HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240195

Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education

Lars KRUPP [a,b,1], Steffen STEINERT [c] Maximilian KIEFER-EMMANOUILIDIS [a,b] Karina E. AVILA [b] Paul LUKOWICZ [a,b] Jochen KUHN [c] Stefan KÜCHEMANN [c] and Jakob KAROLUS [a,b]

[a] *German Research Center for Artificial Intelligence (DFKI)*
[b] *RPTU Kaiserslautern-Landau*
[c] *LMU Munich*

Abstract. The general availability of large language models and thus unrestricted usage in sensitive areas of everyday life, such as education, remains a major debate. We argue that employing generative artificial intelligence (AI) tools warrants informed usage and examined their impact on problem solving strategies in higher education. In a study, students with a background in physics were assigned to solve physics exercises, with one group having access to an internet search engine (N=12) and the other group being allowed unrestricted use of ChatGPT (N=27). We evaluated their performance, strategies, and interaction with the provided tools. Our results showed that nearly half of the solutions provided with the support of ChatGPT were mistakenly assumed to be correct by students, indicating that they overly trusted ChatGPT even in their field of expertise. Likewise, in 42% of cases, students used copy & paste to query ChatGPT — an approach only used in 4% of search engine queries — highlighting the stark differences in interaction behavior between the groups and indicating limited task reflection when using ChatGPT. In our work, we demonstrated a need to (1) guide students on how to interact with LLMs and (2) create awareness of potential shortcomings for users.

Keywords. ChatGPT, Large Language Models, Education, Physics

1. Introduction

LLMs have been omnipresent in media and the public eye since November 2022 when ChatGPT was first presented [1]. With one of the fastest growing user bases ever measured for any application [2,3], it is difficult to estimate its future impact on every aspect of our daily lives.

Especially in sensitive areas, such as education, easily accessible information — true or false — poses challenges for educators and students alike. Recent discussions

[1] Corresponding Author: Lars Krupp, lars.krupp@dfki.de

around ChatGPT often involve its use as an AI support tool in assignments, for homework and in the classroom. Despite research advances, it is still unclear how LLMs, such as ChatGPT, can meaningfully support students in educational contexts [4,5]. To properly design methods that allow for informed usage of these systems, we need to investigate how the users — in our case students — interact with those AI tools and how their usage influences the students' decision making.

LLMs are predictive models that predict the most probable next token based on a series of previously seen tokens they have already seen. As a result, they excel at tasks such as brainstorming [6], writing [7], translation [8], and even programming [9]. Contrarily, disciplines that rely heavily on calculations and reasoning prove more challenging for LLMs [10]. Potentially leading to unforeseen or even negative consequences for students, like incorrect homework [4] or learning an incorrect explanation of a concept. Yet, it remains unclear how interacting with LLMs may give rise to students' misconceptions. Consequently, identifying disparities in students' decision making when using AI tools is essential to understand potential negative consequences.

In our work, we examined the field of physics, specifically how students with a strong background in physics interact with ChatGPT to assist them in solving physics questions. We conducted a study with a total of 39 participants with backgrounds in science, technology, engineering and math (STEM) fields from multiple universities. The experimental group (N=27) had unrestricted access to ChatGPT, while the control group (N=12) had access to a search engine only.

Our findings indicate that participants with the ChatGPT condition overly trusted answers generated by ChatGPT. In particular, students often failed to recognize wrong answers given by ChatGPT and largely relied on a copy & paste strategy to solve the posed physics questions. In contrast, participants in the Search Engine condition showed higher rates of reflection, as indicated by their sparse use of copy & paste, favoring more thought-through solving strategies.

Our work highlights that there are stark disparities in the interaction behavior between the student groups, provoked by the accessibility of ChatGPT for the experimental group. Even students with advanced domain knowledge struggled to differentiate between correct and incorrect answers given by the LLM and could not use the system effectively. Consequently, there is a need for further research to design AI-based support tools in a way that (1) creates awareness of their inherent uncertainty and (2) allows moderated use that encourages critical thinking.

2. Related Work

The field of language models (LM) offers a variety of possible applications in education. For example, they have been used for multiple-choice question generation [11] or answering [12]. However, since we have to expect students to use LLMs like ChatGPT at home, there is a need to figure out how they utilize these powerful new tools unaided.

Recent advances in natural language processing, initiated by the introduction of the transformer architecture [13], have led to significant progress in the field of language models. The different approaches taken by GPT [14] and BERT [15] models proved to be exceptionally successful. Progress has been steady, with a trend towards increasingly larger models, supported by their scaling laws [16], which suggest that larger size generally leads to a better model. ChatGPT [1] brought the technology into the public eye, further accelerating the pace of publications and leading to the development of models such as LLaMA [17], GPT-4 [18], and PaLM-E [19]. Some of which even support multi-modal inputs [19]. Language models have shown their potential in many different areas [7,8,9] and are a topic that also influences education [20].

