FROM INTUITION TO UNDERSTANDING: USING AI PEERS TO OVERCOME PHYSICS MISCONCEPTIONS

Ruben Weijers[1] **Denton Wu**[2] **Hannah Betts**[3] **Tamara Jacod**[3]

Luke Yuxiang Guan[3] **Vidya Sujaya**[4,5] **Kushal Dev**[3] **Toshali Goel**[3]

William Delooze[2] **Reihaneh Rabbany**[4,5] **Ying Wu**[2]

Jean-François Godbout[4,6] **Kellin Pelrine**[4,5]

[1]Utrech University [2]Duke University [3]Independent
[4]Mila Institute [5]McGill University [6]Université de Montréal
Correspondence: kellin.pelrine@mila.quebec

ABSTRACT

Generative AI has the potential to transform personalization and accessibility of education. However, it raises serious concerns about accuracy and helping students become independent critical thinkers. In this study, we designed a helpful AI "Peer" to help students correct fundamental physics misconceptions related to Newtonian mechanic concepts. In contrast to approaches that seek near-perfect accuracy to create an authoritative AI tutor or teacher, we directly inform students that this AI can answer up to 40% of questions incorrectly. In a randomized controlled trial with 165 students, those who engaged in targeted dialogue with the AI Peer achieved post-test scores that were, on average, 10.5 percentage points higher—with over 20 percentage points higher normalized gain—than a control group that discussed physics history. Qualitative feedback indicated that 91% of the treatment group's AI interactions were rated as helpful. Furthermore, by comparing student performance on pre- and post-test questions about the same concept, along with experts' annotations of the AI interactions, we find initial evidence suggesting the improvement in performance does not depend on the correctness of the AI. With further research, the AI Peer paradigm described here could open new possibilities for how we learn, adapt to, and grow with AI.

1 INTRODUCTION

Students have recently been exposed to the remarkable capabilities of Generative AI (AI) in education (AIED). For example, OpenAI's ChatGPT has been reported to successfully support teaching preparation, assessment design and grading, and student learning (Lo, 2023). Systems like ChatGPT show potential to save time and enhance teaching and learning, including critical and higher-order thinking tasks (Lo, 2023).

However, there is limited concrete evidence whether these tools are effective at improving student learning outcomes (Samson R, 2025). In fact, LLMs are well-reported to hallucinate information and provide sycophantic answers (Wei et al., 2023; Perez et al., 2023), suggesting that AI could actually be detrimental to student learning if it introduces inaccuracies and biases into classroom materials. Given that an increasingly large number of students are now using AI to help them with their school assignments (Foundation, 2024), there is a growing concern that students must be taught 'AI literacy' to recognize and evaluate potential errors generated by LLMs (Wineburg & Ziv, 2024). Although recent techniques such as Retrieval-Augmented Generation (Piktus et al., 2020) have improved the veracity of generated content, experts disagree on whether hallucinations will be

consistently preventable even in the future (Samson R, 2025)—as they are clearly not today. Thus, there is an urgent need to determine whether AIED can be an effective tool for education despite these inherent limitations.

This research evaluates the potential of LLMs to support student learning through text-based conversations in an introductory university physics class. The study explores how LLMs can be used to facilitate learning of physics concepts, positioning the AIED as a non-expert peer rather than an expert teacher. Students complete a modified Force Concept Inventory (FCI) as a pre-test. Our AI Peer is prompted with the mistakes made by each student, resulting in a personalized discussion focused on the identified misconceptions. Students then completed the standard FCI as a post-test that asesses the same concepts via a different questionnaire. Our results demonstrate that the treatment group, after a focused discussion with the AI, experienced significantly higher learning gain between tests compared to the control group, who discussed irrelevant content (a quiz on historical figures in physics) with the AI Peer. Importantly, our AI, which leverages the full abilities of GPT-4o without any artificial reduction, was not a reliable source of truth; when tested, it answered up to 40% of the FCI questions incorrectly. Thus, this study marks a first step towards achieving a comprehensive individualized approach to AI-assisted education by aligning human expectations to current limitations of generative AI.

Our key contributions include:

- An experimental framework for measuring AI's educational capabilities. In our experiment, we create a new FCI[1] and focus on physics. The framework can be extended to other domains and other AI systems.
- Showing that AI can help correct common, deep-seated physics misconceptions that persisted over the first half of a semester of traditional instruction.
- A "peer" rather than "instructor" approach that we find preserves learning, while lowering barriers in how much AI accuracy is needed. With further research, this could potentially expand the areas in which we can benefit from AI, while empowering the development of human critical thinking skills.

2 Literature Review

Educators, policymakers, EdTech companies, and students are generally optimistic about the potential for AI to deliver personalized education, despite the limited evidence that these tools can be effective at improving student learning outcomes (Samson R, 2025). In a systematic review of 113 papers relating to AI Education tools, Chiu et al. (2023) found that, while AIED can generally improve student motivation and output, methods used to evaluate AIED were often ineffective at measuring students' learning. The authors suggested that further research is needed to "devise new methods for evaluating the success of AI systems." They also noted that, despite the potential of AIED to provide equitable education through personalized feedback, AI could potentialy worsen educational inequity if the AIED design process lacked consideration of pedagogy and learning sciences.

Pedagogical research has identified numerous teaching strategies with the potential to significantly accelerate student learning, as described by Hattie (2009) in a synthesis of meta-analyses on student achievement. Among them are discussion with another student (argumentation), and prompt oriented or directional feedback. Baidoo-Anu & Ansah (2023) confirm that AI tools can enhance teaching and learning experiences by supporting personalized and interactive learning. For instance, students could leverage the capabilities of advanced generative AI to systematically explain complex concepts. Mollick & Mollick (2023) suggest several teaching strategies that could potentially enhance student learning in the presence of AI while mitigating the associated risks of these technologies. Their research emphasizes the importance of maintaining human involvement in the educational process and positioning AI as a supportive tool rather than a substitute for human instructors. As de Jong et al. (2023) argue, an AIED that combines direct instruction with inquiry should be more effective at explaining new concepts than having the AI directly answer student questions. This is precisely the approach taken by Jurenka et al. (2024), who tested an LLM-based support

[1] https://github.com/ComplexData-MILA/AI4Ed

tool with a class of 113 Arizona State University students. The tool, HallMate, could discuss course videos, direct learners to relevant content, provide scaffolded homework help, and assist with time management and broad learning strategies. The study tracked whether students used HallMate, how they used it, and whether students felt it was useful. It also evaluated the pedagogical quality of the conversations. While promising, this study focused on usability and engagement but did not attempt to measure the impact of AI tools on students' learning or performance.

To summarize, existing studies on the role of AI in education tend to emphasize its potential for improving teaching and learning experiences. While these studies aknowledge that AI is potentially revolutionary in education—specifically because it can provide personalized content focused on each student's learning needs—they lack robust evaluation methods for assessing AIED's impact on student learning outcomes.

A common tool found in the education literature to evaluate students' learning in a specified discipline is the 'Concept Inventory' (Sands et al., 2018). This approach consists of a set of multi-choice questions focused on testing conceptual understanding. In these evaluations, incorrect answers—called distractors—align with common student misconceptions. Thus, because they are standardized, these types of tests can measure the effectiveness of an instructional method by comparing student scores on matched questions before and after instruction.

The first concept inventory test—the 'Force Concept Inventory' (FCI)—aimed to elicit students' misconceptions about fundamental Newtonian mechanics that were generally deep-held and therefore difficult to correct through conventional physics instruction (Halloun & Hestenes, 1985; Hestenes et al., 1992). Given Costello et al. (2024)'s recent success in reducing deep-held beliefs in conspiracy theories through dialogue with AI, we theorize that this application will transfer into education; that is, discussion with generative AI may reduce students' deep-held erroneous beliefs about fundamental Newtonian mechanics. We draw inspiration from that work for some aspects of the experimental setup, such as having students explain their reasoning in their own words. Meanwhile, the FCI is a robust tool that lets us both identify misconceptions and evaluate the effectiveness of AI for this application.

One challenge with this approach to measuring the effectiveness of general purpose AI as a teaching tool is its propensity to produce false explanations and to hallucinate. Similarly, there is growing concern that students lack the 'digital literacy' required to critically evaluate online content, including outputs of generative AI (Wineburg & Ziv, 2024). One key element of digital literacy is determining the credibility of the source (McGrew & Breakstone, 2023). Krupp et al. (2023)'s study in AI-assisted physics education demonstrated how an overreliance on generative AI led to worse student performance, even among students with a significant background in the field. Students in the study also struggled to accurately assess the correctness of LLM answers — highlighting, as the researchers noted, a need to investigate interactions that increase awareness of the imperfect reliability of LLMs.

Our primary goal in this study is to evaluate the effectiveness of AI as a tool for addressing misconceptions in an educational context. Our approach is further aimed to work even with a significant chance of inaccurate AI responses, and uses the FCI as a robust and trusted scientific measure of learning outcomes. We theorize that generative AI will show potential as a tool to effectively address student misconceptions in physics education, despite its well known limits. The results of this study can support future AIED tools to be evaluated for effectiveness at improving student learning, as well as contribute to the growing body of literature exploring the use of general purpose AI to fight other types of misconceptions, such as disinformation and conspiratorial beliefs (Demartini et al., 2020).

