Math Education With Large Language Models: Peril or Promise?

Harsh Kumar*
University of Toronto
Toronto, Canada
harsh@cs.toronto.edu

David M. Rothschild
Microsoft Research
New York City, United States
david@researchdmr.com

Daniel G. Goldstein
Microsoft Research
New York City, United States
dgg@microsoft.com

Jake M. Hofman
Microsoft Research
New York City, United States
jmh@microsoft.com

ABSTRACT

The widespread availability of large language models (LLMs) has provoked both fear and excitement in the domain of education. On one hand, there is the concern that students will offload their coursework to LLMs, limiting what they themselves learn. On the other hand, there is the hope that LLMs might serve as scalable, personalized tutors. Here we conduct a large, pre-registered experiment involving 1200 participants to investigate how exposure to LLM-based explanations affect learning. In the experiment's learning phase, we gave participants practice problems and manipulated two key factors in a between-participants design: first, whether they were required to attempt a problem before or after seeing the correct answer, and second, whether participants were shown only the answer or were also exposed to an LLM-generated explanation of the answer. Subsequently, all participants were tested on new test questions to assess how well they had learned the underlying concepts. Overall we found that LLM-based explanations positively impacted learning relative to seeing only correct answers. The benefits were largest for those who attempted problems on their own first before consulting LLM explanations, but surprisingly this trend held even for those participants who were exposed to LLM explanations before attempting to solve practice problems on their own. An accompanying qualitative analysis revealed that these boosts in performance were indeed due to participants adopting the strategies they were shown, and that exposure to LLM explanations increased the amount people felt they learned and decreased the perceived difficulty of the test problems.

CCS CONCEPTS

• **Human-centered computing → Laboratory experiments**; *Empirical studies in interaction design*; • **Applied computing → Computer-assisted instruction**.

KEYWORDS

education, large language models, math.

1 INTRODUCTION

Recent advancements in Large Language Models (LLMs) present both potential benefits and challenges for math education. There is the fear that students could use LLMs as substitutes for studying, having these tools do their homework for them at the cost of learning how to do it themselves. However, there is also hope that an LLM could act as a personalized and customizable tutor that can deepen their comprehension of the material. While there is public discussion on this topic, there is little to no empirical evidence as to whether LLMs help or hinder in these ways [35].

In this paper, we present a large, pre-registered experiment on how the use of LLMs affects learning in the domain of mathematics. We use this experiment to investigate the following:

- When doing practice questions for a math test, how does the type of explanation people receive (answers alone or answers with LLM-generated explanations) affect performance on subsequent test questions?
- How does the relationship between explanation type and performance change when people i) attempt questions before seeing explanations or ii) see explanations before attempting questions?

In the experiment, we give participants a round of practice problems and manipulate two key factors in a between-participants design: first, whether they are required to attempt a problem before or after seeing the correct answer, and second, whether participants are shown only the answer or are also exposed to an LLM-generated explanation of the answer. Subsequently, all participants are tested on new test questions to assess how well they have learned the underlying concepts. The primary outcome measure is their performance on this test set.

Overall, we find that LLM-based explanations positively impacted learning (relative to seeing only correct answers), regardless of whether participants consulted them before or after attempting practice problems. Pre-prompting LLMs to act as tutors with customized instructions and strategies may be even more helpful. In the following sections, we look at the related work, followed by more details of the experimental design and results.

2 BACKGROUND AND RELATED WORK

This paper builds upon the longstanding research into education and the much more recent literature on how people use and are affected by LLM-based tools.

2.1 Role of feedback and explanation in learning

Long before LLM-based tools, researchers were exploring methods to improve learning with feedback and explanations [20, 33, 63].

Feedback. Feedback plays a vital role in the learning process as it minimizes the gap between actual and desired knowledge

*Work done during an internship at Microsoft Research.

[7]. Numerous factors influence feedback efficacy, as identified in feedback research [13, 67]. Content is arguably the most critical aspect of any feedback method [37]. Fundamentally, feedback must convey the correctness of a student's response (i.e., correct vs. incorrect), however studies reveal little to no benefit of providing verification feedback compared to no feedback [54, 56, 61]. Incorporating the correct answer in the feedback message significantly enhances feedback efficacy compared to verification feedback alone, as it offers the information students require to rectify their errors [54, 55, 61, 73, 77].

Instructors and researchers often assume that additional information in feedback messages will improve learning. *Elaborative feedback* describes feedback types that are more complex than correct answer feedback [37]. Surprisingly, studies comparing elaborative and correct answer feedback found little or no benefit to increasing feedback message complexity [5]. For instance, *explanation feedback*, which explains why a response is correct or incorrect, showed no benefits compared to correct answer feedback [26, 38, 49, 57, 62, 69, 77]. [12] suggested that these null effects might result from assessing learning on the final test by repeating initial test questions. An experiment involving university students demonstrated that explanation feedback improved performance on new inference questions in a test. In online quizzes with closed questions, [21] demonstrated how elaborate feedback can enhance learning, recommending at least providing explanations for incorrect student answers. [23] indicated that elaborating on correct answer feedback with examples improved conceptual learning. Elaborative feedback is commonly incorporated into instructional methods, such as intelligent tutoring systems [34, 51] and computer-assisted instruction programs [25, 40].

Explanation. In *Explanation and Cognition*, [32] emphasized that explanations are pervasive, emerging from childrens' early developmental questions and extending into all aspects of adult life. Explanations represent one of the most common instructional strategies [42], with applications in classroom teaching [47], human tutoring [16], and learning from texts [17, 28]. Despite their importance, the question of what makes an explanation effective is still an open question in educational research [74, 79].