LLMs offer great potential for advancing standard practices and research in education [20]. Several possible applications have been previously suggested, such as personalized learning, lesson planning, assessment and evaluation, to familiarize students with challenges and opportunities of LLMs [20]. Furthermore, a number of studies exist that investigate the use of chatbots based on different technologies in education [21]. The use of chatbots in education offers several advantages, such as serving as a pedagogical tool to help students with disabilities and to help different social groups to close the educational gap that may exist between them [22]. However, none of the systems examined in these works are based on a LLM despite several authors seeing great potential for LLM-based chatbots in the educational domain [23]. It should be noted that LLMs show some weaknesses. Until now, they lack higher-order thinking skills, and their outputs strongly depend on the data they have been trained on, sometimes leading to unreliable outputs [24].

In physics education, there are conflicting reports on the ability of LLMs to solve physics tasks. On the one hand, a few studies have observed inconsistent behavior in ChatGPT's answers to physics questions [25,26]. These studies showed that ChatGPT often provides incorrect answers to physics questions and concluded that it is unsuitable as a physics tutor or for cheating on homework. Bitzenbauer used this apparent weakness of ChatGPT to foster students' critical thinking skills by having them generate answers to a question and discuss them critically, leading to an improved perceived usefulness of ChatGPT [24]. On the other hand, other studies demonstrate the strength of ChatGPT 3.5 and 4.0 to solve conceptual multiple-choice questions in physics [27,28]. ChatGPT was able to solve 28 out of 30 items of the force concept inventory correctly [28]. Kieser et al. even found that ChatGPT 4.0 is able to mimic different students' difficulties when answering conceptual questions, which opens the opportunity for data augmentation, personalized support for students that is sensitive to different difficulties, and support for teachers during task creation [29]. The latter opportunity was studied by Küchemann et al., who compared the characteristics and quality of created physics tasks by prospective physics teachers either using ChatGPT or a textbook. Their findings indicate, that participants who used ChatGPT embedded the tasks less frequently in a real-world context and that most ChatGPT generated tasks were used without modification [30]. These findings point towards the affordances of using ChatGPT in education and the overreliance on AI [31] of participants when using it.

While these articles provide interesting findings and show that using ChatGPT for answering questions present great demands on students, the results were either not verified with real students [25,26] or the problem solving strategies when using ChatGPT were not studied [24,30].

3. Methodology

The related work highlights the existing uncertainties regarding the use of LLMs in general and specifically in the context of physics education. However, to date, little work has been conducted that allows for moderated and informed usage of such models. We argue that informed usage of generative models is crucial, particularly in educational areas. Our work contributes an investigation specifically into how students interact with LLMs and whether they are aware of their shortcomings. In a mixed-method evaluation conducted online and at two universities (RPTU Kaiserslautern-Landau and LMU Munich), we tasked students with solving given physics problems. Using a between-subject study design, we assessed students' performance and interaction strategies when having access to different support tools.

As a baseline condition, we had students use an internet search engine (SEARCH ENGINE). This setup represent the de facto standard prior to the advent of LLMs [32,33]. In the CHATGPT condition, students were able to freely use ChatGPT. We recorded the students' physics knowledge with a pretest (no support tools allowed) and their performance in the main test, as well as inquired about their impressions when interacting with ChatGPT through questionnaires and an exit interview (see Figure 1). Our research was guided by two main research questions:

RQ1: *What is the performance of students when being allowed to use ChatGPT in comparison to the students who used a search engine?* One main inquiry of our work focused on whether ChatGPT allowed students to perform better when solving the physics questions. We further analyzed the students' interaction protocols with both tools (SEARCH ENGINE, CHATGPT) to investigate how effectively they used the tool.

RQ2: *What are predominant strategies when interacting with ChatGPT compared to search engines?* On a meta-level, we were interested in what solving strategies students employed when using ChatGPT and how they differed from the ones used with search engines. From the conducted exit interviews, in combination with the students' interaction protocols, we distilled predominant strategies when interacting with either tool.

3.1. Physics Question Acquisition

For our **main test**, we selected four physics questions. To fulfill the requirement that all questions are solvable with school knowledge, we chose questions that require knowledge of six topics of physics taught in school. By choosing four tasks from the International Physics Olympiad [34], a high school competition, we ensured that the tasks were suitable, yet challenging, for university students. All tasks were reviewed by two physics university educators and considered adequate in terms of difficulty and time required for university students. The task texts were adapted in such a way that no picture is necessary for the solution and it was verified that ChatGPT cannot solve the tasks directly and the search engine does not show a page containing the solutions, but both can give hints for obtaining the solution.