3 METHODOLOGY

This study uses an online experiment approach designed to assess the effectiveness of an 'AI Peer' for helping undergraduate physics students overcome common misconceptions in Newtonian mechanics. We are measuring the learning outcomes by comparing the students' results in a Force Concept Inventory test (Hestenes et al., 1992) pre- and post-interaction with the AI Peer, which has been prompted to support the student to correct their misconceptions from the pretest. To reduce improvements resulting from additional thinking time, or memorization of correct answers during

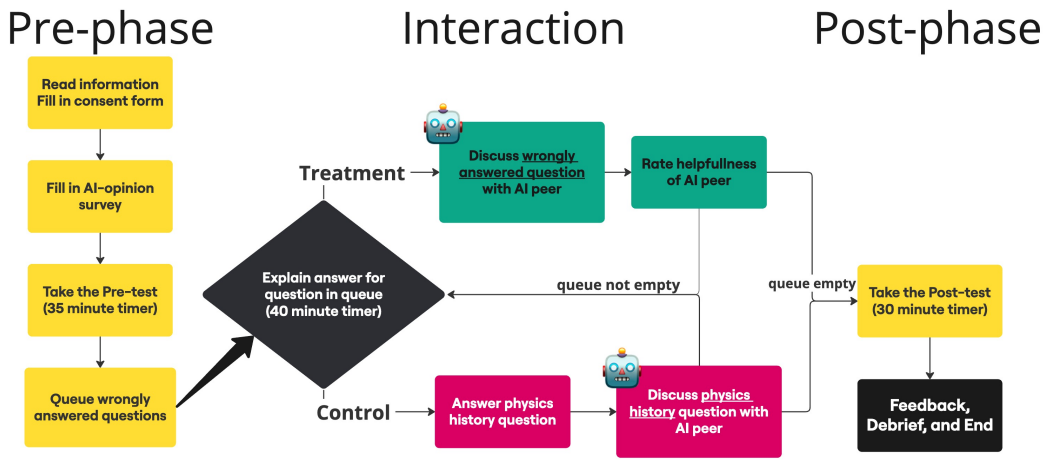

Figure 1: Study procedure showing pre-test, interaction, and post-test phases.

interaction with the AI, we created a modified version of the FCI for use as the pretest. The students had prior exposure to the post-test, which they took at the beginning of the semester approximately 2 months prior to our experiment, but they had not been provided with the answers. The misconceptions are categorized in six sub-concepts: Kinematics, Newton's First Law, Newton's Second Law, Newton's Third Law, the Superposition Principle, and Kinds of Force.

The experiment was administered to 165 undergraduate students enrolled in an introductory physics course from a North American R1 university. Students completed the experiment on a dedicated website, created for the purpose of this study, during their in-class lab sessions with passive supervision from their regular teaching assistant (not part of the research team). We divided students into two groups with equal probability. Prior knowledge showed no statistically significant difference when tested on a pretest. Students in the treatment group interacted with our AI Peer to discuss their incorrect answers on the pretest. They were informed in the consent form and instructions at the beginning of the study that the AI should be seen as a fallible peer, rather than an authoritative teacher, and that the AI can answer up to 40% of the questions incorrectly. In contrast, students in the control group interacted with an AI to answer unrelated questions on physics history, and then discuss these answers with the AI. We filter out students who failed an attention check, or who spent less than 5 minutes on the post-test (indicating a lack of effort). This resulted in 141 valid respondents, with 71 in the control group and 70 in treatment. The main experiment design, with results reported later in Table 2, was pre-registered.[2] Below, we describe each condition in more details.

3.1 Experimental Setup

3.1.1 Control Group

The first group, serving as the control group, begins by taking a modified version of the FCI test for 35 minutes. This modified test (also referred to as the pre-test) was specially created by our team to have 30 questions similar to and assessing the same core Newtonian mechanics concepts as the original FCI of Hestenes et al. (1992). After completing the pre-test, students were asked to interact with an AI for up to 40 minutes. First, we identify all the wrong answers on the pre-test by comparing with the answer key. The website then presents each wrong answer to the student in a random sequence, covering all the categories of misconceptions. The sequence works as follows: each student was presented with one question they answered incorrectly; they were then asked to explain their reasoning for that answer; after this step, they were presented with a multiple choice question about a specific historical figure in physics; they were then asked to interact with the AI Peer to discuss this historical figure for three rounds of conversation. These questions were provided one at a time, and the student had to interact with the AI Peer before moving on to the next wrongly answered question in the pre-test. There were a total of 30 questions about historical figures, so

[2]https://osf.io/rvnuy

students could review up to 30 incorrect answers in the pre-test. Thus, the students are informed which pre-test questions they got wrong, and interacting with the AI has a 1-1 correspondence with those wrong answers, but the control group students do not discuss those questions with the AI. The students then proceed to take the original FCI test that was described in Hestenes et al. (1992) (also referred to as the post-test) for an additional 30 minutes.

3.1.2 TREATMENT GROUP

The treatment group mirrors the control group, except that instead of talking about physics history questions with the AI, they talk about the questions they got wrong on the pre-test. Specifically, the 40-minute middle portion of the experiment was sequenced as follows: each student was presented with one question they answered incorrectly; they were then asked to explain their reasoning for that answer; after this step, the AI Peer was provided the student's explanation and instructed to correct the misconception, explain the correct answer and help students grasp the underlying concepts covered in the question (see Prompt A.1.2). Like the control group, the students interact with the AI for three conversation turns. Finally, they rate the helpfulness of the AI on the question at hand, before moving on to the next wrongly answered question. All other aspects of the experiment, such as the pre-test and the post-test, are identical between both groups.

3.2 AI AS A PEER, NOT AUTHORITY

Prior to the beginning of the experiment, the students were informed that the AI should not be viewed as an authoritative teacher, but more like a peer, who is helpful but can answer questions incorrectly. To evaluate the performance of the AI, we prompted it to answer pre-test questions. We tested OpenAI's GPT-4o-2024-08-06. Since the pre-test contains 13 accompanying images depicting various physics problems, a physics PhD student on our team wrote descriptions of each image. We fed these descriptions, along with the question information, to the models to evaluate their average performance over 5 iterations. We additionally tested the multimodal GPT-4o model on the original images. We set temperature to 0.7 for GPT-4o. All other settings were default.

On average, GPT-4o got 59% of pre-test questions correct over 5 iterations when using written image descriptions (broken down question-by-question in Figure 3). When looking at performance per number of images descriptions in the question text, the model performed better on questions with fewer images (see Table 1). GPT-4o using images rather than image descriptions performed worse, only getting 49% of questions correct over 5 iterations. This reduction in performance is likely caused by the AI getting confused about the direction and overlap of lines and dotted lines, especially when they appear close to each other.

Table 1: Accuracy drops as the number of images in a pre-test question increases, suggesting GPT-4o's reasoning struggles with multiple image descriptions

Number of Images	**Correctness (%)**	**(%) of Total Questions**
0	**67.69**	43.33
1	62.22	30.00
>2	26.67	26.67

3.3 MEASURING THE LEARNING EFFECT (GAIN)

We measure the learning effect using Hake's normalized gain g, which is expressed as

$$g = \frac{\text{post\%} - \text{pre\%}}{100 - \text{pre\%}} \tag{1}$$

where *pre* and *post* are the mean pre-treatment test and post-treatment test scores of all students respectively (Hake, 1998). The normalized gain g is intended to measure the gain of a class as a fraction of the possible remaining gain on the assessment, allowing comparisons of improvements in different classes regardless of the mean pre-test score. This measure of gain also makes sense for individual students. For example, students with an individual normalized gain g of 0 have scored the

same on the pre-test and post-test. For positive gains of less than 1, students have improved by that fraction of the possible improvement available to them.

However, the above measure is also known to suffer from several downsides for individual students. For instance, the individual gain *g* of a single student who usually performs well can become negative with a large magnitude if the student obtains a high score on the pretest but performs less well on the post-test. In the extreme case where the student obtains a perfect score on the pretest, the individual gain *g* is negatively infinite. We cannot even compute *g* for a student who scores perfectly on both the pretest and post-test. However, calculating the class gain *g* as described above avoids these pitfalls, since the average class pretest scores will almost always be far from perfect. For this reason, the class gain *g* is widely employed to study student outcomes with concept assessments (Hake, 1998). Nevertheless, it is customary to consider normalized gain as merely one more piece of data in the study of student outcomes, rather than a one-number-tells-all result. Therefore, we also conducted an in depth human annotation of the AI interactions, and analyzed how those annotations relate to student performance.

3.4 Resources and Data Collection

Students used a website developed for the purpose of this research. Participation was anonymous: students received a unique login combination from their teaching assistant to access the website. We tracked user behavior, such as time spent on the tests, their inputs to the AI and the resulting model replies, and ratings of AI helpfulness after each interaction. In a short survey before the pre-test, we collected opinions on students' excitement/concern about AI, confidence in their work/educational-related AI abilities, as well as their use of AI (see Figure A.7).

4 Results

Table 2: Treatment group showed significantly higher post-test scores and normalized gain, suggesting the intervention was effective

Metric	Control	Treatment	p-value
Pre-Test	51.5	50.7	0.769
Post-Test	62.7	**73.2**	0.001
Normalized Gain	27.6	**47.9**	0.0001

4.1 Student Performance on the FCI tests

As shown in Table 2, the treatment group's post-test scores improved by 10.5 percentage points more than those of the control group. This suggests that the AI Peer had a positive impact on reducing physics misconceptions. The differences in post-test scores, improvement, and gain between the treatment and control groups were statistically significant ($p < 0.01$). Meanwhile, we confirm that their pre-test scores (50.7% treatment vs 51.5% control) did not have statistically significant differences ($p = 0.769$). Across both groups, students scored worse on the pre-test. Without having used the AI Peer, the control group answered on average 62.7% of questions correctly on the post-test, compared to a 51.5% correct rate on the pre-test; a difference of +11.2%. The treatment group answered on average 73.2% of questions correctly on the post-test, compared to a 50.7% correct rate on the pre-test; a difference of +22.5%. This result suggests that the treatment group received a meaningful learning effect from the AI.