[79] proposed a framework for evaluating the effectiveness of instructional explanations, defining them as explanations deliberately designed for teaching in educational contexts. By reviewing and synthesizing empirical work across various research fields, they identified four key characteristics of effective explanations: (a) adaptability, (b) focus on concepts and principles, (c) consideration of learners' ongoing cognitive activities, and (d) complementing, rather than replacing, learners' cognitive processes. Additionally, the authors noted that instructors' (i) ability to assess students' understanding [8, 65] and (ii) their epistemological beliefs about teaching [75] might play crucial roles in providing effective explanations. Supporting tutors/instructors in delivering effective explanations has also been addressed [59, 64, 78].

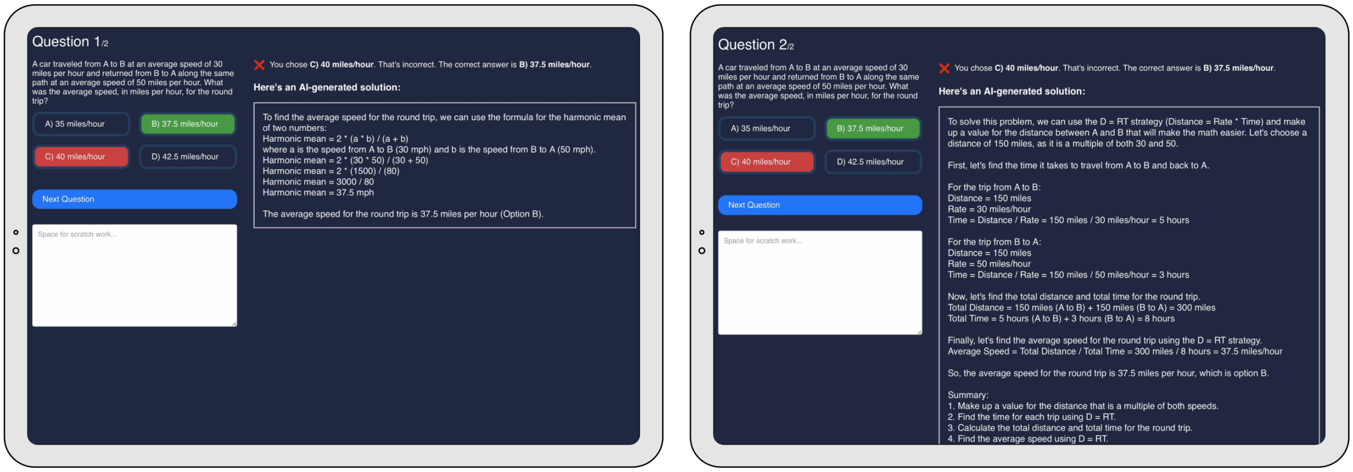
Moreover, several cognitive and non-cognitive factors in learners may negatively impact the effective use of explanations. For instance, [22] suggested that non-cognitive variables such as interest, motivation, and attitude can significantly influence the positive outcomes of instructional explanations. [43] argued that to facilitate learners' processing of explanations, the topic should be problematized to encourage productive engagement, thereby promoting the construction of conceptual knowledge. [79] also emphasized that instructional explanations and their specific functions may vary greatly depending on the context in which they are provided [76]. [36] proposed a framework for effective instructional explanations in scientific teaching. [50] compared two types of explanations, functions (what a thing is for) and mechanisms (how a thing works), and examined the order in which they may be most effective. [1] demonstrated how computer-based learning environments might compel learners to process instructional explanations to address gaps in their understanding.

2.2 LLM performance in standardized tests

Performance of LLMs on Question Answering (QA) tasks have been extensively studied [41, 44]. [3] have shown fine-tuned models, such as Vicuna and ChatGPT, attain near-perfect scores in QA tasks, surpassing models that lack supervised fine-tuning. [68] has shown that Flan-PaLM beats the existing state-of-the-arts in medical QA datasets. In terms of performance on academic and professional exams, GPT-4 has been shown to land in the top 80th percentile for both SAT Math and GRE Quantitative Reasoning [53]. [27] showed the performance of ChatGPT on the United States medical licensing exam and discussed its implications on medical education and assessment. ChatGPT was shown to achieve the equivalent of a passing score for a third-year medical student and provide logical and informational context across majority of the answers.

2.3 Human-AI teaming

Research in human-centered AI has shown when AI systems *explain* their answers, the performance of human-AI team reaches a higher level compared to the performance of human alone [14, 45, 58]. [6] have further shown how human-AI team can achieve *complementary performance*, where it outperforms both the AI and human acting alone in a decision-making task. [72] showed complementarity in a visual classification task with explanations. In a similar visual classification task, [29] found that counterfactual explanations, in addition to training examples, improved performance of people. [15] showed that leveraging AI to discover and teach optimal cognitive strategies could be a promising approach to improve human judgment and decision making. [66] analyzed 5.8 million moves by professional Go players and found that humans began to make significantly better decisions and novel (i.e., previously unobserved) moves following the advent of superhuman AI. In the context of solving multiple-choice questions with support from LLMs, [9] conducted online experiments to show human+LLM team outperforms LLM, which in turn outperforms humans alone for questions drawn from practice tests for exams targeted at high-school, undergraduate and professional students [30]. [39] did a comprehensive survey of empirical studies in human-AI decision making. [52] conducted an online experiment to evaluate the impact of ChatGPT on productivity and other measures related to writing tasks.



Question 1/2

A car traveled from A to B at an average speed of 30 miles per hour and returned from B to A along the same path at an average speed of 50 miles per hour. What was the average speed, in miles per hour, for the round trip?

✕ You chose **C) 40 miles/hour**. That's incorrect. The correct answer is **B) 37.5 miles/hour**.