We then designed a **pretest** containing the previously the selected topics. For this, we acquired items from multiple sources [35,36,37] and created our own questions, one of which was inspired by [38].

3.2. Procedure

The study itself was split into multiple parts, as shown in Figure 1. After providing informed consent and an in-depth explanation of the study procedure, the study started

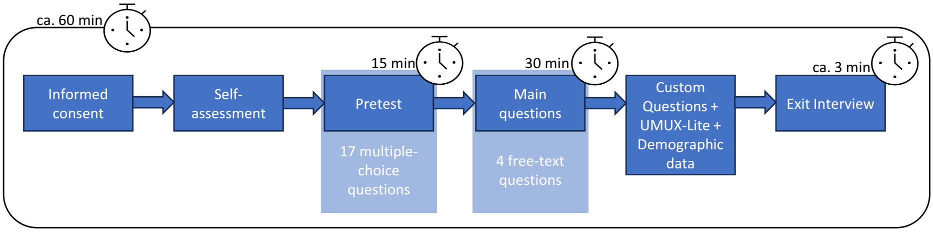


Figure 1. The study procedure timeline in detail. First a self-assessment was conducted, followed by a short pretest and finally the main test, where use of the support tool was allowed. Afterwards followed a questionnaire and depending on the condition an exit interview.

with a self-assessment where participants could rate their physics and ChatGPT knowledge and how often they use ChatGPT (see Section 3.3). Following that, participants had 15 minutes to solve the 17 multiple-choice pretest questions worth 1 point each (max points = 17). Afterwards, participants were allowed to use a modified user interface of ChatGPT or a search engine to help them solve the four physics questions (max points = 12) given a time frame of 30 minutes. The written part of the survey ended with a short questionnaire, including the affinity for technology interaction scale [39] to assess participants' views on technology, the UMUX-Lite scale [40] to assess usability, and custom questions on perceived accuracy and quality of the tools' answers, as well as demographics. Throughout the course of the survey, the order of all questions remained unchanged, ensuring the same experience for all participants. For participants attending in person at the university of Kaiserslautern-Landau (N=20), we additionally recorded a short (2-5min) exit interview. After the study, participants were reimbursed with the equivalent of $11 or course credit for a voluntary seminar (N=7). Ethical approval for this study was obtained from the Ethics Committee at the German Research Center for Artificial Intelligence (DFKI).

3.3. Participants

For our baseline condition (SEARCH ENGINE), we acquired 13 participants by providing them the option to do the survey online using university mailing lists from both universities. Of these, 12 participants (Age $\bar{x}$=23.6 y, s=2.6 y; 10m, 2n/a; 3 in person, 9 online) fully completed the survey. The students (5 physics, 7 non-physics STEM) were on average in their eighth semester ($\bar{x}$=7.4, s=4.3), scored eight points in the pretest ($\bar{x}$=8.2, s=3.9, max=12), had an above-average self-reported physics knowledge ($\bar{x}$=62.8, s=25.1) coupled with below average experience when using ChatGPT ($\bar{x}$=40.5, s=33.4)[2].

For the second condition of our study (CHATGPT), we initially recruited 30 participants from two different universities (RPTU Kaiserslautern-Landau, LMU Munich). with a background in physics. They were recruited using mailing lists, posters, and by advertising the study in lectures. Of these 27 participants (Age $\bar{x}$=22.6 y, s=4.0 y, 25m, 2f, 27 in person, 0 online) fully completed the survey. Participants were, on average, in their sixth semester ($\bar{x}$= 5.3, s=3.3). Participants (17 physics, 10 non-physics STEM) scored on average nine points in the pretest ($\bar{x}$=9.2, s=3.2, max=15). Using ANOVA, we found a statisti-

[2] Self-assessed physics knowledge and experience using ChatGPT were input on a visual analog scale between 0 and 100.

cally significant difference for the pretest score between physics and non-physics STEM students ($F(1,37)$=11.8, $p <$.002,η^2=.24)[3] but not the CHATGPT and SEARCH ENGINE conditions. Further, students reported an above-average perceived physics knowledge ($\bar{x}$=58.7, s=18.6) and below average experience with ChatGPT ($\bar{x}$=42.2, s=24.5).

3.4. Apparatus

For the CHATGPT condition, we used ChatGPT 3.5 turbo with client side modifications using JavaScript, including a rating scale (good, neutral, bad) to appear with each answer provided by ChatGPT to allow participants to directly voice their opinions. Furthermore, we implemented a download button to be able to save the conversation[4].

