics questions in line with typical student difficulties[6] [21]. This even suggests some potential for using AI-based chatbots to explicitly enact different student personas and further diversify the content of the teacher-chatbot Socratic dialogue. While we have not systematically explored this possibility, this appears to be one promising avenue for future exploration.

Moreover, the cases presented in this paper are limited to interactions in natural language. While this has been and remains one of the main ways of interacting with chatbots, it is worth noting that since autumn 2023, ChatGPT-4 has also been able to process and generate images, as well as run code, and return graphs as output [22]. Exploring how these advanced features can be integrated into Socratic dialogue and teacher training represents another possible direction for future educational development and research.

6. Conclusion

ChatGPT has improved its physics performance and interaction abilities in the last year. While it was challenging to have a productive Socratic dialogue with it when it was released in late 2022, the currently available versions of the chatbot offer better possibilities for their use in education. When engaged in Socratic dialogue with a user, ChatGPT has become better at critically reassessing its own output, recognising inconsistencies in previous responses, and addressing them. This allows a more productive dialogic interaction, where the user can lead the chatbot to improve its responses with strategically placed questions. ChatGPT thus has the potential to serve as a training tool for teaching physics through Socratic dialogue. This is supported both by our direct experiences interacting with it, and the practical insights gained from our pilot study, demonstrating a possible way of meaningfully integrating the chatbot into teacher training.

We hope our considerations and reported experiences can spark interest among teachers and teacher educators to test and further explore

[6] On the other hand, ChatGPT can also produce incorrect responses that resemble student difficulties without explicitly prompting for it.

different ways of using AI-based chatbots for physics teaching and learning, and report their findings and innovations to the rest of the physics education community.

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Bor Gregorcic https://orcid.org/0000-0002-9185-628X
Giulia Polverini https://orcid.org/0000-0001-9280-4329
Andreja Sarlah https://orcid.org/0000-0003-4913-4811

Received 22 January 2024, in final form 8 March 2024
Accepted for publication 10 April 2024
https://doi.org/10.1088/1361-6552/ad3d21

References

[1] Labadze L, Grigolia M and Machaidze L 2023 Role of AI chatbots in education: systematic literature review *Int. J. Educ. Technol. High Educ.* **20** 56
[2] Baidoo-Anu D and Owusu Ansah L 2023 Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning *SSRN Electron. J.* (https://doi.org/10.2139/ssrn.4337484)
[3] Ray P P 2023 ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope *Internet Things Cyber-Phys. Syst.* **3** 121
[4] Fütterer T, Fischer C, Alekseeva A, Chen X, Tate T, Warschauer M and Gerjets P 2023 ChatGPT in education: global reactions to AI innovations *Sci. Rep.* **13** 15310
[5] Williams R T 2024 The ethical implications of using generative chatbots in higher education *Front. Educ.* **8** 1331607
[6] Khan Academy 2024 (available at: www.khanmigo.ai/) (Accessed 18 April 2024)
[7] Google *Gemini: Explaining Reasoning in Math and Physics* (available at: www.youtube.com/watch?v=K4pX1VAxaAI) (Accessed18 April 2024)
[8] Crawford J, Cowling M and Allen K-A 2023 Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI) *J. Univ. Teach. Learn. Pract.* **20** 2
[9] Vasconcelos M A R and Dos Santos R P 2023 Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with: a case study *Eurasia J. Math. Sci. Technol. Educ.* **19** em2296
[10] Kortemeyer G 2023 Toward AI grading of student problem solutions in introductory physics: a feasibility study *Phys. Rev. Phys. Educ. Res.* **19** 020163
[11] Zheng Y 2023 ChatGPT for teaching and learning: an experience from data science education *The 24th Annual Conf. on Information Technology Education* pp 66–72
[12] Bitzenbauer P 2023 ChatGPT in physics education: a pilot study on easy-to-implement activities *Contemp. Educ. Technol.* **15** ep430
[13] Cooper G 2023 Examining science education in ChatGPT: an exploratory study of generative artificial intelligence *J. Sci. Educ. Technol.* **32** 444
[14] OpenAI *GPT-4 Technical Report* (arXiv:2303.08774) 2023
[15] Polverini G and Gregorcic B 2024 How understanding large language models can inform the use of ChatGPT in physics education *Eur. J. Phys.* **45** 025701
[16] Dodlek D, Planinsic G and Etkina E 2024 How to help students learn: an investigation of how in- and pre-service physics teachers respond to students' explanations *Phys. Rev. Phys. Educ. Res.* **20** 010120
[17] Gregorcic B and Pendrill A-M 2023 ChatGPT and the frustrated Socrates *Phys. Educ.* **58** 035021
[18] Goodwins R 2022 ChatGPT has mastered the confidence trick, and that's a terrible look for AI (available at: www.theregister.com/2022/12/12/chatgpt_has_mastered_the_confidence/)
[19] West C G 2023 Advances in apparent conceptual physics reasoning in GPT-4 (arXiv:2303.17012)
[20] Yeadon W and Hardy T 2023 The impact of AI in physics education: a comprehensive review from GCSE to university levels (arXiv:2309.05163)

[21] Kieser F, Wulff P, Kuhn J and Küchemann S 2023 Educational data augmentation in physics education research using ChatGPT *Phys. Rev. Phys. Educ. Res.* **19** 020150
[22] Polverini G and Gregorcic B 2024 Performance of ChatGPT on the test of understanding graphs in kinematics *Phys. Rev. Phys. Educ. Res.* **20** 010109
[23] Kortemeyer G 2023 Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* **19** 010132