Challenges and Opportunities of Moderating Usage of Large Language Models in Education

Lars Krupp[a,b] LARS.KRUPP@DFKI.DE
Steffen Steinert[c] STEINERT.STEFFEN@PHYSIK.UNI-MUENCHEN.DE
Maximilian Kiefer-Emmanouilidis[a,b] MAXIMILIAN.KIEFER@DFKI.DE
Karina E. Avila[b] KAVILA@RPTU.DE
Paul Lukowicz[a,b] PAUL.LUKOWICZ@DFKI.DE
Jochen Kuhn[c] JOCHEN.KUHN@LMU.DE
Stefan Küchemann[c] S.KUECHEMANN@LMU.DE
Jakob Karolus[a,b] JAKOB.KAROLUS@DFKI.DE
[a] *German Research Center for Artificial Intelligence (DFKI)*
[b] *RPTU Kaiserslautern-Landau*
[c] *LMU Munich*

Abstract

The increased presence of large language models (LLMs) in educational settings has ignited debates concerning negative repercussions, including overreliance and inadequate task reflection. Our work advocates moderated usage of such models, designed in a way that supports students and encourages critical thinking. We developed two moderated interaction methods with ChatGPT: hint-based assistance and presenting multiple answer choices. In a study with students (N=40) answering physics questions, we compared the effects of our moderated models against two baseline settings: unmoderated ChatGPT access and internet searches. We analyzed the interaction strategies and found that the moderated versions exhibited less unreflected usage (e.g., copy & paste) compared to the unmoderated condition. However, neither ChatGPT-supported condition could match the ratio of reflected usage present in internet searches. Our research highlights the potential benefits of moderating language models, showing a research direction toward designing effective AI-supported educational strategies.

Keywords: ChatGPT, Large Language Models, Education, Physics, LLM usage

1. Introduction

There are ongoing debates on the usage of ChatGPT in schools and universities (University of Cambridge, 2023). Fueling these debates is GPT4's passing of the Bavarian A-levels (BR, 2023). Likewise, research has already shown potential negative impacts on learning methods (Santos, 2023). Furthermore, it has been shown that students critically miss reflection when interacting with LLMs to help them answer physics questions (Krupp et al., 2023). Finding the right compromise between moderating the usage of LLMs and leveraging their untapped potential remains a challenge.

To address this challenge, we designed and implemented different variants of output moderation for ChatGPT and evaluated those in a user study where students were tasked

to answer physics questions. We constructed two moderated tools: the Multi-Response Bot (MRB), which returns three different answers, and the Hint Bot (HB), which only gives hints instead of answering questions. We further implemented two baseline tools: a Search Engine (SE) and the Classic Bot (CB), an unmoderated version of ChatGPT. This allowed us to observe different interaction strategies and their success for each moderated tool and compare them with each other and the baseline.

Our results show that moderation can improve reflection and student performance positively. Both the MRB and the HB decreased how often the copy & paste strategy was used compared to the CB, indicating an improved rate of reflection.

In this work, we contribute an investigation into the effects of moderating the usage of LLMs in education on students' performance and problem solving strategies. Moderation can foster reflection and critical thinking but is subject to design constraints such as limited usability and poor understanding of LLMs by users. Our work opens up a research direction toward designing effective AI-supported educational strategies, leveraging the advantages of LLMs while still allowing for reflection and critical thinking from students.

2. Related Work

The domain of language models (LMs) is rapidly expanding with diverse applications relevant to education. They have been successfully employed for tasks like generating multiple-choice questions (Raina and Gales, 2022) or providing answers to them (Zhang et al., 2022). Given the accessibility and power of LMs like ChatGPT, it is probable that students will harness these tools at home. Furthermore, as shown by Krupp et al. (Krupp et al., 2023), unmoderated access to LMs, such as ChatGPT, leads to low amounts of reflection in students. They tend to trust the chatbot even in their domain of expertise. This issue highlights a need to evaluate how access to LMs can be moderated to support students efficiently.