For the SEARCH ENGINE condition, we set up a website through which participants could use Google while we were able to collect their search queries.

When students participated in person, we further recorded the ChatGPT conversation log, the participants' ratings and conducted an exit interview. Additionally, all participants were allowed to use a non-programmable calculator, pen, and paper throughout the study.

3.5. Measures

To allow for a holistic picture of how students interact with CHATGPT, we measured student performance through different factors, conducted exit interviews, and analyzed the full student interaction protocols as described in the following section.

Student Performance To evaluate participant performance, a grading schema was created by two physics university educators. Using this schema, two other physicists scored the given answers for the four main questions, independently from each other, awarding between zero and three points per question and participant. We evaluated the inter-rater reliability by calculating the average Cohen's Kappa (κ=0.72) over all main questions, which indicates a substantial reliability [41]. Through discussion both raters reached an agreement in cases where their initial rating differed. The resulting final scores show student performance in answering the main questions. Further, we determined how participants reached their final answers, indicating their problem solving strategy. If the final result of a question was present in the interaction protocol with ChatGPT related to that question, we assigned "extracted from ChatGPT" as strategy. Otherwise, it was counted as "own answer". Questions that were not answered were counted as "none". When it was not evident how the answer was obtained, we assigned "random guess" as strategy.

Interaction with the support tools We analyzed the interaction of the participants with their respective tool (ChatGPT or search engine). For the CHATGPT condition, this includes all prompts from participants, respective answers from ChatGPT and associated ratings from participants. Furthermore, for the SEARCH ENGINE search queries were analyzed.

[3] Effect sizes are given using η^2 (Partial Eta Squared): small (> .01), medium (> .06), large (> .14).

[4] At the time of the study, this feature was not yet available.

Perceived Correctness of ChatGPT Answers Having two physicists additionally rate all answers given by ChatGPT for correctness enabled us to compare how students rate answers and their actual correctness. With this information, we were able to calculate the false positive rate (FPR), i.e. positively voted answers that are incorrect, and the true positive rate (TPR), i.e. positively rated answers that are correct. In our analysis, we focus on these metrics as they highlight how often information from ChatGPT was assumed to be correct.

Interaction Types Additionally, we created codes to represent the strategies with which participants created their prompts by categorizing each individual prompt into a coding, comparing and merging them as needed until a consistent representation emerged.

Custom questions As mentioned in Section 3.2, we administered the ATI [39] and UMUX-Lite [40] as well as two custom questions to inquire about the participants' impression on ChatGPT correctness accuracy and quality.

Exit Interviews We conducted exit interviews with 20 participants that were assigned the CHATGPT to further examine qualitative aspects of their interaction. Questions during the interview included asking what strategies were used, how the tool was used and how confident participants were in the correctness of their results.

4. Results

4.1. Student Performance

On average, participants scored $\bar{x}$=1.04 points (s=1.43) out of the maximum achievable 12 points in the CHATGPT condition. Most points (nearly 90%) were achieved in questions Q1 and Q3. We found a large positive correlation between the final score and the pretest score, using Kendall's rank correlation (τ=.37, p=.02). No further correlations with respect to the final score were found, in particular for the self-assessed physics knowledge, and study program related demographics such as study subject and semester.

Analyzing how final answers were obtained, we observed that the most prominent strategy was "extracted from ChatGPT" being used in 62% of all cases. Following this, 28% of participants arrived at their "own answer", 9% of questions were not answered ("none") and 1% made a "random guess".

For the SEARCH ENGINE, participants scored $\bar{x}$=1.83 points (s=1.27) on average. Here too, most points (around 95%) were achieved in questions Q1 and Q3. Using Kendall's rank correlation, we found a statistically significant medium positive correlation (τ=.27, p=.03) between the main test score and self-assessed physics knowledge, but none for the pretest score.

Further, we conducted a one-way ANOVA after rank-aligning the data [42] to investigate whether there are significant differences between our two conditions (SEARCH ENGINE, CHATGPT) with regard to the students' performances in the main test. We found that students in the SEARCH ENGINE condition performed significantly better (F(1,37)=5.5, p=.02, η^2=.13)[3]. Progress in the study program (semester number), course of study (physics, non-physics STEM) and self-rated physics knowledge did not impact the final score as confirmed by ANCOVA tests.

4.2. Interactions with ChatGPT

In total, participants working with ChatGPT created 272 prompts, 165 of which were rated (see Section 3.4). Overall, participants rated 47% of ChatGPT answers positively, indicating that they deemed them to be correct. 29% were rated neutral, and 24% negative, indicating that participants were unsatisfied with them.