Here's an AI-generated solution:

To find the average speed for the round trip, we can use the formula for the harmonic mean of two numbers:
Harmonic mean = 2 * (a * b) / (a + b)
where a is the speed from A to B (30 mph) and b is the speed from B to A (50 mph).
Harmonic mean = 2 * (30 * 50) / (30 + 50)
Harmonic mean = 2 * (1500) / (80)
Harmonic mean = 3000 / 80
Harmonic mean = 37.5 mph

The average speed for the round trip is 37.5 miles per hour (Option B).

Question 2/2

A car traveled from A to B at an average speed of 30 miles per hour and returned from B to A along the same path at an average speed of 50 miles per hour. What was the average speed, in miles per hour, for the round trip?

✕ You chose **C) 40 miles/hour**. That's incorrect. The correct answer is **B) 37.5 miles/hour**.

Here's an AI-generated solution:

To solve this problem, we can use the D = RT strategy (Distance = Rate * Time) and make up a value for the distance between A and B that will make the math easier. Let's choose a distance of 150 miles, as it is a multiple of both 30 and 50.

First, let's find the time it takes to travel from A to B and back to A.

For the trip from A to B:
Distance = 150 miles
Rate = 30 miles/hour
Time = Distance / Rate = 150 miles / 30 miles/hour = 5 hours

For the trip from B to A:
Distance = 150 miles
Rate = 50 miles/hour
Time = Distance / Rate = 150 miles / 50 miles/hour = 3 hours

Now, let's find the total distance and total time for the round trip.
Total Distance = 150 miles (A to B) + 150 miles (B to A) = 300 miles
Total Time = 5 hours (A to B) + 3 hours (B to A) = 8 hours

Finally, let's find the average speed for the round trip using the D = RT strategy.
Average Speed = Total Distance / Total Time = 300 miles / 8 hours = 37.5 miles/hour

So, the average speed for the round trip is 37.5 miles per hour, which is option B.

Summary:
1. Make up a value for the distance that is a multiple of both speeds.
2. Find the time for each trip using D = RT.
3. Calculate the total distance and total time for the round trip.
4. Find the average speed using D = RT.

(a) ***Try First*** **+** ***Stock LLM*** **condition.**

(b) ***Try First*** **+** ***Customized LLM*** **condition.**

Figure 1: Example questions from the practice phase in different LLM conditions.

2.4 LLM-based tools for learning

Intelligent Tutoring Systems (ITSs) have a rich history in providing personalized feedback and explanations to students [2]. Advances in generative models have resulted in increased use of these models in ITSs [4, 10, 18, 46, 80]. Through a commentary, Kasneci et al. [31] highlighted the opportunities and challenges of LLMs for education. Tack et al. [71] proposed an AI "teacher" test to estimate the pedagogical abilities of generative models. In a comparative study of responses generated by a human teacher, GPT-3 [11], and Blender [60], the authors found that human performance exceeds both models when it comes to helping the student. Within the context of computing education, Macneil et al. [48] compared three different types of explanations of a code snippet, generated using Codex, with respect to the usefulness of these explanation types.

This paper builds upon the existing research on the role of feedback and explanation in learning, intelligent tutors and human-AI collaboration. We probe the design space of using LLMs for learning by manipulating the sequence in which learners engage with content - either by attempting to solve problems before accessing LLM-generated solutions or vice versa. Our approach provides a unique understanding of how the timing of feedback can play a role in AI-augmented learning, offering fresh perspectives on the integration of LLMs as educational aids.

3 EXPERIMENTAL DESIGN

We conducted a pre-registered[1] online experiment that consisted of two phases: a *practice* phase and a *test phase*. As depicted in Figure 2, it was a 2x3 between subjects design with randomization in the practice phase. During the *practice* phase, participants were presented with questions that were designed to mirror the math section of the Scholastic Assessment Test (SAT). They were randomly assigned to either attempt these questions without any help (*Try First*) or view the answers before attempting them (*See Answer First*). Additionally, they received one of three types of explanations:

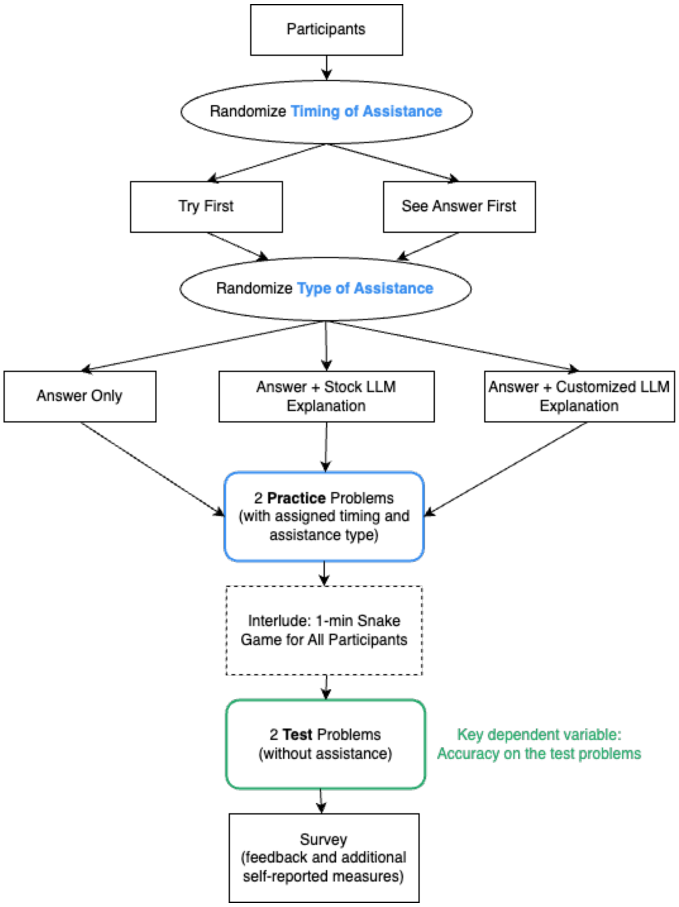

Participants
Randomize Timing of Assistance
Try First
See Answer First
Randomize Type of Assistance
Answer Only
Answer + Stock LLM Explanation
Answer + Customized LLM Explanation
2 Practice Problems (with assigned timing and assistance type)
Interlude: 1-min Snake Game for All Participants
2 Test Problems (without assistance)
Key dependent variable: Accuracy on the test problems
Survey (feedback and additional self-reported measures)

Figure 2: Schematic of the experiment design.