The application of Large Language Models (LLMs) like ChatGPT in physics education has garnered considerable attention, although with mixed results. Several studies have reported inconsistencies in ChatGPT's responses to physics questions (Gregorcic and Pendrill, 2023; Santos, 2023). Predominantly, the model exhibited a tendency to present incorrect answers, leading some researchers to consider it ill-suited for roles like physics tutoring or aiding in homework. However, Bitzenbauer turned this apparent shortcoming into an opportunity by encouraging students to critically evaluate ChatGPT's responses, thereby enhancing their critical thinking skills (Bitzenbauer, 2023). Contrary to the shortcomings mentioned above, other research showcases the proficiency of later GPT versions (3.5 and 4) in tackling conceptual multiple-choice physics questions (West, 2023b,a). Notably, ChatGPT successfully answered most of the force concept inventory items (West, 2023a). Furthermore, Kieser et al. postulate GPT 4's potential in mimicking student difficulties, which could pave the way for tailored student support and enhanced task creation for educators (Kieser et al., 2023). An interesting phenomenon observed is the tendency for prospective physics teachers to rely heavily on ChatGPT. Küchemann et al. (2023) found that in a comparative study, these educators often used tasks provided by ChatGPT verbatim, without adaptation. Additionally, the study revealed a lesser inclination to contextualize tasks within real-world scenarios when using ChatGPT instead of traditional textbooks. Furthermore, Krupp et al. (Krupp et al., 2023) have first evaluated what strate-

gies students employ when having unrestricted access to ChatGPT. They have shown that using copy & paste is the most common strategy, indicating overreliance and that it leads to worse results than using a search engine when used to answer complex physics questions.

3. Methodology

In our work, we present a first look into how LLM output moderation can affect the strategies students use when interacting with them and their results when answering physics questions. We employed a between-subject design where we changed the available tool for students to use. Students participating in our online study were randomly assigned to one of four conditions. Two baseline conditions are represented through the classic ChatGPT chatbot (CB) and an internet search engine (SE), representing a common support tool for homework (Lenhart et al., 2001). For our experimental conditions, we designed and implemented different variants of output moderation for ChatGPT: the Multi-Response Bot (MRB) and the Hint Bot (HB). This model- and domain-agnostic approach allowed us to use an LLM without trainable weights, where fine-tuning is not possible. As opposed to using a specifically trained physics LLM, this model-agnostic approach allows for greater generalization of our moderation strategies across LLMs.

The MRB was designed to generate three different answers to each question asked using prompt engineering. We visualized the answer options next to each other from which the participant could select their preferred answer, which would become part of the conversation history. To make this decision, participants had to read, understand, and compare all options with each other, which inherently fosters critical thinking (Bitzenbauer, 2023).

In contrast, the HB was built to provide hints on how to solve a question. We used prompt engineering to explicitly forbid it from giving final solutions and to encourage it to give hints and approaches instead. This induced behavior can be seen as a form of flexible scaffolding, which should be advantageous to students (Anghileri, 2006).

4. Study

Experienced university-level physics educators carefully curated challenging physic questions for an initial pre-test (eight multiple choice questions) and the main test (two free-text questions). They drew inspiration from a previous International Physics Olympiad (Leibniz-IPN, 2023) and standardized tests (Afif et al., 2017; Lichtenberger et al., 2017).

During the study (see Figure 1), we collected interaction protocols (see Section 4) for each condition and analyzed whether the interaction in the experimental conditions indicated increased reflection compared to our baseline conditions.

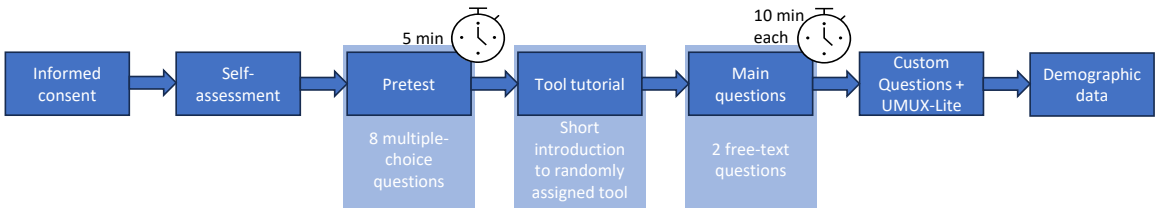

Figure 1: Study procedure depicting timed initial pretest and main questions.

A total of 40 (Age $\bar{x}$=26.2y, s=6.83y, 29m, 10f, 1 not specified) participants fully completed our online study. Table 1 details their allocation to conditions, pretest scores and their self-assessed physics knowledge[1]. Of the participants, 50% studied physics, 27.5% studied non-physics STEM fields, and 22.5% studied something else but confirmed prior physics knowledge.

Table 1: Participant numbers, pretest scores and self-rated physics knowledge per condition.