Contrarily, our expert physicists only rated approximately 22% as correct, highlighting a mismatch in expectations. This effect is visible throughout all main questions, as depicted in Figure 2. To further analyze intersections in believes of students and experts, we looked at perceived correctness (see Section 3.5). We obtained a high false-positive rate of 57%, i.e., over half of all the answers provided by ChatGPT were believed to be correct by participants but rated incorrect by experts. The true-positive rate of 91%, however, indicates that participants rated most correct answers positive.

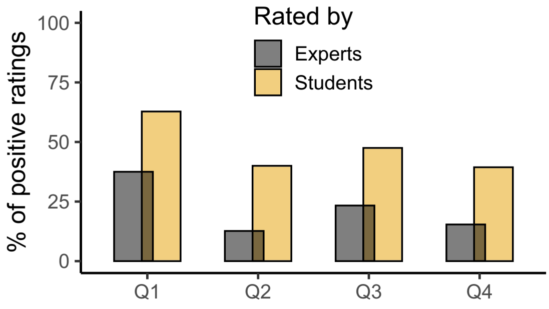


Figure 2. The proportion of positively rated ChatGPT answers to students' prompts visualized for each of the main questions and broken down for students and experts.

Interaction Types We identified four main interaction types based on the reviewed ChatGPT interaction logs from all participants: *copy & paste*, *preprocessing*, *postprocessing*, and *transformation*. The individual interactions are described in more detail below.

Copy & Paste is the most prominent interaction type, where participants transferred the physics question directly to ChatGPT without any changes.

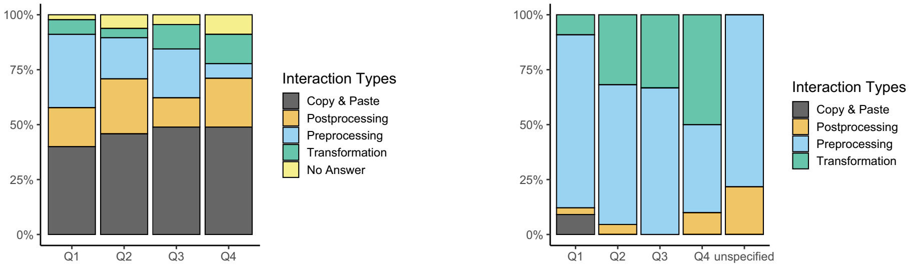
Preprocessing is characterized by students trying to reduce the question complexity and using simple priming strategies. They divide a question into multiple parts (P10), ask for formulas (P4), or try to prime the model to improve their results when asking physics questions (P14).

Postprocessing builds on already existing answers given by ChatGPT. The participants try to obtain explanations for parts of a question (P1) or correct mistakes they found in the given answer (P12).

Transformation is an interaction type where students used ChatGPT to apply some kind of *transformation* on the data, including translation into another language (P6) and summarizing results (P3).

Interaction Strategies During the study, we noticed that students built their individual strategies to solve the given physics questions based on these interaction types. For example, a participant might start with priming ChatGPT (*preprocessing*), followed by *copy & pasting* the question and, ultimately, asking for an explanation of some part of the answer (*post-processing*).

Overall, *copy & paste* was the most used interaction strategy, being used 84 times. *Preprocessing*, the next most common strategy, was used 37 times, followed by *post-processing* (36) and *transformation* (16). In Figure 3a, the distribution of used interaction strategies per question is visualized.



(a) Distribution of interaction types for each question of the main test for the CHATGPT condition.

(b) Distribution of interaction types for each question of the main test for the SEARCH ENGINE condition.

Figure 3. Distribution of interaction types per question for both conditions. Interaction types that could not be assigned to a question since their content did not include any identifying markers were put into the unspecified category.

4.3. *Interactions with the Search Engine*

To be able to compare how participants of both groups interacted with their respective tool, we describe interaction types and strategies when using the search engine here. The distribution of used interaction strategies is visualized in Figure 3b.

Interaction Types We divided the interactions done with the search engine into the same four types as the interactions with ChatGPT (Section 4.2). This allows for easy comparison between the two conditions. There are some minor updates to the interaction types, as some interactions seen when using a search engine were not present when using ChatGPT. *Preprocessing* for the SEARCH ENGINE condition mainly consists of asking for formulas and calculations, while *postprocessing* only encompasses asking for explanations. In the *transformation* interaction, the interaction types “finding answers to related problems” and “trying to find the initial question using keyword search” were added. There were no changes to the *copy & paste* interaction.