[1] https://aspredicted.org/H34_SZX

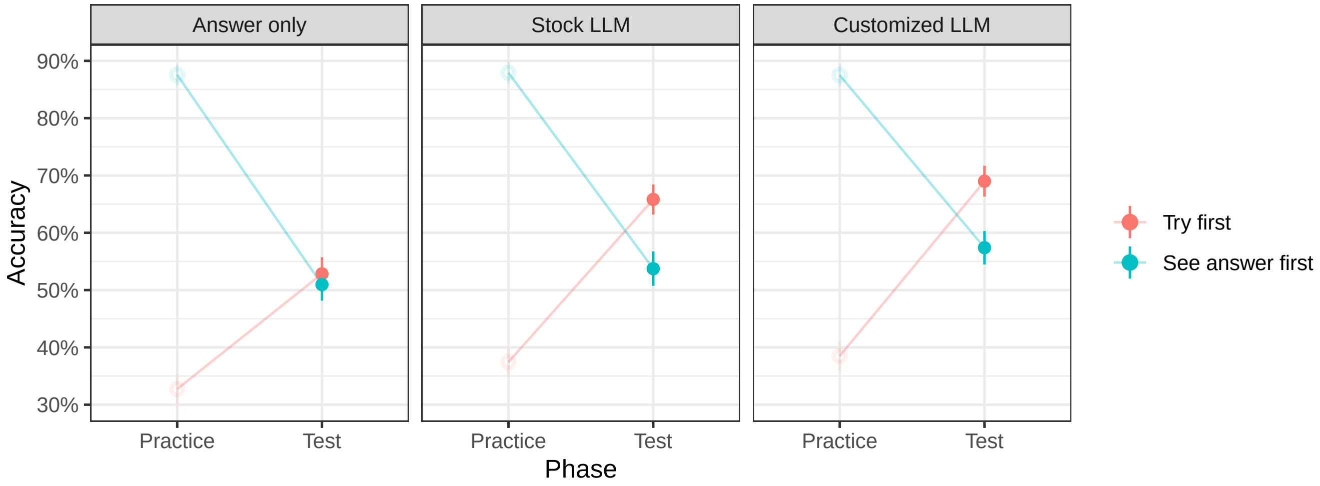


Figure 3: Accuracy by condition. Empty light circles show accuracy in the practice phase and solid dark circles show accuracy in the test phase. Error bars show one standard error above and below these averages.

- *None*: Only the correct answer was provided.
- *Stock LLM*: Correct answers were accompanied by a GPT-4 generated explanation.
- *Customized LLM*: Correct answers were accompanied by explanations generated by GPT-4 with a specialized pre-prompt. The pre-prompt, which was hidden from particiapnts, included customized problem solving strategies (see Figure 6 in the Appendix).

After completing the *practice* phase, participants took a short break by playing a game of Snake in their browser for one minute. This served as a distractor, allowing some level of forgetting regarding the problems and solutions they just encountered.

In the *test phase*, participants were given slightly altered versions of the same types of questions they faced during the *practice* phase. The wording remained the same, but the numbers used in the questions were changed. These test questions were presented without any assistance or feedback, and the set of numbers shown in the practice versus test phases was randomized for each participant. After the test phase, participants were asked to report how difficult they found the test questions (very easy, somewhat easy, somewhat difficult, very difficult), how many of the two test questions they thought they got right, how much they learned from the practice problems (nothing, a little, a lot). Participants also provided a free-text response with a few sentences about the strategy they used to answer each of the questions from the test phase.

3.1 Domain and Stimuli

The questions the participants saw were designed to mirror material from the math section of the Scholastic Assessment Test (SAT). Each participant solved two out of four possible types of questions. The four distinct types of questions for our experiment were:

- Calculating the average speed of a multi-segment trip.
- Solving for two unknowns with two constraints.
- Determining the parity of an algebraic expression.
- Identifying the missing measurement from an average.

Figure 1 shows an example question (calculating the average speed of a multi-segment trip) from the practice phase for the different LLM conditions. We used GPT-4 to generate all explanations for the LLM conditions. For the *Customized LLM* condition, we used a hidden pre-prompt (Figure 6) with customized problem-solving strategies based on best practices for standardized test tutoring.

3.2 Participants

We recruited 1,202 (99% of whom passed an attention check) participants from Amazon Mechanical Turk.[2] We chose this sample size based on the results of previous pilots so that we had approximately 80% power in detecting differences between conditions at a level of significance of 5%. Participants were randomly assigned to one of the six conditions. Each participant received a flat payment of $3.30.

4 RESULTS

4.1 Measure of Accuracy on the Test Problems

Following our pre-registration, we used a mixed-effects logistic regression to model each participant's selected option for each question in the test phase, with main effects for and an interaction between *order* and *explanation type*, random effects for participants, and fixed effects for question type as a control. We report the results of the pre-planned contrasts below.

As depicted in Figure 3, in the practice round participants in the *See Answer First* condition who were shown the correct answer before responding did better than those in the *Try First* condition. However, this trend was reversed in the test phase: participants who attempted practice problems before viewing the solution (solid

[2] In education research, previous studies have shown that MTurk responses offer a quality comparable to traditional methods, including samples from undergraduate students and communities [24]. Crowdworkers have been shown to be reliable proxies for online learners [19].