	# Participants	Pretest score		Physics knowledge	
		$\bar{x}$	s	$\bar{x}$	s
SE	12	4.67	1.61	70.7	17.8
MRB	10	4.6	2.17	61.9	26.6
HB	10	4.1	2.38	66.5	27.0
CB	8	4.62	1.77	60.5	17.9

To measure the **student performance**, we had two physics educators create a rating scheme for the two main questions. Then, two different physicists independently rated the answers given for each question based on that scheme, reaching a consensus through discussion. An additional evaluation of the inter-rater reliability (κ=0.72) over the main questions indicated a substantial reliability (Landis and Koch, 1977).

Furthermore, we analyzed participant **interaction with the support tools** by evaluating the prompts and search queries given to the tools and the interaction protocols of the different bots. By coding how participants interacted with their assigned tool, we were able to extract a number of interaction strategies. The coding was done independently by two researchers, who then came to an agreement after a discussion.

Finally, we analyzed the **custom questions**, for which we asked participants to evaluate the quality and correctness of their given tool and conducted the UMUX-Lite questionnaire (Lewis et al., 2013).

5. Results

Student Performance We awarded up to three/four points for the two main questions, respectively. Across all conditions, the average final score was 1.9. Participants using SE were most successful ($\bar{x}$=2.5, s=1.38), followed by MRB ($\bar{x}$=2.3, s=2.21), CB ($\bar{x}$=1.62, s=1.77) and HB ($\bar{x}$=1.1, s=0.57). We found a significant correlation between the pretest score and main test score ($p < 0.05$, τ=0.31), main test score and self-assessed physics knowledge ($p < 0.01$, τ=0.32) and pretest score and self-assessed physics knowledge ($p < 0.001$,τ=0.42). This correlation shows that our pretest is adequate and that self-assessed physics knowledge is a good indicator of success in answering the main questions. A one-way ANOVA (after rank alignment (Wobbrock et al., 2011)) found no statistically significant differences between the conditions.

Interaction with the Support Tools In total, participants asked 151 queries to the different support tools. 40 of which were asked using SE, 52 using MRB, 44 using HB and

1. Input on a visual scale between 0 and 100.

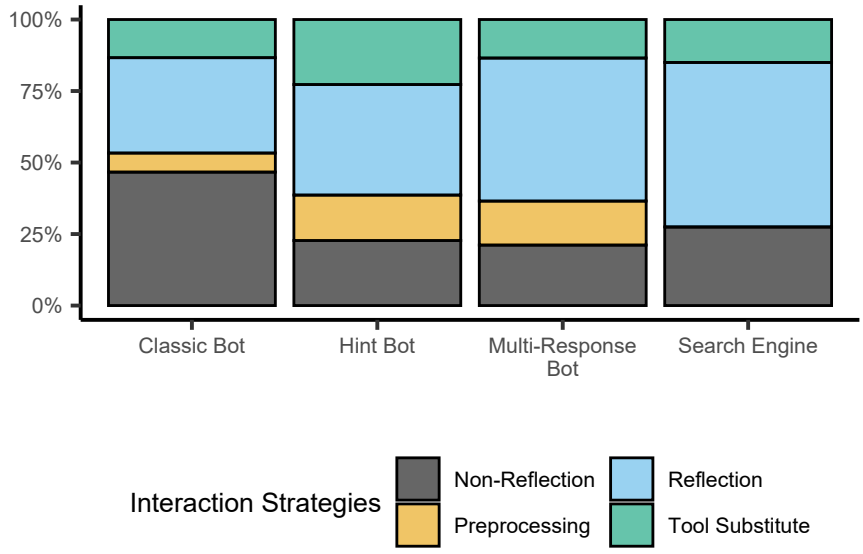

Figure 2: Distribution of interaction strategies for each condition.

15 using CB. Each of those tool interactions was labeled independently by two researchers. Subsequently, the researchers identified four overarching interaction strategies through discussions. Each strategy encompasses a number of different interaction types that fall under this strategy. A distribution of strategies per condition is illustrated in Figure 2.

Non-reflection encompasses interaction types that exhibit no active task reflection of the students, such as using copy & paste (including partial questions), or trying to locate the question in the internet. This strategy is employed to a comparable degree for SE (28%), MRB (21%) and HB (23%) and nearly twice as often when using CB (47%).

Preprocessing includes all interaction types in which participants used priming, tried to change the bot behavior, or reformulated the question. These strategies were mostly used for HB (16%) and MRB (15%) and less for CB (7%). They were not used for SE (0%).

Reflection includes all interaction types that require some level of task reflection by the students: conceptualizing a question to ask for a related formula, prompting the bot for an explanation, as well as correcting the bot after the answer was given. We observed high values for SE (58%) and MRB (50%) and lower reflection rates for HB (39%) and CB (33%).