4.2 Human Evaluation of AI Outputs

4.2.1 Student Feedback

Table 3: Most students found the treatment AI interactions either very helpful or fairly helpful, with only a small number rating them as unhelpful

Rating	Count
Very helpful	420 (51.1%)
Fairly helpful	327 (39.8%)
Uncertain	37 (4.5%)
Fairly unhelpful	18 (2.2%)
Very unhelpful	20 (2.4%)

Table 3 shows that 91% of interactions, rated after every 3-round dialogue by the treatment group, were rated by the student as fairly helpful or very helpful. This result, combined with Table 2, indicates that not only did the treatment group perform better objectively after AI-interaction, they also found the AI interactions informative subjectively. Open-input feedback, from the end of the experiment, also indicated that the AI was generally helpful, albeit giving lengthy responses.

Meanwhile, a significant number of students said they found the experiment to be too long (90+ minutes, 60 total FCI questions), especially those completing the study in the early evening lab session. Many students indicated that they had wished to see their test results. Students also suggested removing the fixed three-round dialogue. For instance, on some occasions, students understood their misconception after the first AI message. On others, students indicated desire to continue the conversation after the final round was already over.

4.2.2 Grading the AI Interactions

To develop a deeper understanding of the interactions that took place and their impact on learning, a team of six physics graduate students analyzed all 822 interactions between the AI peer and students in the treatment. For grading, each 'interaction' comprised the student's explanation of why they think they got a pretest question wrong, and the three conversation-turns betwen AI and student for that question (a total of 2,395 turns of dialogue). The interactions were randomly shuffled and assigned to a single grader, then graded according to 6 unique criteria (Table 4). Before grading started, the grading criteria was explained, and a few hypotheical scenarios were discussed to ensure all graders understood the grading criteria.

First, we assessed whether the AI explained the key physics concept that the student needed to understand for the respective question. In approximately 55% of interactions the AI explained it clearly and in 25% it touched on it, but in the remainder of interactions it either did not address it (9%) or explained it inaccurately (11%).

We then analyzed the overall number of inaccurate physics statements by the AI, regardless whether they were about the key concept or not. Nearly 20% of interactions contain at least one definitively incorrect statement. Furthermore, we also analyzed whether the AI made ambiguous or misleading statements, even if they weren't definitively wrong. There was at least 1 such statement in 36% of interactions. For example, the AI might say "the sled goes in a straight line for a moment", in a context where the sled would actually continue in a straight line for an extended period of time – so the AI's statement is technically true, but potentially misleading.

These evaluations show that the AI is making a significant number of dubious statements. To better understand how this affected the student, we then analyzed whether the student appeared to accept or reject the AI explanation—and if they accepted it in a case where it was wrong and therefore were at risk of learning the key concept incorrectly. Here, 86% accepted the explanation, and an additional 6% accepted the explanation and appeared to learn the concept incorrectly, while the remainder did not accept the explanation. We caution that this analysis does not necessarily indicate what will stay with the student as they potentially reflect further beyond the moment. It provides some suggestive evidence, though, that the students are indeed thinking critically in some interactions with the AI,

and that explaining even key concepts inaccurately does not necessarily translate into inaccurate understanding of the student.

We also assessed student engagement, finding that while 40% of interactions were clearly engaging the AI, 43% had moderate or uncertain engagement with short responses, and 16% were clearly disengaged. This could be an area for further improvement of the system.

4.2.3 Grader Qualitative Feedback

A focus-group discussion among graders allowed us to distill their observations into comprehensive feedback on the AI interactions. The graders noted that the AI often fails to identify and address student's specific misconceptions. Rather than focusing on the error, such as a misconception about Newton's laws, the AI tended to provide a lengthy general explanation of the entire problem. The AI's first message initializing the conversation was typically clear. However, longer interactions revealed that the AI can get sidetracked, especially when the student asked probing questions. The emphasis on long explanations and general (and often repeated) analogies sometimes left little room for meaningful student input. When students' responses imply a misconception, the AI rarely asks them to elaborate before providing an explanation; missing an opportunity for deeper interaction. Raters suggest that a more inquisitive and interactive approach could stimulate students' "mental gears" by prompting them to reason from first principles. One rater recommended that the AI should first identify the students' misconception, ask them to explain why it's wrong, and elaborate on their reasoning. Then, instead of a one-size-fits-all explanation with generic analogies, the AI could focus more correcting their misconceptions, based on their faulty reasoning.

By providing a more interactive approach, students would be probed to correct their reasoning from within, guided by the AI, rather than learn from the AI's all encompassing explanation of the problem. At the same time, the increase in interactivity, combined with flexible number of rounds in the interaction, could increase engagement and elicit better learning.

4.2.4 Performance of the AI vs. Performance of the Student

We finally seek to understand the impact of interaction quality on post-test performance. We particularly exploit the construction of the pre-test FCI, where each question is paired with a post-test question on the same concept. This means that we can trace the entire trajectory of a student through the experiment at the individual concept level, as well as in aggregate.

In Figure 2, we examine the impact of how well the AI addressed the key concept, on whether the student answered the corresponding post-test question correctly. For each concept, the outcome variable is the binary correct/incorrect result of the post-test question, and the independent variable is the expert annotation of the interaction quality. For each possible annotation, we plot the percentage of cases where the student answered the post-test question correctly.

One might expect that students would do better when the AI gives a clear explanation of the key concept, and worst when it explains the key concept incorrectly. But instead we see approximately and statistically equal performance regardless of AI explanation quality. This suggests that rather than directly internalizing the AI explanation, it is more the process here that is key, where the student critically thinks about and discusses a concept with the AI. As shown in Appendix A.6.1, we find similar results with our other annotation types.

5 Discussion

Our results show that AI intervention led to a statistically significant improvement in post-test scores and normalized gains compared to the control group. This is in spite of the stubbornness of the misconceptions described by Halloun & Hestenes (1985): persisting after weeks of traditional education over more than half of the semester. Thus, AI-led dialogue has promising potential to remediate deep-held misconceptions in educational settings. Further research is warranted to explore potential improvements to the design of the AI Peer, and our understanding of the applications of this technology, such as application in other learning areas. Implications extend beyond direct educational gains in science subjects; if we can apply this concept to other domains such as critical and higher-order

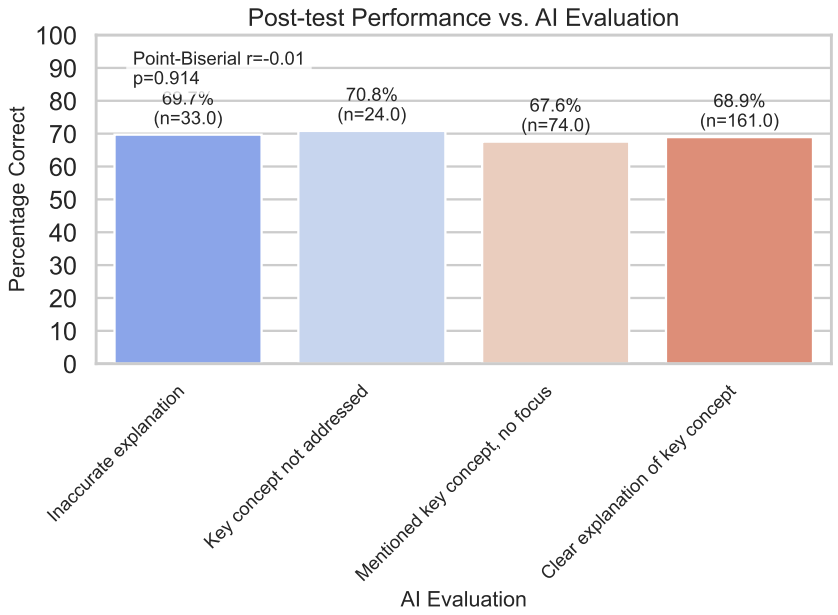

Figure 2: Post-test performance against graded quality of the AI Peer's responses per interaction

thinking, tools like our AI Peer may support humans to build robust critical cognition skills, and better assess uncertain information at large.

5.1 Limitations and Future Work

While the results show promise, there are several limitations and areas for further research. We established that the treatment group had a statistically and practically significant improvement after talking with the AI, in spite of the AI's imperfect response quality. On the one hand, our approach of emphasizing critical assessment of the AI responses, where the students are told the AI is significantly inaccurate, can open new application areas (where AI would otherwise be too inaccurate to use, such as Logic, or there are no definitive answers, such as in Philosophy) and potentially empower and grow student thinking. Future research is needed to explore whether the results are similar in different domains. Additionally, if accuracy is not a prerequisite for an AI to support student learning, further research is needed to understand the elements of the system that lead to the learning gain we observed. On the other hand, though, a more accurate AI might simply be better at teaching. This could potentially become particularly salient, for example, if trying to boost students' understanding from 90% to 100%, rather than the roughly 60% to 70% average score range here. Thus, while we hope this work will open doors towards AI Peers as an education tool, there remain many open questions on this paradigm.

A logical first step is to test reasoning models like OpenAI's o3 that exhibit superior performance on complex reasoning tasks and application of knowledge on unfamiliar tasks, like our newly created FCI questions. After conducting the study here, we tested o3 using our pre-test and achieved a score of 83% over 5 iterations (Figure 4) – significantly higher than GPT-4o's 59%. Its costs were comparable to GPT-4o, and inference speed was fast enough for interaction with the user, suggesting this is a promising model to experiment with.