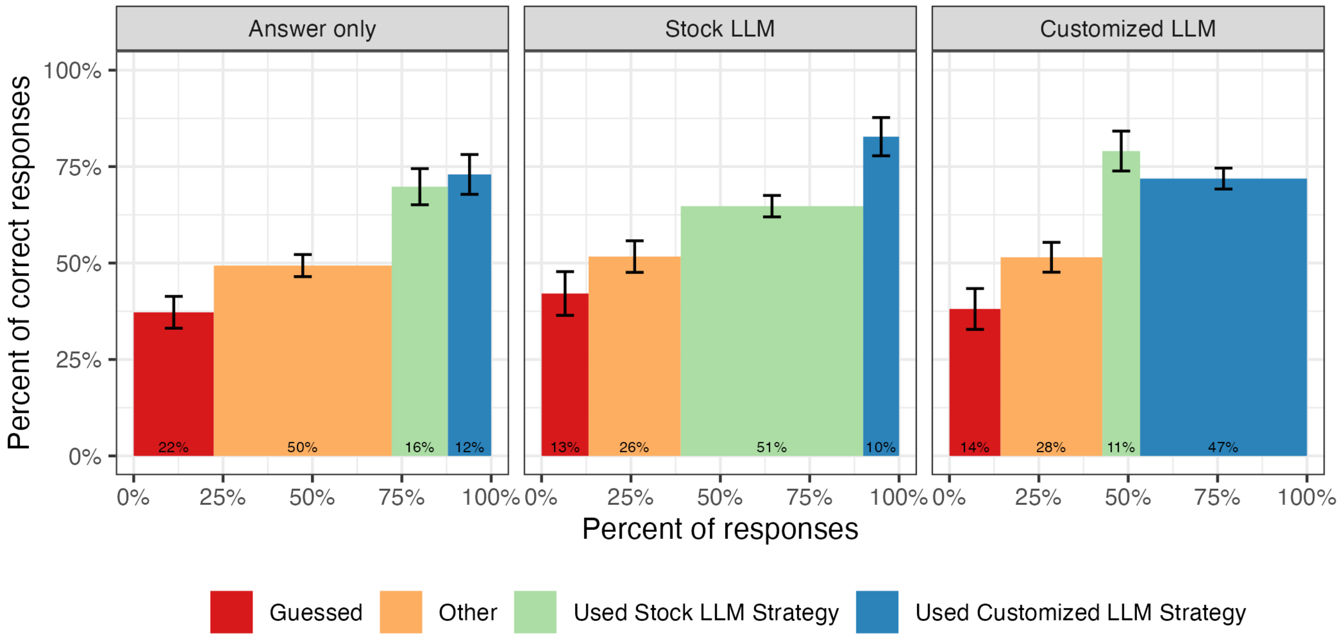


Figure 4: Strategies used by participants for solving test problems. The width of each rectangle (annotated on the bottom of each rectangle) shows the percent of responses that used each strategy in each condition, while the height shows the corresponding average accuracy when that strategy was used. Error bars show one standard error above and below these averages. Increased performance is largely due to participants adopting the strategies they were exposed to by the LLM.

red points) performed better in the test phase compared to those who saw the answers prior to attempting questions in the practice phase (solid blue points). Furthermore, we see that within the *Try First* condition, there were substantial benefits to both the *Stock* and *Customized LLM* explanations over seeing just the correct answer—while slightly more than half of test question responses were correct in the *Answer only* condition, more than two-thirds of responses were correct when participants were shown LLM explanations ($z = -3.46, p < 0.001$ for *Answer only* vs. *Stock LLM*; $z = -4.20, p < 0.001$ for *Answer only* vs. *Customized LLM*).

In the *See Answer First* condition, we see that exposure to the customized LLM results in a higher accuracy compared to seeing the correct answer only ($z = -1.67, p = 0.05$), but find no statistically significant lift from the stock LLM compared to just seeing the correct answer (z = -0.68, p = 0.25). Interestingly, comparing performance in the *See Answer First* + *Customized LLM* (on the far right) to *Try First* + *Answer only* (on the far left), we find no statistically significant difference between conditions ($z = -1.15, p = 0.12$).

4.2 Strategies Followed by the Participants

After solving the test problems we asked participants to describe the strategy they used for each problem in their own words. We did this to learn more about whether they adopted the strategies they were shown in the practice rounds, and if so how effective different strategies were in improving performance. To investigate this we used an iterative coding scheme to categorize each of the free text responses. First, we manually labeled a sample of 200 responses to determine the set of strategies that participants tried for each question type and whether they explicitly mentioned following the strategy provided in the practice round or not. For instance, for the average speed calculation problem shown in Figure 1, strategy labels included the *Stock LLM*'s "harmonic mean" approach, the *Customized LLM*'s "make up a distance" strategy, and some alternative strategies that were not directly shown to participants (e.g., "algebra with D=RT", which can be time consuming but yields a correct response, and "averaging speeds", which results in an incorrect answer). We also included labels for when participants indicated that they had simply guessed. Then we created a few shot prompt for GPT-4 with examples of each strategy label for each question type and ran each response through the GPT-4 API to generate labels at scale for all 2400 responses (2 responses for each of the 1200 participants). To check the quality of these labels, we verified that the GPT-generated labels agreed with the manual labels on the sample of 200 responses that we coded ourselves.