Finally, the **tool substitute** interaction strategy include interaction types indicating uses as a calculator, translation tool, or to rearrange a formula. It was most often observed for HB (23%), followed by SE (15%), MRB (14%) and CB (13%).

Custom Questions We used the UMUX-Lite questionnaire (Lewis et al., 2013) and calculated the SUS score (Brooke et al., 1996) from it to obtain the usability score for each condition. We recorded the highest score for SE ($\bar{x}$=73.0, s=13.1), indicating good usability. The lowest was $\bar{x}$=47.3 (s=17.4) for MRB, indicating poor usability (Jeff Sauro, 2019). Using a one-way ANOVA (after confirming normality), we found a significant main

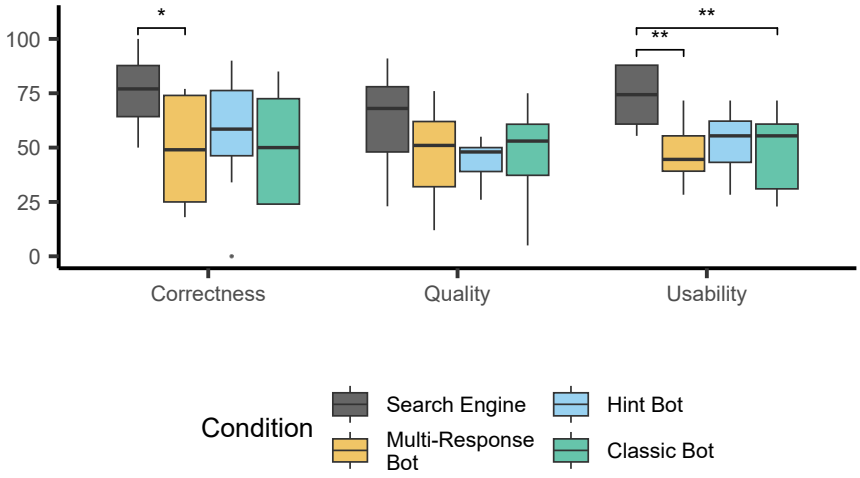

Figure 3: Rated correctness, quality and usability for each condition.

effect for the type of support tool ($F(3, 36) = 6.7, p < 0.01$). Post-hoc pairwise comparisons (tukey-adjusted p values) revealed significant differences between the SE and the MRB as well as the SE and the CB, respectively. No further significant differences were found.

Additionally, students rated the system for correctness (see Figure 3), where we found a significant main effect for the type of support tool ($F(3, 36) = 3.8, p < 0.05$). Post-hoc pairwise comparisons (tukey-adjusted p values) revealed a significant difference between the SE ($\bar{x}$=77.2, s=16.5) and the MRB ($\bar{x}$=48.6, s=21.3).

Looking at the answer quality (see Figure 3) of the systems, no significant differences were found, with the average reported quality decreasing from the SE condition ($\bar{x}$=62.1, s=23.7) to the HB condition ($\bar{x}$=43.1, s=9.7).

6. Discussion

Our results show that student performance when using the Multi-Response Bot (MRB) came closest to the results achieved by students when using the Search Engine (SE), achieving a mean final score of 2.3 compared to the 2.5 of SE. We believe this to be the case due to the multiple answers provided by the chatbot. This behavior shows the participants that there is not one perfect answer and forces them to think critically about each of the given responses to evaluate which one is best (Bitzenbauer, 2023). The same pattern is visible when looking at the ratio of reflection for both conditions (see Figure 2). The fact that participants using those conditions were forced to make a decision on which answer to take led to a positive impact on their critical thinking. This would further indicate that LLMs when used the right way, can exhibit a similar positive influence on students' task reflection compared to using a search engine.

6.1. Tool-Specific Interaction Strategies - And Why They Matter

For the Hint Bot (HB) condition, we found that multiple people tried to actively change the bot's behavior, forcing it to deliver actual answers; something it should actively avoid doing. This is an indicator of frustration with the system since students tried to actively prime it to get their way. To decrease frustration when interacting with a chatbot, we believe having the chatbot answer the question is essential. This might be implemented using a button to toggle different behaviors (give hints, give answers) that students can use when required or by automatically detecting if a question that requires an answer was asked or not.

For the Multi-Response Bot (MRB), we noticed that a lot of questions were asked multiple times. We suspect that this strategy has to do with the participants realizing that they could get three more answers to the question they asked, which might be seen as advantageous to answering the main questions. To increase the answer diversity using different LLMs in the backend would make sense. Furthermore, using prompt engineering, different behaviors could be leveraged for each answer. For example, one of the three multi-response answers could be given by the Hint Bot (HB).