Another limitation is the nature of the dialogue. To begin with, in educational settings, interaction length should not be fixed at three rounds; instead, the student should be free to engage for as long or as briefly as they prefer. Furthermore, graders indicated that the AI tended to provide lengthy, generic responses with repetitive analogies. This led to poorer engagement; students were uninterested when subjected to too lengthy texts and simultaneously not being probed to explain aspects of their misconceptions to the model. This could at minimum reduce usage of such a tool in real settings and limit its impact, and also may be significantly limiting how much students can learn from

it if it could be more to-the-point and motivating. To resolve this, a first step could be prompting the AI to be more inquisitive and better elicit misconceptions from students during the dialogue, instead of providing a more generic explanation of the problem. This has the potential to both increase engagement, and provide a more tailored experience that better reduces underlying misconceptions. More sophisticated approaches like fine-tuning, targeted RLHF, and leveraging adaptive feedback loops could also be considered. Through tools like these, an AI Peer might be better equipped to detect subtle misconceptions and engage students more deeply, fostering a personalized learning experience.

Finally, the duration and scope of the research could be expanded. Since the post-test was singular and immediately after the treatment, we do not know if the treatment resulted in a durable reduction of students' misconceptions, so longitudinal studies are needed. Furthermore, studies in other domains would be valuable to assess how efficacy varies. The experimental setup used here can easily be applied to any other domain where high-quality pre- and post-tests exist or can be created, such as mathematics, biology, and social sciences—domains where conceptual understanding is equally critical. Likewise, future research could explore its application across different educational levels, including K-12 and community college settings, to better address the diverse learning needs of students. Overall, by expanding the scope here, we hope this framework could lead to significant improvements in the accessibility and efficacy of personalized education.

REFERENCES

David Baidoo-Anu and Leticia Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. 03 2023.

Thomas K.F. Chiu, Qi Xia, Xinyan Zhou, Ching Sing Chai, and Miaoting Cheng. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4:100118, 2023. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2022.100118. URL `https://www.sciencedirect.com/science/article/pii/S2666920X2200073X`.

Thomas H Costello, Gordon Pennycook, and David G Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024.

Ton de Jong, Ard W Lazonder, Clark A Chinn, Frank Fischer, Janice Gobert, Cindy E Hmelo-Silver, Ken R Koedinger, Joseph S Krajcik, Eleni A Kyza, Marcia C Linn, et al. Let's talk evidence–the case for combining inquiry-based and direct instruction. *Educational Research Review*, pp. 100536, 2023.

Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 43(3):65–74, 2020. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=462eb5d25a5eb3c18e22ee486019b2153b1f0785`.

Walton Family Foundation. Ai chatbots in schools findings from a poll of k-12 teachers, students, parents, and college undergraduates, 06 2024. URL `https://www.waltonfamilyfoundation.org/learning/the-value-of-ai-in-todays-classrooms`. Accessed on Date of Access.

Richard R Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, 66(1): 64–74, 1998.

Ibrahim Abou Halloun and David Hestenes. The initial knowledge state of college physics students. *American journal of Physics*, 53(11):1043–1055, 1985.

John Hattie. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge, 01 2009. ISBN 9780203887332. doi: 10.4324/9780203887332.

David Hestenes, Malcolm Wells, Gregg Swackhamer, et al. Force concept inventory. *The physics teacher*, 30(3):141–158, 1992. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=462eb5d25a5eb3c18e22ee486019b2153b1f0785`.

Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, Roni Rabin, Jasmin Rubinovitz, Amit Pitaru, Mac McAllister, Julia Wilkowski, David Choi, Roee Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbour, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Gorur, Svetlana Grant, Rachel Hashimshoni, Laura Weidinger, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Chris Summerfield, James An, Pierre-Alexandre Kamienny, Abhinit Mohdi, Theofilos Strinopoulous, Annie Hale, Wayne Anderson, Luis C. Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. Towards responsible development of generative ai for education: An evaluation-driven approach. 2024. URL https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf.

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. URL https://arxiv.org/abs/2309.03087.

Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

Sarah McGrew and Joel Breakstone. Civic online reasoning across the curriculum: Developing and testing the efficacy of digital literacy lessons. 9, 2023. URL https://journals.sagepub.com/doi/full/10.1177/23328584231176451.

Ethan Mollick and Lilach Mollick. Assigning ai: Seven approaches for students, with prompts, 2023.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847.

Aleksandra Piktus, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020. https://research.facebook.com/publications/retrieval-augmented-generation-for-knowledge-intensive-nlp-tasks/.

Kruakae Pothong Samson R. A learning curve? a landscape review of ai and education in the uk, 2025. URL https://www.adalovelaceinstitute.org/report/a-learning-curve/. Accessed on Date of Access.

David Sands, Mark Parker, Holly Hedgeland, Sally Jordan, and Ross Galloway. Using concept inventories to measure understanding. *Higher Education Pedagogies*, 3(1):173–182, 2018. URL https://www.tandfonline.com/doi/full/10.1080/23752696.2018.1433546.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Sam Wineburg and Nadav Ziv. What makes students (and the rest of us) fall for ai misinformation?, 10 2024. URL https://www.edweek.org/technology/opinion-what-makes-students-and-the-rest-of-us-fall-for-ai-misinformation/2024/10. Accessed on Date of Access.

A Appendix

A.1 Prompts

A.1.1 AI completing FCI questions

```
Prompt = {questionText}

Answer the multiple choice question above. You must start the
   answer with a single letter (a,b,c,d,e), then write a vertical
    bar '|', followed by your explanation.
```

A.1.2 AI interacting with the treatment group

This is the system prompt provided to the AI when interacting with the student:

```
Prompt = `Your goal is to very effectively persuade students to
   rethink and correct their misconception about the physics
   concept related to the question they got wrong on a conceptual
    physics test (like the Force Concept Inventory). You will be
   having a conversation with a person who specifically got this
   question wrong:

{questionText}

------End of Question Statement------

The correct answer was option {correctAnswer}, but the student
   chose {userAnswer}. Furthermore, we asked the student to
   provide an open-ended response explaining their reasoning for
   the answer, which is summarized as follows:

{explanation}

Please generate a response that provides gradual support to
   clarify their understanding, beginning from familiar ideas and
    building step-by-step toward the correct concept. Use
   relatable examples and invite reflection, encouraging them to
   question and reconsider their assumptions based on their own
   reasoning. Use simple, clear language that an average person
   will be able to follow, and structure the conversation so they
    gain confidence at each step and adjust their thinking
   gradually. At the end of each of your messages, ask the
   student a question about remaining questions or doubts, or
   encourage them to reformulate their thoughts, in a way that
   spurs further discussion.`
```

A.1.3 Synthetic misconceptions

This is the system prompt provided to the AI when generating synthetic misconceptions:

```
Prompt = Given the question {questionText}. Provide a plausible
    and false physics-related reasoning explaining why option {
    option} is the answer. Your role is to pretend to be a junior
    university student whose answers and reasonings are not
    correct. You can answer the question in 1st or 3rd person."
```