With strategy labels generated for each question type, we analyzed which strategies participants adopted and the efficacy of these strategies, aggregated across questions, as shown in Figure 4. Specifically, we looked at whether participants followed one of the strategies recommended by the LLMs, if they used an alternative strategy, or if they simply guessed. In the *Answer only* condition (the far left facet), the majority of participants resorted to guessing or using other alternative strategies, resulting in an average accuracy below 50%. This is visually represented by the predominant red and orange bars. In contrast, when examining the *Stock LLM*

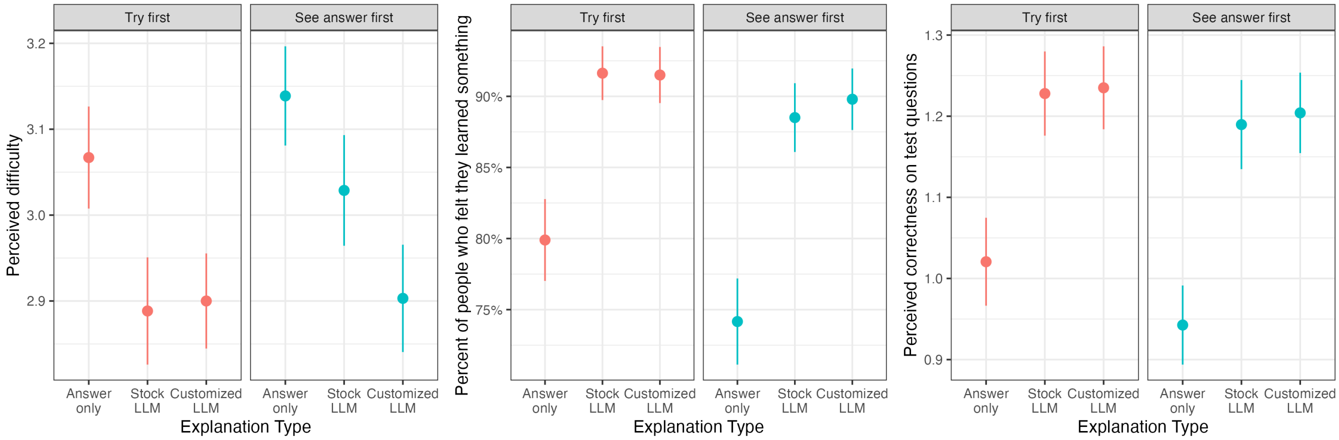


Figure 5: Plots of participants' self-reported perception and performance. The left panel shows perceived difficulty ratings (on a 4-point likert scale), the middle panel indicates the percentage of participants who felt they learned something, and the right panel presents the perceived correctness on test questions (either 0, 1, or 2), all segmented by *Try First* and *See Answer First* conditions, and across three explanation types: *Answer only*, *Stock LLM*, and *Customized LLM*.

and *Customized LLM* conditions (middle and right facets, respectively), we see a significant shift. We see substantially less guessing and use of alternative strategies, indicated by the smaller red and orange regions. Instead, there is a notable increase in the blue and green bars, showing that participants predominantly adopted the strategies aligned with the LLM explanations they were shown. Furthermore, performance for those who adopted either of these strategies is markedly higher than for those guessed or used alternative approaches, with average accuracies under these strategies ranging from 60-80%.

4.3 Post-task Survey Measures

Figure 5 presents participants' perceptions and self-reported performance across different experimental conditions. In the *Try First* condition with *Stock LLM* and *Customized LLM* explanations, participants rated the test problems with average difficulty of around 2.9 on a 4-point Likert scale (ranging from "very easy" to "very difficult"), which is lower than the average difficulty rating for those who had *Answer only* support (the far left facet). In the *See answer first* condition, participants in the *Customized LLM* condition found the test problems to be less difficult compared to those in the *Stock LLM* condition. Moving to the middle facet which shows the percentage of participants who felt they learned something from the practice problems, we see over 85% participants, irrespective of timing of assistance, in *Try first* and *See answer first* conditions who felt they learned something. This is substantially higher than the participants in the *Answer only* conditions. We see a similar trend when asked about the perceived correctness (how many of the 2 test questions the participant thought they got right). The participants in *Stock LLM* and *Customized LLM* reported solving higher number of test questions correctly compared to participants in *Answer only* condition (the far right facet).

These responses suggest that participants not only found the test problems to be less difficult when supported by LLM explanations but also felt more confident about what they had learned and their test performance.

5 DISCUSSION

Through a large, pre-registered experiment we found that LLM-based explanations positively impacted learning relative to seeing only correct answers. The benefits were largest for those who attempted problems on their own first before consulting LLM explanations: while slightly more than half of test question responses were correct for those who weren't exposed to LLM explanations, more than two-thirds of responses were correct for those who saw LLM explanations. Surprisingly this trend held even for those participants who were exposed to LLM explanations before attempting to solve practice problems on their own, although the effects were less pronounced. An accompanying qualitative analysis revealed that these boosts in performance were indeed due to participants adopting the strategies they were shown, and that exposure to LLM explanations increased the amount people felt they learned and decreased the perceived difficulty of the test problems.

Taken together, these results hold promise for the use of LLMs as personal tutors and educational tools. That said, there are of course several limitations to this work and many areas for future investigation. First, GPT-4 provided correct answers and coherent explanations for the question types that we examined in this study, but this is far from guaranteed. LLMs are known to make errors and provide incorrect explanations, which could have negative effects in a tutoring scenario if students are unaware of these limitations. Future work could investigate issues of overreliance for LLM tutoring, similar to work that has been done in the domain of LLM-based information retrieval [70].

Second, we have examined just one domain (SAT math problems) with one response format (multiple choice). A simple extension of this work would be to repeat the same questions but with free response answers instead of multiple choice, which might reveal even larger differences between conditions. For example, for the multi-segment trip problem in Figure 1, we found that many participants in the *Answer only* condition learned to "game" the multiple choice format by simply choosing the answer that was just lower in magnitude than the arithmetic average. This produced a correct answer, but, as many participants who took this approach wrote in their qualitative description of the strategy they used, these participants did not necessarily know why this was the case. A free response format could provide additional insights. Likewise, we conducted this study using math problems because the underlying strategies can be learned fairly quickly, it is relatively straightforward to test generalization (e.g., by simply changing the numbers in a practice problem to create a test problem), and it is easy to objectively score responses as correct or incorrect. We have already started to develop parallel studies in other domains (e.g., reading comprehension). While these have proven more difficult to construct clean experimental designs for, our pilot studies show promise for potential benefits of LLM assistance in the learning process.