6.2. Interface Paradigms Dictate Usage Patterns

Users employed a more systematic approach when given the Search Engine (SE). We believe this behavior originates from the inherent nature of the search engine interface. The familiarity with this interface allows users to extract information with higher precision but requires initial reflection to formulate an appropriate search query. This behavior highlights an important design component for education support tools. Likewise, we believe that teaching users how to interact with LLMs, thus familiarizing them with the intricacies of such models, would enable them to develop their own informed strategies as they recognize their benefits.

7. Conclusion

In this work, we have shown two possible LLM moderation approaches: giving multiple responses and hints. We compared them to using a search engine or an unmoderated LLM in the context of supporting students to answer physics questions. We found that LLM moderation can improve the students' interaction behavior and amount of critical thinking compared to the unmoderated LLM and can potentially be a valuable approach to using LLMs in education. However, their usage is still subject to design constraints, such as poor usability of chatbot-based LLMs. Tandem solutions can overcome these weaknesses of current LLM interactions, leveraging the individual strengths of our moderation methods.

Our findings help guide the current debate on LLMs and their usage in education, highlighting ways to design effective AI-supported educational methods, leveraging their benefits while limiting negative repercussions.

Acknowledgments

This research is supported by the European Union's Horizon 2020 Programme under ERCEA grant no. 952026 HumanE-AI-Net and by the German Federal Ministry of Education and Research (BMBF) through the project KI4TUK.

References

Nur Faadhilah Afif, Muhammad Gina Nugraha, and Achmad Samsudin. Developing energy and momentum conceptual survey (emcs) with four-tier diagnostic test items. In *AIP Conference Proceedings*, volume 1848. AIP Publishing, 2017.

Julia Anghileri. Scaffolding practices that enhance mathematics learning. *Journal of Mathematics Teacher Education*, 9:33–52, 2006.

Philipp Bitzenbauer. Chatgpt in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, 15(3):ep430, 2023.

BR. ChatGPT: So gut hat die ki das bayerische abitur bestanden https://www.br.de/nachrichten/netzwelt/chatgpt-ki-besteht-bayerisches-abitur-mit-bravour,TfB3QBw, 2023.

John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

Bor Gregorcic and Ann-Marie Pendrill. Chatgpt and the frustrated socrates. *Physics Education*, 58(3):035021, 2023.

Jeff Sauro. 5 ways to interpret a SUS score https://measuringu.com/interpret-sus-score/, 2019.

Fabian Kieser, Peter Wulff, Jochen Kuhn, and Stefan Küchemann. Educational data augmentation in physics education research using chatgpt. *arXiv preprint arXiv:2307.14475*, 2023.

Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. Unreflected acceptance – investigating the negative consequences of chatgpt-assisted problem solving in physics education, 2023.

Stefan Küchemann, Steffen Steinert, Natalia Revenga, Matthias Schweinberger, Yavuz Dinc, Karina E Avila, and Jochen Kuhn. Physics task development of prospective physics teachers using chatgpt. *arXiv preprint arXiv:2304.10014*, 2023.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Leibniz-IPN. ScienceOlympiaden https://www.scienceolympiaden.de/, 2023.

Amanda Lenhart, Maya Simon, and Mike Graziano. The internet and education: Findings of the pew internet & american life project. 2001.

James R Lewis, Brian S Utesch, and Deborah E Maher. Umux-lite: when there's no time for the sus. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2099–2102, 2013.

Andreas Lichtenberger, Clemens Wagner, Sarah I Hofer, Elsbeth Stern, and Andreas Vaterlaus. Validation and structural analysis of the kinematics concept test. *Physical Review Physics Education Research*, 13(1):010115, 2017.

Vatsal Raina and Mark Gales. Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*, 2022.

Renato P dos Santos. Enhancing physics learning with chatgpt, bing chat, and bard as agents-to-think-with: A comparative case study. *arXiv preprint arXiv:2306.00724*, 2023.

University of Cambridge. ChatGPT (we need to talk) https://www.cam.ac.uk/stories/ChatGPT-and-education, 2023.

Colin G West. Advances in apparent conceptual physics reasoning in chatgpt-4. *arXiv preprint arXiv:2303.17012*, 2023a.

Colin G West. Ai and the fci: Can chatgpt project an understanding of introductory physics? *arXiv preprint arXiv:2303.01067*, 2023b.

Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 143–146, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1978963.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*, 2022.