Interaction Strategies The relations between how often different strategies were used changed considerably from the CHATGPT condition to the SEARCH ENGINE. Here, the most used strategy was *preprocessing* with 64 uses, followed by *transformation* with 17 uses, *postprocessing* (8) and *copy & paste* (3).

4.4. Custom Questions

We calculated the average ATI [39] score of all study participants ($\bar{x}$=4.35, s=0.79) showing above-average technical affinity allowing them to adequately interact with the given tools. Additionally, we used the UMUX-Lite [40] questionnaire to calculate a parity score for SUS [43] for the CHATGPT ($\bar{x}$=73.05, s=9.95) and the SEARCH ENGINE condition ($\bar{x}$=66.23, s=11.62). Both indicate an above-average system usability. Further, participants rated ChatGPT answers for correctness at $\bar{x}$=58.0 (s=18.59) and their quality at $\bar{x}$=69.26 (s=16.21) on a visual analog scale from 0 to 100. The search engine answer correctness was rated $\bar{x}$=59.6 (s=22.8) and its answer quality $\bar{x}$=55.5 (s=28.7). We found no significant differences between the two conditions for all custom questions.

4.5. Exit Interview

We recorded the audio of the CHATGPT exit interviews (59:30 min) and transcribed them using Whisper [44]. To analyze the exit interviews, we used the approach by Blandford et al. [45]. Two researchers coded all interviews separately and merged a final coding tree. From a final discussion, the following themes surfaced: STRATEGIES, INTERACTION, and REFLECTION as presented in detail below.

Strategies While a diverse set of strategies was employed by the participants, most of them mentioned copy & pasting a question in their exit interview. Different reasons for this were given, such as wanting to see how ChatGPT would deal with the question (P4) or that they did not know how to address the physics question (P7). Other strategies included using ChatGPT like a search engine, e.g., asking for formulas (P1) as it was more convenient. Some strategies indicated a higher level of reflection, such as prior physics problem conceptualization and asking targeted questions (P10). Similarly, ChatGPT was used to explore options for possible solutions and approaches. Here, students identified valuable pieces in ChatGPT answers and showed the ability to detect mistakes and inconsistencies in its argumentation (P4). Though, participants also stated that they had to compromise between speed and correctness of their solutions due to the time constraints. While motivated initially, they tried to offload more work to ChatGPT if time was running out (P1).

Interaction When interacting with ChatGPT, participants identified a need to use informed queries. Some tried to achieve this by extracting the most relevant parts of a question from it (P3). Others found that longer texts worked poorly, implying a need for concrete queries to work around this issue (P20) or requiring participants to dig deeper into an answer given by ChatGPT (P12). Interestingly, some participants described their interaction/conversation with ChatGPT as human-like, that the answers looked nice and were very well elaborated (P5). However, selected participants feared that this could delude unaware users (P20).

Reflection A number of participants were aware that it is important to reflect on the answers given by ChatGPT, rigorously reviewing them for correctness (P20) and identifying mistakes made by ChatGPT (P1). Especially participants with background knowledge about LLMs were aware of ChatGPT's weaknesses with regard to physics content and knew what to look out for (P18). Contrarily, for most physics questions, participants showed no sign of actively engaging with the exercises, limiting their reflection (P14).

5. Discussion

Our study provides concrete evidence that students demonstrated vastly different problem solving strategies when having access to ChatGPT and heavily relied on its answers, even struggling to determine their validity. In the following section, we elaborate on these findings and highlight open research questions for the responsible use of LLMs in education.

5.1. Overreliance on ChatGPT Answers Leads to Low Scores

Scored student performance **(RQ1)** was worse than initially expected (Section 4.1) given our curated selection of exercises. Students using the CHATGPT condition performed significantly worse compared to students in the SEARCH ENGINE condition. Moreover, our study revealed that students in the CHATGPT condition had difficulties detecting if answers generated by ChatGPT were correct or not, as indicated by the high false positive rate of 57%. The unreflected acceptance of presented answers is worrying as it might lead from singular misinformation to general misconception and showcases that **there is a definite need to research interactive mechanisms to increase awareness of the uncertainty of LLMs**. Contrarily, most search engines are less likely to suffer from this drawback, as presented results are not formulated as definite answers, a design aspect that could potentially inform the design of future interfaces for LLMs.

5.2. Copy & Paste Is The Most Prominent Strategy for ChatGPT Users

This overreliance also manifested when analyzing the employed interaction strategies for the two different tools (Sections 4.2 and 4.3). In the CHATGPT condition, 42% of search prompts are based on *copy & paste*, highlighting the limited reflection during problem solving. We did observe some participants testing out informed strategies like priming, reducing the question complexity, or correcting ChatGPT **(RQ2)**. However, the low ratio of these informed strategies shows **the necessity of teaching users how to use LLMs effectively**, allowing them to write prompts to achieve accurate results. The novelty — and thus unfamiliarity — of the interface often enticed users to use the most convenient option (*copy & paste*) available **(RQ2)**.