Third, we focused on short-term learning and retention in this experiment. While the problems we used were difficult enough for participants that even a short break between practice and testing surfaced differences in performance across conditions, it would be interesting to see how differences play out on a longer timescale. For instance, for the multi-segment trip problem, one could imagine that participants in the *Stock LLM* condition were able to remember the harmonic mean formula for a few minutes but might not retain it weeks or months later. In contrast, the *Customized LLM* strategy of assigning a number to the unspecified distance in the problem might be easier to retain and apply in other settings over time. Likewise, it would be interesting to see how LLM assistance affects learning more holistically, when applied over the course of an entire semester, for instance, instead of just for a given set of problems.

Finally, we conducted these experiments with a set of high-quality, well-intentioned participants who were randomized into either being exposed to LLMs or not in a relatively confined mode of interaction. There are many differences outside of the lab setting that could produce different outcomes from what we find here. Most obviously, perhaps, is that less well-intentioned participants (e.g., students who are simply looking for the quickest way to turn in a homework assignment) might simply copy and paste LLM solutions without reflecting on or trying to learn from them. This could, of course, lead to worse outcomes across the board. Or it could be the case that there are interesting effects coupled to when students do (or don't) make the decision to use LLMs for assistance—perhaps students have a good sense for when LLMs will (and won't) help in learning new material, or maybe this is itself a skill to be learned. The same could extend to how students interact with LLMs. Whereas we simply showed students pre-generated LLM explanations, the ability to engage with an LLM interactively to learn new material could serve to further enhance learning. For example, explanations could be tailored to students mistakes, or provide an opportunity to further probe the details of a problem or related aspects of a subject.

In closing, we hope this work offers a template for identifying when LLMs should (or shouldn't) be used in education, and for highlighting the ways in which we can design and use LLMs to benefit both students and teachers.

ACKNOWLEDGMENTS

Thanks to Mark Whiting and Duncan Watts for their help in recruiting participants.

REFERENCES

[1] Vincent AWMM Aleven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science* 26, 2 (2002), 147–179.

[2] John R Anderson, C Franklin Boyle, and Brian J Reiser. 1985. Intelligent tutoring systems. *Science* 228, 4698 (1985), 456–462.

[3] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. *arXiv preprint arXiv:2306.04181* (2023).

[4] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484* (2023).

[5] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of educational research* 61, 2 (1991), 213–238.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[7] Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice* 5, 1 (1998), 7–74.

[8] Hilda Borko, Vicky Mayfield, Scott Marion, Roberta Flexer, and Kate Cumbo. 1997. Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher education* 13, 3 (1997), 259–278.

[9] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540* (2022).

[10] Anja Boxleitner. 2023. Integrating AI in Education: Opportunities, Challenges and Responsible Use of ChatGPT. *Education: Opportunities, Challenges and Responsible Use of ChatGPT (September 9, 2023)* (2023).

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[12] Andrew C Butler, Namrata Godbole, and Elizabeth J Marsh. 2013. Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology* 105, 2 (2013), 290.

[13] Deborah L Butler and Philip H Winne. 1995. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research* 65, 3 (1995), 245–281.

[14] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[15] F. Callaway, Y. Jain, B. Opheusden, P. Das, G. Iwama, S. Gul, P. Krueger, F. Becker, T. Griffiths, and F. Lieder. 2022. leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences* 119 (2022). Issue 12. https://doi.org/10.1073/pnas.2117432119

[16] Michelene TH Chi, Stephanie A Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G Hausmann. 2001. Learning from human tutoring. *Cognitive science* 25, 4 (2001), 471–533.

[17] Elaine B Coleman, Ann L Brown, and Inna D Rivkin. 1997. The effect of instructional explanations on learning from scientific texts. *The Journal of the Learning Sciences* 6, 4 (1997), 347–365.

[18] Tyler Cowen and Alexander T Tabarrok. 2023. How to learn and teach economics with large language models, including GPT. *Including GPT (March 17, 2023)* (2023).

[19] Dan Davis, Claudia Hauff, and Geert-Jan Houben. 2018. Evaluating Crowdworkers as a Proxy for Online Learners in Video-Based Learning Contexts. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 42 (nov 2018), 16 pages. https://doi.org/10.1145/3274311

[20] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest* 14, 1 (2013), 4–58.

[21] Natalie Enders, Robert Gaschler, and Veit Kubik. 2021. Online quizzes with closed questions in formal assessment: How elaborate feedback can promote learning.

Psychology Learning & Teaching 20, 1 (2021), 91–106.

[22] Randi A Engle and Faith R Conant. 2002. Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and instruction* 20, 4 (2002), 399–483.

[23] Bridgid Finn, Ruthann Thomas, and Katherine A Rawson. 2018. Learning more from feedback: Elaborating feedback with examples enhances concept learning. *Learning and Instruction* 54 (2018), 104–113.

[24] D Jake Follmer, Rayne A Sperling, and Hoi K Suen. 2017. The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher* 46, 6 (2017), 329–334.

[25] Andrew S Gibbons and Peter G Fairweather. 1998. *Computer-based instruction: Design and development.* Educational Technology.

[26] David A Gilman. 1969. Comparison of several feedback methods for correcting errors by computer-assisted instruction. *Journal of Educational Psychology* 60, 6p1 (1969), 503.

[27] A. Gilson, C. Safranek, T. Huang, V. Socrates, L. Chi, R. Taylor, and D. Chartash. 2023. how does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *Jmir medical education* 9 (2023), e45312. https://doi.org/10.2196/45312

[28] Susan R Goldman. 2004. Cognitive aspects of constructing meaning through and across multiple texts. *Uses of intertextuality in classroom and educational research* 317352 (2004).