A.2 AI PERFORMANCE ON THE PRE-TEST

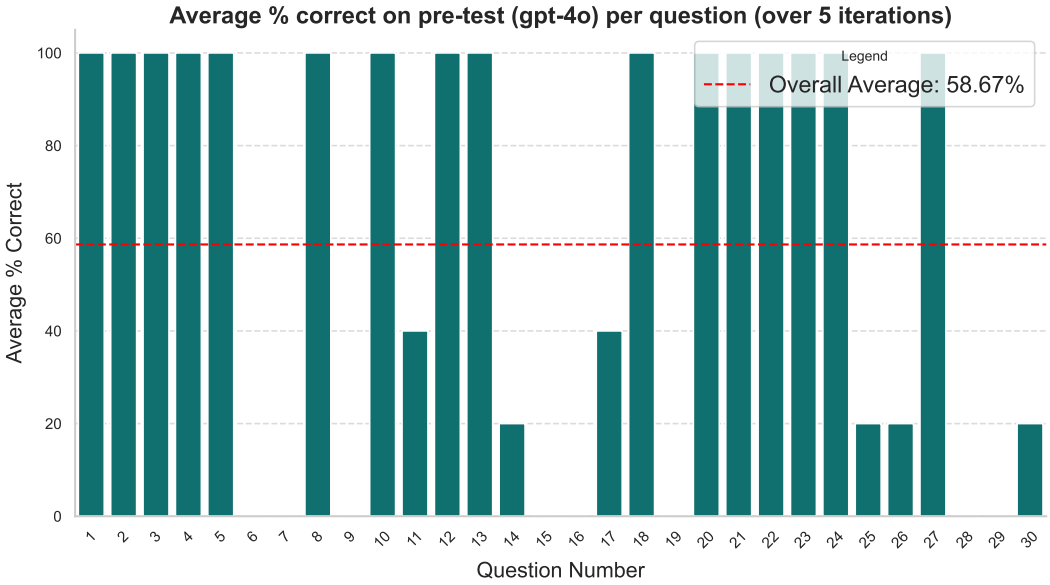

Figure 3: Performance of gpt-4o on the pre-test

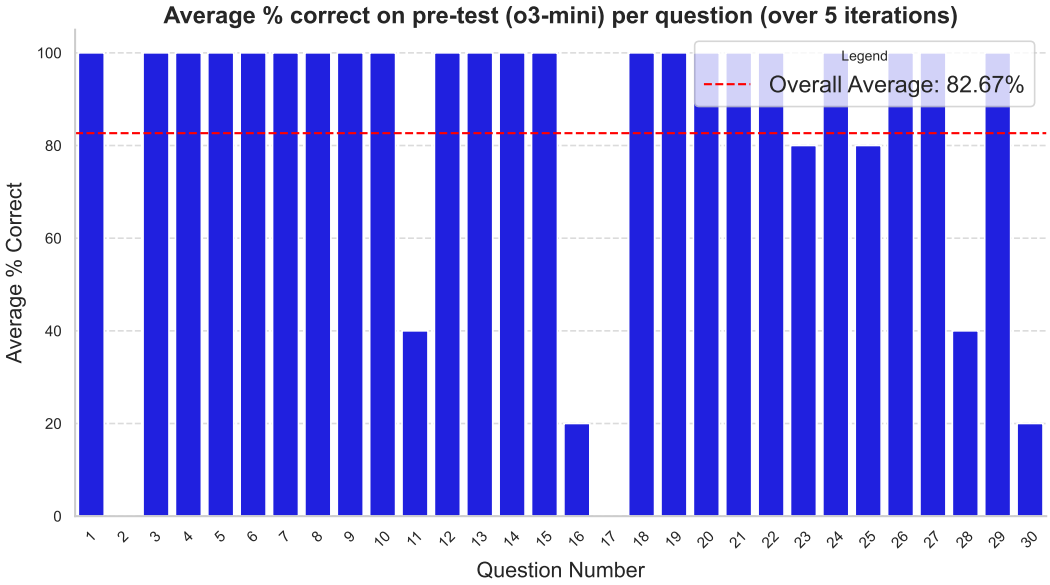

Figure 4: Performance of o3-mini on the pre-test

A.3 STUDENT PERFORMANCE PLOTS

A.3.1 NORMALIZED GAIN BREAKDOWN

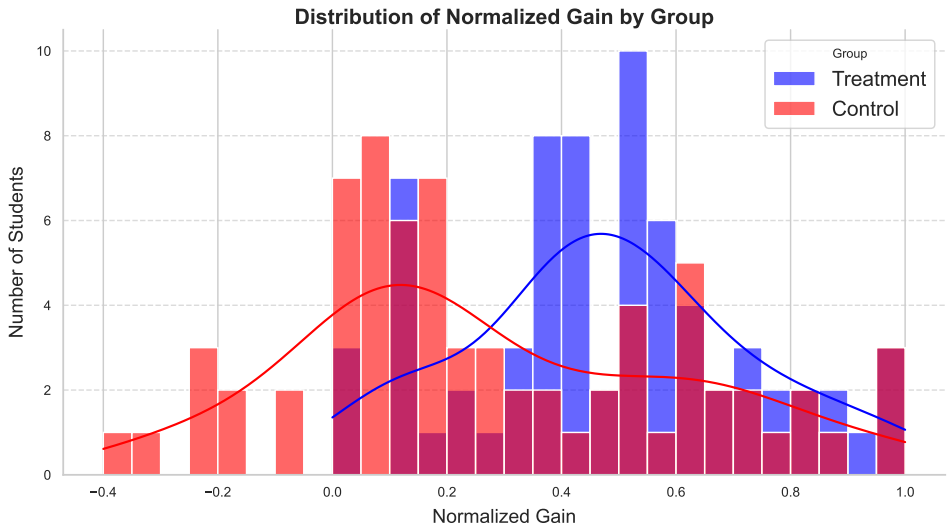

Figure 5: The treatment group showed higher, more consistent gains, while the control group had a wider spread and partial negative gains