5.3. Limitations

Overall, we expected students to score better given the careful curation of our exercises through physics education researchers (see Section 3.1). In hindsight, our questions might have been too difficult for a realistic assessment of how students interact with ChatGPT. However, this result also shows that proper training on how to use LLMs such as ChatGPT might be necessary to achieve good results.

The number of total participants that took part in our study was relatively small. To alleviate this, we made the SEARCH ENGINE condition of the survey available online as well, allowing us to gather more participants. However, due to the online environment, it is possible that the answer quality was lower compared to in person participants. Though, if that were the case, we can assume that the difference between the two conditions would have been even more prominent.

While our recorded results are limited to one specific ChatGPT version (3.5) available at the time of writing, we believe that the implications of this work hold true for future iterations of ChatGPT. Although improved capabilities of ChatGPT can deliver potentially more correct answers, this does not change the fact that students overly trusted its answers and showed little reflection on their assigned tasks, limited learning effects.

5.4. The Potential of Moderated Use of LLMs

Our analysis revealed a need to think about the design of educational systems that use LLMs. We need to moderate interaction with language models such as ChatGPT in a way that students can profit from the vast abilities of such tools while simultaneously reducing the negative impact it can have on the students' learning progress. Informed use can be achieved by raising awareness of LLM caveats and educate users on how to best use them. Though we argue that, to leverage the full potential of these models, we should strive to achieve moderated use: a usage that allows students to interact with ChatGPT as a guidance teacher or sparing partner to formulate and explore ideas to solve a physics problem. Such a system should **carefully guide students towards the solution, introducing necessary concepts but allowing critical thinking and reflection** while still being enjoyable and effective to use. If we can demonstrate the benefits of moderated LLMs compared to unrestricted LLMs to students in terms of their ability to learn and progress, we can certainly change and evolve the current ways of teaching. A possible way to moderate LLMs would be to change their output behaviour using prompt engineering as we have done in a different work [46].

6. Conclusion

In this work, we analyzed the impact of ChatGPT on problem solving strategies of students. We found that students who used ChatGPT performed significantly worse compared to those using a search engine. Furthermore, stark differences in user interaction manifested, where ChatGPT users mainly relied on copy & pasting questions and answers, while search engine users used more refined strategies such as searching for formulas. This highlights missing reflection and limited critical thinking as two of the main issues when using LLMs in education. To combat this, we — first and foremost — suggest to inform students more adequately of the shortcomings of these models. Though ultimately, we want to converge towards moderated LLMs, specifically designed to support students in a meaningful way by encouraging critical thinking.

Acknowledgements

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

[1] OpenAI. Introducing ChatGPT https://openai.com/blog/chatgpt; 2022.

[2] Reuters. ChatGPT sets record for fastest-growing user base - analyst note https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/; 2023.

[3] Reuters. Meta's Twitter rival Threads surges to 100 million users faster than ChatGPT - analyst note https://www.reuters.com/technology/metas-twitter-rival-threads-hits-100-mln-users-record-five-days-2023-07-10/; 2023.

[4] Yeadon W, Hardy T. The Impact of AI in Physics Education: A Comprehensive Review from GCSE to University Levels. arXiv preprint arXiv:230905163. 2023.

[5] Revell T, Yeadon W, Cahilly-Bretzin G, Clarke I, Manning G, Jones J, et al. ChatGPT versus Human Essayists: An Exploration of the Impact of Artificial Intelligence for Authorship and Academic Integrity in the Humanities. Research Square. 2023.

[6] Salikutluk V, Koert D, Jäkel F. Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems. In: HHAI 2023: Augmenting Human Intellect. IOS Press; 2023. p. 153-67.

[7] Yuan A, Coenen A, Reif E, Ippolito D. Wordcraft: story writing with large language models. In: 27th International Conference on Intelligent User Interfaces; 2022. p. 841-52.

[8] Wang L, Lyu C, Ji T, Zhang Z, Yu D, Shi S, et al. Document-level machine translation with large language models. arXiv preprint arXiv:230402210. 2023.

[9] Kashefi A, Mukerji T. Chatgpt for programming numerical methods. Journal of Machine Learning for Modeling and Computing. 2023.

[10] Yang Z, Ding M, Lv Q, Jiang Z, He Z, Guo Y, et al. Gpt can solve mathematical problems without a calculator. arXiv preprint arXiv:230903241. 2023.