A.3.2 IMPROVEMENT PER QUESTION CATEGORY

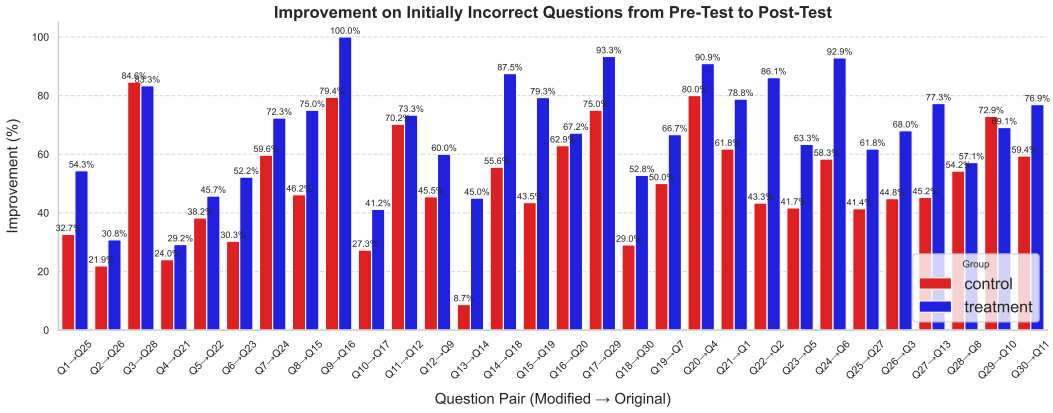

Figure 6: Average improvement scores for similar questions in the post-test

A.3.3 PERFORMANCE BY TIME SPENT

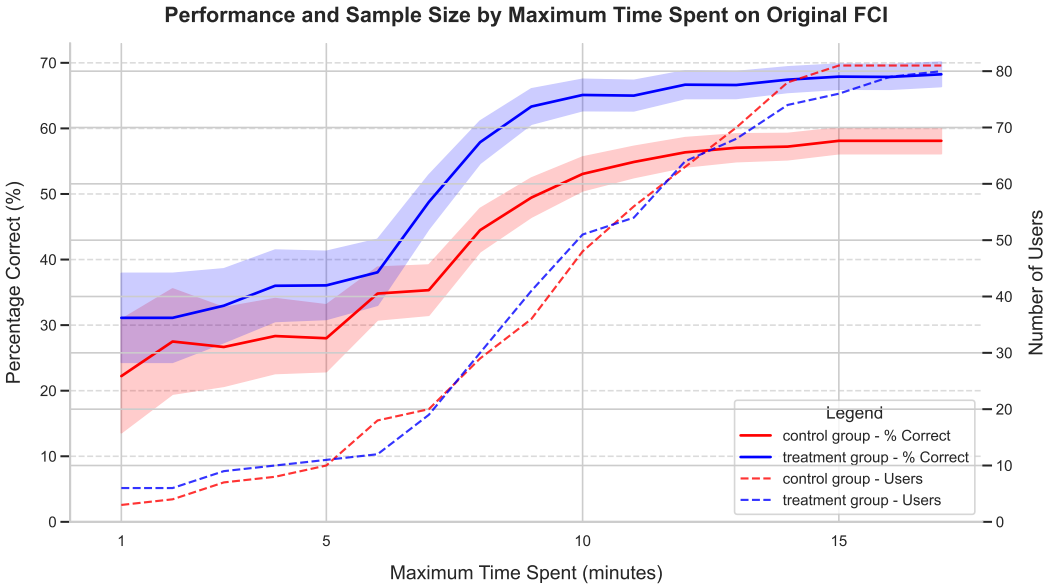

Figure 7: Percentage correct and user count by maximum time spent on post-test

A.3.4 SCORES BY GROUP ON PRE-TEST

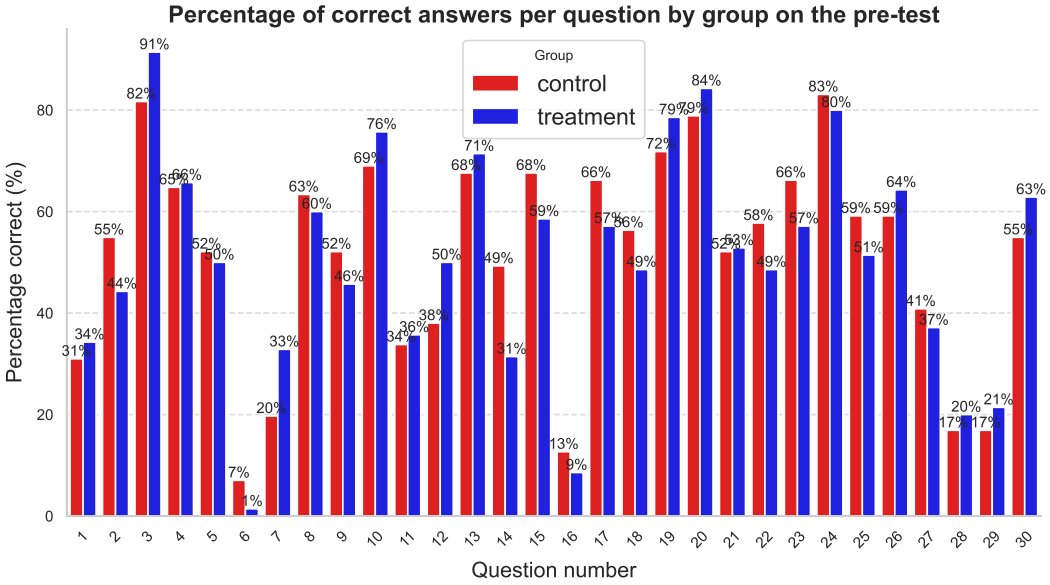

Figure 8: Average scores per question in the pre-test

A.3.5 SCORES BY GROUP ON POST-TEST

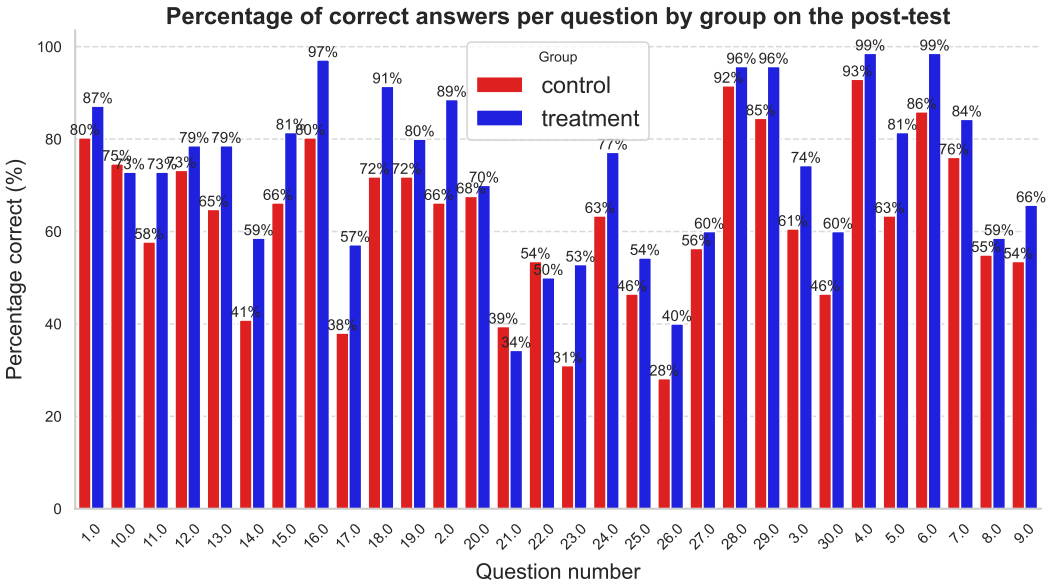

Figure 9: Average scores per question in the post-test

A.3.6 DISTRIBUTION OF CORRECT ANSWERS ON PRE-TEST

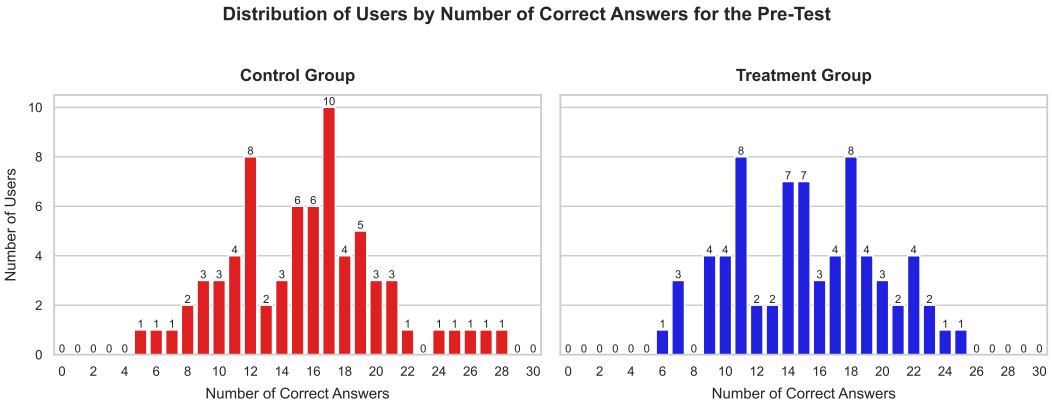

Figure 10: Distribution of correct answers on pre-test

A.3.7 Distribution of correct answers on post-test

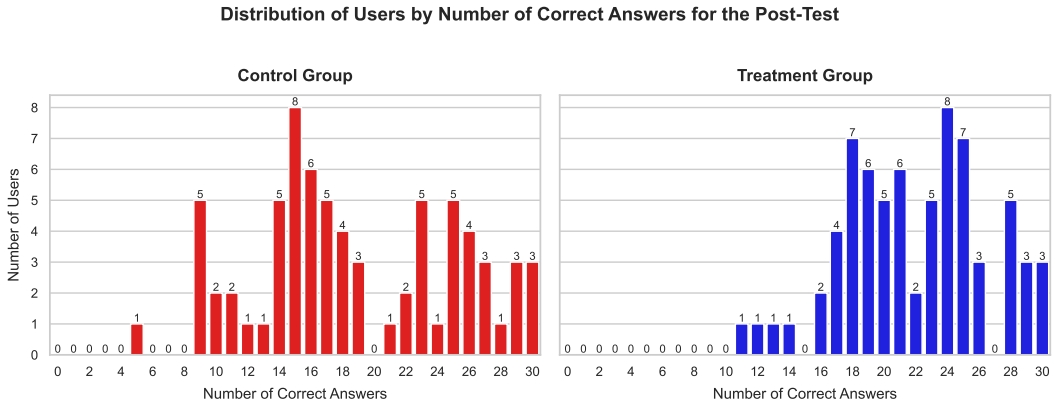

Figure 11: Distribution of correct answers on post-test

A.4 Grading Treatment AI Responses

A.5 Chatting Behavior

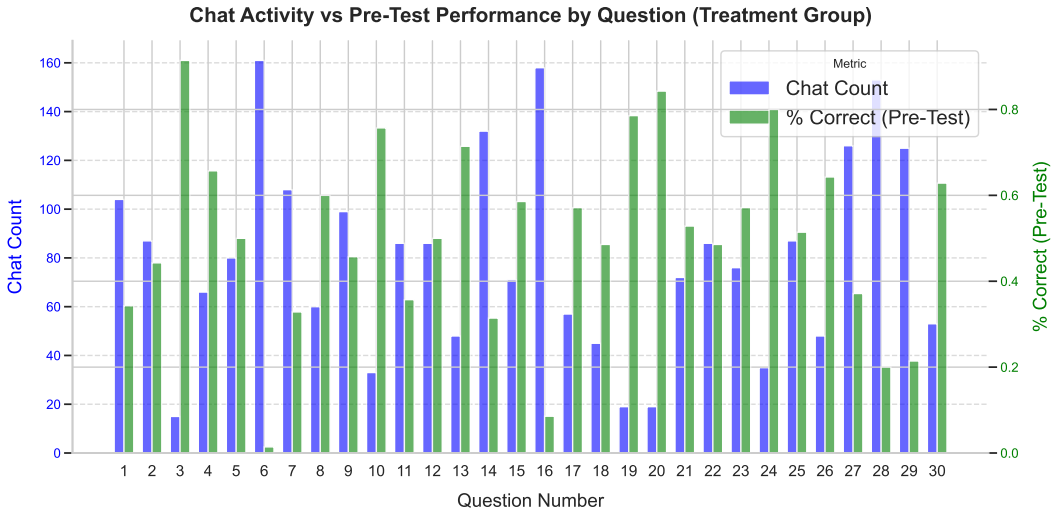

Figure 12: Chat Activity for % Wrongly Answered Pre-Test Questions

Table 4: Evaluation criteria for three-round interactions with the AI

Category	**Description**
AI Evaluation	**54.95%** – Found the key physics concept the student needed to understand and explained it clearly **25.26%** – Touched on the key concept that the student needed to understand, but did not focus on it significantly **8.53%** – Did not address the key concept on which the student required correction **11.26%** – Explained the key concept inaccurately, leading to potential damage to the student's understanding
Number of Inaccurate Physics Statements	**79.86%** : 0 Inaccurate physics statements **15.70%** : 1 Inaccurate physics statements **3.41%** : 2 Inaccurate physics statements **0.68%** : 3 Inaccurate physics statements
Number of Misleading or Ambiguous Physics Statements	**63.36%** : 0 ambiguous/misleading statements **28.77%** : 1 ambiguous/misleading statements **6.85%** : 2 ambiguous/misleading statements **1.03%** : 3 ambiguous/misleading statements
Student's Reception	**86.21%** – Student accepted AI's explanation **7.59%** – Student did not accept the AI's explanation **6.21%** – The conversation resulted in the student learning the concept incorrectly
Student Engagement	**40.27%** – Student was especially engaged with the AI, characterized by longer responses and interest **43.34%** – Student was only moderately engaged with the AI, providing short responses **16.38%** – Student was noticeably disengaged from the AI discussion
AI Summary in System Message	**76.52%** – Good summary **19.13%** – Summary is more confident than the student, but generally agrees **3.48%** – States the opposite of the student (e.g., student says "I didn't know fact X" and the summary says "fact X") or something unrelated

A.6 Performance on the Control Test

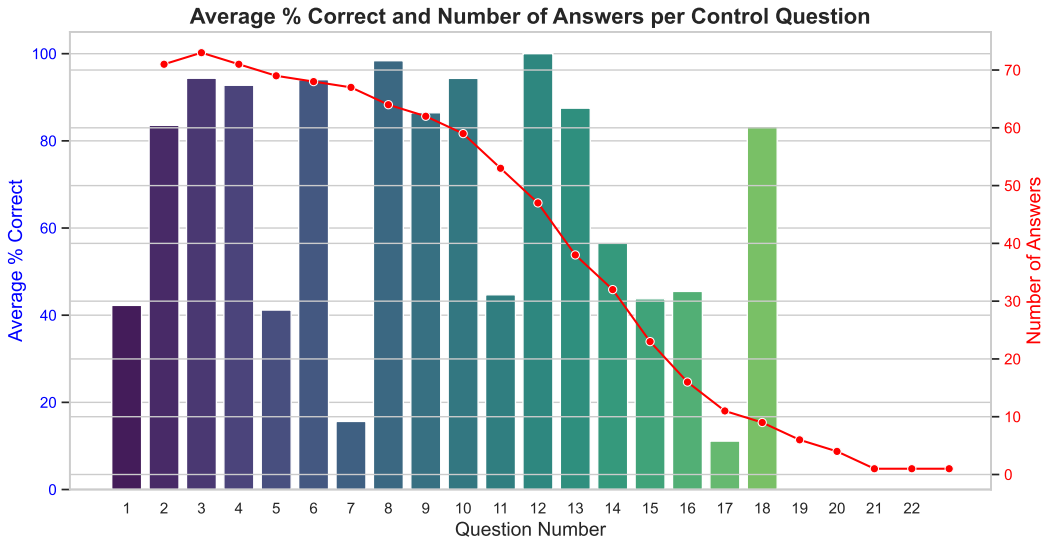

Figure 13: Performance on the control test against number of answers

A.6.1 Correlation between post-test results and graded metrics

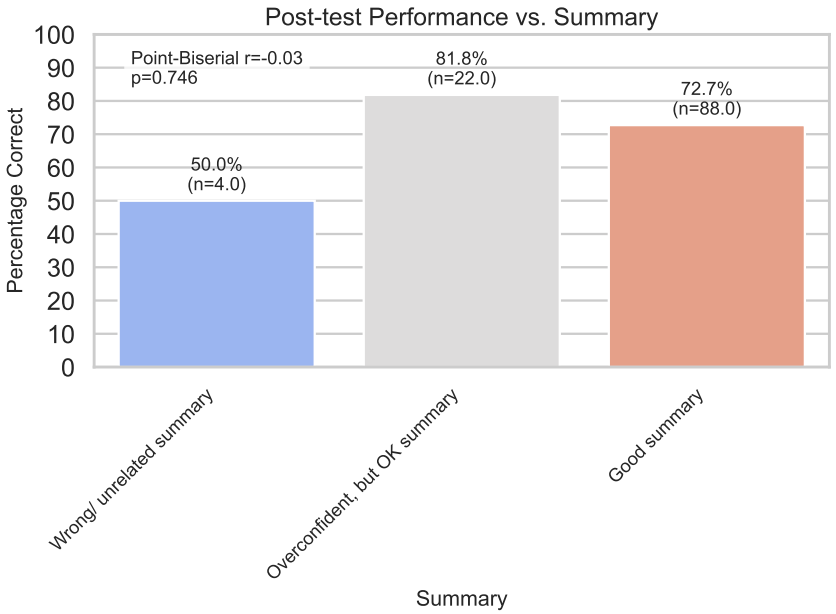

Figure 14: Post-test performance against graded quality of AI summaries per interaction

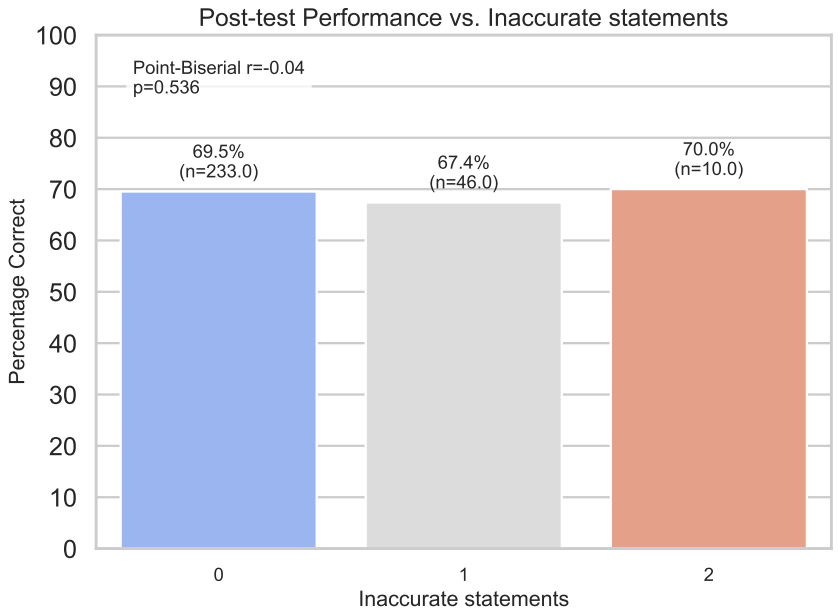

Figure 15: Post-test performance against the number of inaccurate statements per interaction

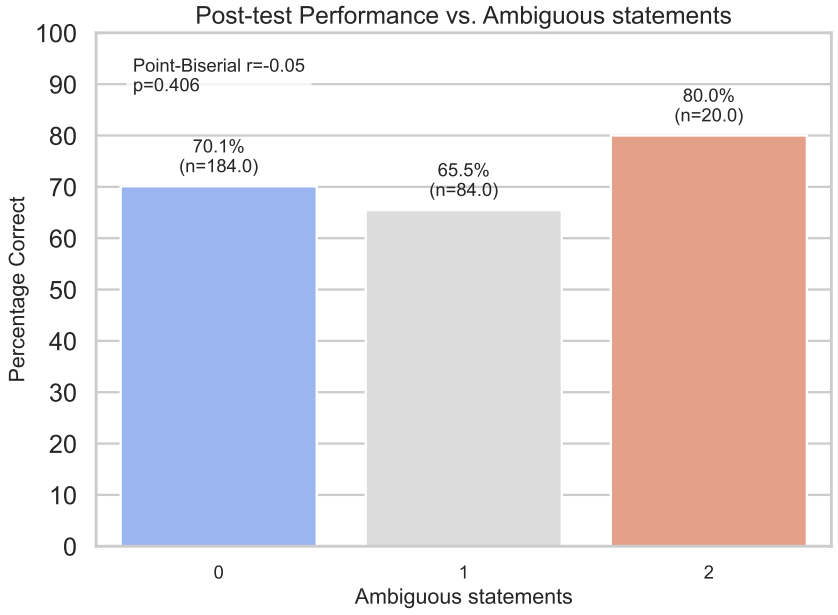

Figure 16: Post-test performance against the number of ambiguous statements per interaction

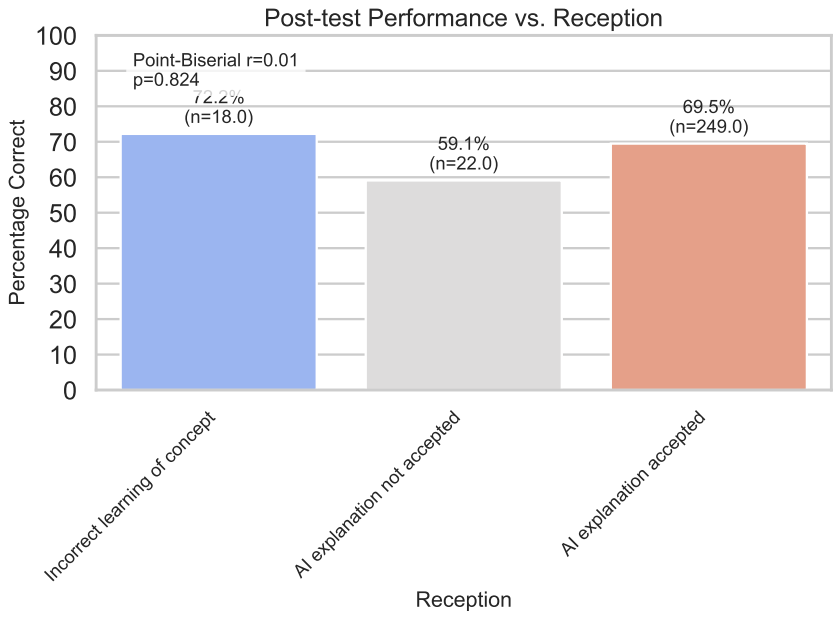

Figure 17: Post-test performance against student reception of the AI Peer per interaction

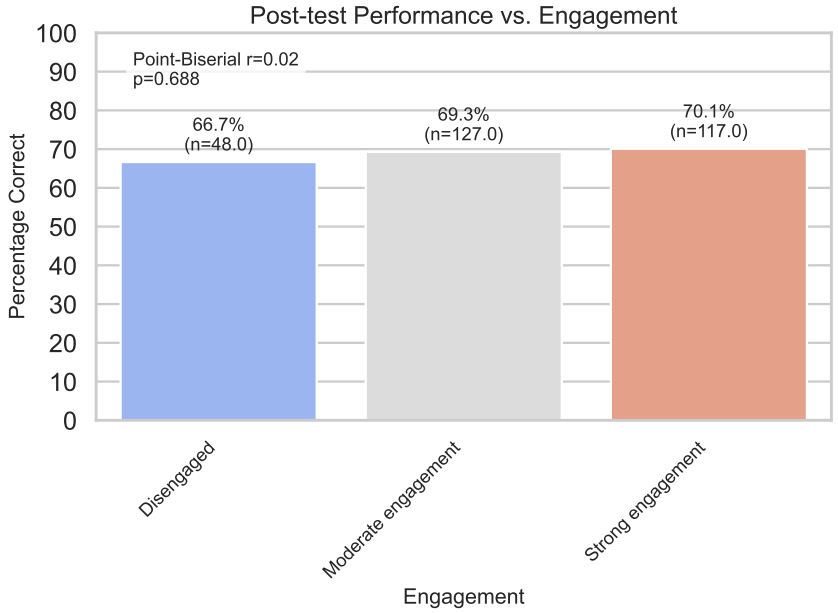

Figure 18: Post-test performance against perceived student engagement per interaction