[11] Raina V, Gales M. Multiple-Choice Question Generation: Towards an Automated Assessment Framework. arXiv preprint arXiv:220911830. 2022.

[12] Zhang X, Bosselut A, Yasunaga M, Ren H, Liang P, Manning CD, et al. Greaselm: Graph reasoning enhanced language models for question answering. arXiv preprint arXiv:220108860. 2022.

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[14] Radford A, Narasimhan K, Salimans T, Sutskever I, et al.. Improving language understanding by generative pre-training. OpenAI; 2018.

[15] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[16] Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv:200108361. 2020.

[17] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

[18] OpenAI. GPT-4 technical report. arXiv preprint arXiv:230308774. 2023.

[19] Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:230303378. 2023.

[20] Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274.

[21] Kuhail MA, Alturki N, Alramlawi S, Alhejori K. Interacting with educational chatbots: A systematic review. Education and Information Technologies. 2023;28(1):973-1018.

[22] Pérez JQ, Daradoumis T, Puig JMM. Rediscovering the use of chatbots in education: A systematic literature review. Computer Applications in Engineering Education. 2020;28(6):1549-65.

[23] Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching. 2023;6(1).

[24] Bitzenbauer P. ChatGPT in physics education: A pilot study on easy-to-implement activities. Contemporary Educational Technology. 2023;15(3):ep430.

[25] Gregorcic B, Pendrill AM. ChatGPT and the frustrated Socrates. Physics Education. 2023;58(3):035021.

[26] Santos RPd. Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agents-to-Think-With: A Comparative Case Study. arXiv preprint arXiv:230600724. 2023.

[27] West CG. AI and the FCI: Can ChatGPT project an understanding of introductory physics? arXiv preprint arXiv:230301067. 2023.

[28] West CG. Advances in apparent conceptual physics reasoning in ChatGPT-4. arXiv preprint arXiv:230317012. 2023.

[29] Kieser F, Wulff P, Kuhn J, Küchemann S. Educational data augmentation in physics education research using ChatGPT. arXiv preprint arXiv:230714475. 2023.

[30] Küchemann S, Steinert S, Revenga N, Schweinberger M, Dinc Y, Avila KE, et al. Physics task development of prospective physics teachers using ChatGPT. arXiv preprint arXiv:230410014. 2023.

[31] Gajos KZ, Mamykina L. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In: 27th international conference on intelligent user interfaces; 2022. p. 794-806.

[32] Affum MQ. The effect of internet on students studies: a review; 2022.

[33] Lenhart A, Simon M, Graziano M. The Internet and Education: Findings of the Pew Internet & American Life Project.. ERIC; 2001.

[34] Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik. ScienceOlympiaden https://www.scienceolympiaden.de/; 2023.

[35] Afif NF, Nugraha MG, Samsudin A. Developing energy and momentum conceptual survey (EMCS) with four-tier diagnostic test items. In: AIP Conference Proceedings. vol. 1848. AIP Publishing; 2017. .

[36] Sharma S, Sharma K. Concepts of force and frictional force: the influence of preconceptions on learning across different levels. Physics Education. 2007;42(5):516.

[37] Mashood K, Singh VA. An inventory on rotational kinematics of a particle: unravelling misconceptions and pitfalls in reasoning. European Journal of Physics. 2012;33(5):1301.

[38] Steif PS, Dantzler JA. A statics concept inventory: Development and psychometric analysis. Journal of Engineering Education. 2005;94(4):363-71.

[39] Lezhnina O, Kismihók G. A multi-method psychometric assessment of the affinity for technology interaction (ATI) scale. Computers in Human Behavior Reports. 2020;1:100004.

[40] Lewis JR, Utesch BS, Maher DE. UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI conference on human factors in computing systems; 2013. p. 2099-102.

[41] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.

[42] Wobbrock JO, Findlater L, Gergle D, Higgins JJ. 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. New York, NY, USA: ACM; 2011. p. 143-6.

[43] Brooke J, et al. SUS-A quick and dirty usability scale. Usability evaluation in industry. 1996;189(194):4-7.

[44] OpenAI. Introducing Whisper https://github.com/openai/whisper; 2022.

[45] Blandford A, Furniss D, Makri S. Qualitative HCI Research: Going behind the Scenes. Synthesis lectures on human-centered informatics. 2016;9(1):1-115.

[46] Krupp L, Steinert S, Kiefer-Emmanouilidis M, Avila KE, Lukowicz P, Kuhn J, et al. Challenges and Opportunities of Moderating Usage of Large Language Models in Education. arXiv preprint arXiv:231214969. 2023.