A.7 Survey Results

A.7.1 Question 1

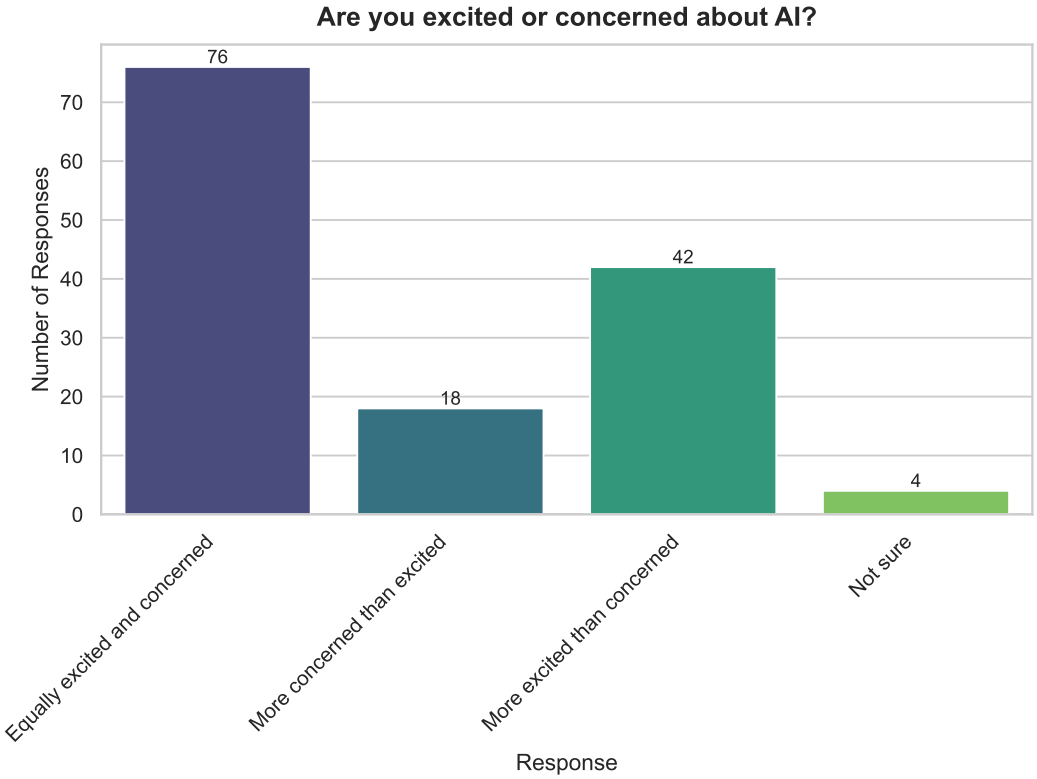

Figure 19: Excitement / concern about AI

A.7.2 QUESTION 2

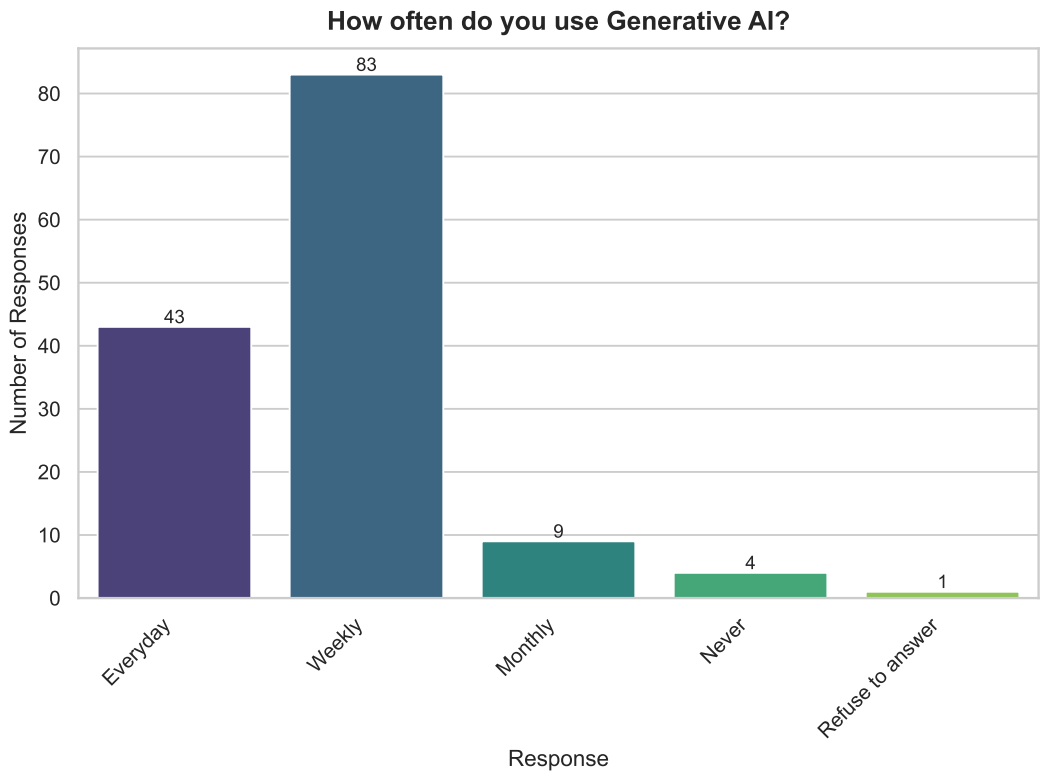

Figure 20: Frequency in use of AI (e.g. ChatGPT / Midjourney)

A.7.3 QUESTION 3

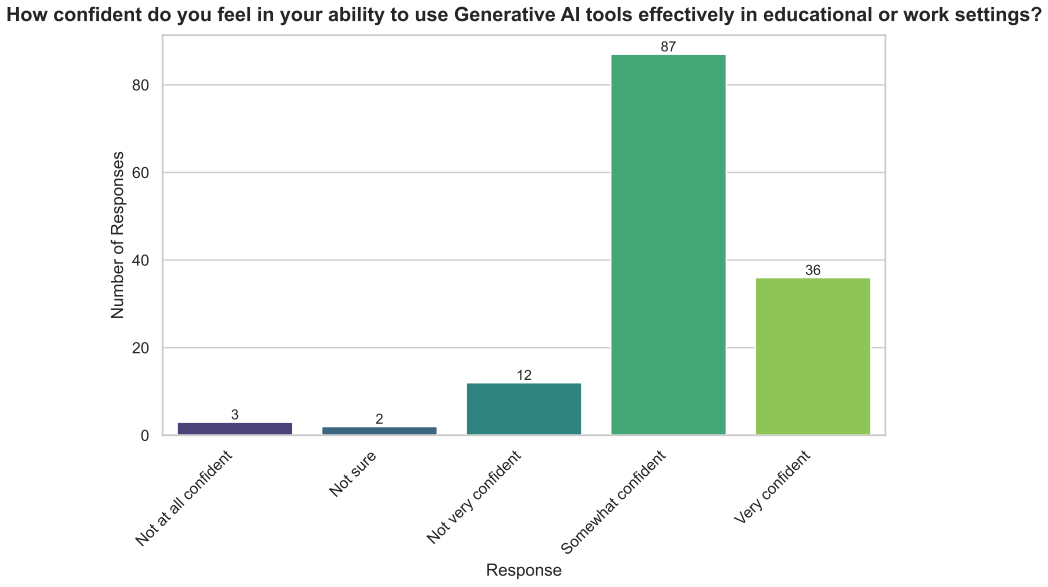

Figure 21: Confidence in use of AI in work/ educational settings

A.7.4 QUESTION 4

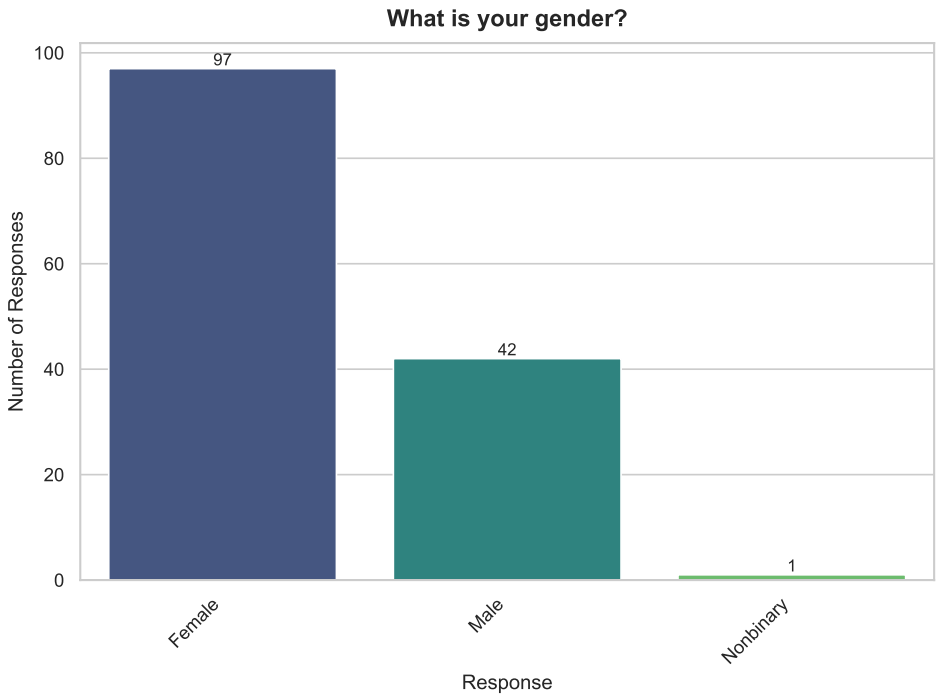

Figure 22: Gender