[29] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2376–2384. https://proceedings.mlr.press/v97/goyal19a.html

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[31] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.

[32] Frank C Keil and Robert Andrew Wilson. 2000. *Explanation and cognition.* MIT press.

[33] Paul Kirschner and Carl Hendrick. 2020. *How learning happens: Seminal works in educational psychology and what they mean in practice.* Routledge.

[34] Kenneth R Koedinger, Albert Corbett, et al. 2006. *Cognitive tutors: Technology bringing learning sciences to the classroom.* na.

[35] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Unreflected Acceptance–Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. *arXiv preprint arXiv:2309.03087* (2023).

[36] Christoph Kulgemeyer. 2018. Towards a framework for effective instructional explanations in science teaching. *Studies in Science Education* 54, 2 (2018), 109–139.

[37] Raymond W Kulhavy and William A Stock. 1989. Feedback in written instruction: The place of response certitude. *Educational psychology review* 1 (1989), 279–308.

[38] Raymond W Kulhavy, Mary T White, Bruce W Topp, Ann L Chan, and James Adams. 1985. Feedback complexity and corrective efficiency. *Contemporary educational psychology* 10, 3 (1985), 285–291.

[39] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[40] Jill H Larkin and Ruth W Chabay. 2021. *Computer assisted instruction and intelligent tutoring systems: Shared goals and complementary approaches.* Routledge.

[41] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. arXiv:2305.18486 [cs.CL]

[42] Gaea Leinhardt et al. 2001. Instructional explanations: A commonplace for teaching and location for contrast. *Handbook of research on teaching* 4 (2001), 333–357.

[43] Gaea Leinhardt and Michael D Steele. 2005. Seeing the complexity of standing to the side: Instructional dialogues. *Cognition and Instruction* 23, 1 (2005), 87–163.

[44] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[45] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[46] Muneer M Alshater. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. *Available at SSRN* (2022).

[47] Liping Ma. 2010. Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.

[48] Stephen MacNeil, Andrew Tran, Arto Hellas, Joanne Kim, Sami Sarsa, Paul Denny, Seth Bernstein, and Juho Leinonen. 2023. Experiences from Using Code Explanations Generated by Large Language Models in a Web Software Development E-Book. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 931–937. https://doi.org/10.1145/3545945.3569785

[49] B Jean Mandernach. 2005. Relative effectiveness of computer-based and human feedback for enhancing student learning. *The Journal of Educators Online* 2, 1 (2005), 1–17.

[50] Amanda M McCarthy and Frank C Keil. 2023. A right way to explain? function, mechanism, and the order of explanations. *Cognition* 238 (2023), 105494.

[51] Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments* 29, 1 (2021), 142–163.

[52] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. https://doi.org/10.1126/science.adh2586 arXiv:https://www.science.org/doi/pdf/10.1126/science.adh2586

[53] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[54] Harold Pashler, Nicholas J Cepeda, John T Wixted, and Doug Rohrer. 2005. When does feedback facilitate learning of words? *Journal of experimental psychology: Learning, Memory, and Cognition* 31, 1 (2005), 3.

[55] Gary D Phye and Cheryl E Sanders. 1994. Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology* 19, 3 (1994), 286–301.

[56] Letha Plowman and James B Stroud. 1942. Effect of informing pupils of the correctness of their responses to objective test questions. *The Journal of Educational Research* 36, 1 (1942), 16–20.

[57] Doris R Pridemore and James D Klein. 1995. Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology* 20, 4 (1995), 444–450.

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[59] Julian Roelle, Kirsten Berthold, and Alexander Renkl. 2014. Two instructional aids to optimise processing and learning from instructional explanations. *Instructional Science* 42 (2014), 207–228.

[60] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* (2020).

[61] WJ Roper. 1977. Feedback in computer assisted instruction. *Programmed learning and educational technology* 14, 1 (1977), 43–49.

[62] Julius M Sassenrath and Charles M Garverick. 1965. Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology* 56, 5 (1965), 259.

[63] R Keith Sawyer. 2005. *The Cambridge handbook of the learning sciences.* Cambridge University Press.

[64] Alan H Schoenfeld. 2010. How and why do teachers explain things the way they do? *Instructional explanations in the disciplines* (2010), 83–106.

[65] Lorrie Shepard. 2001. The role of classroom assessment in teaching and learning. (2001).

[66] Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths. 2023. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences* 120, 12 (mar 2023). https://doi.org/10.1073/pnas.2214840120

[67] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.

[68] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* (2023), 1–9.

[69] Marieke HSB Smits, Jo Boon, Dominique MA Sluijsmans, and Tamara Van Gog. 2008. Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments* 16, 2 (2008), 183–193.

[70] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. *arXiv preprint arXiv:2307.03744* (2023).

[71] Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. arXiv:2205.07540 [cs.CL]

[72] Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *Advances in Neural Information Processing Systems* 35 (2022), 34287–34301.

[73] Robert MW Travers, Keith R Van Wagenen, Danielle H Haygood, and Mary McCormick. 1964. Learning as a consequence of the learner's task involvement under different conditions of feedback. *Journal of Educational Psychology* 55, 3 (1964), 167.

[74] DF Treagust and AG Harrison. 1999. The genesis of effective scientific explanations for the classroom. *Researching teaching: Methodologies and practices for understanding pedagogy* (1999), 28–43.

[75] Jan H Van Driel, Douwe Beijaard, and Nico Verloop. 2001. Professional development and reform in science education: The role of teachers' practical knowledge. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 38, 2 (2001), 137–158.

[76] Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction* 21, 3 (2003), 209–249.

[77] Michael M Whyte, Dolores M Karolick, Milton C Nielsen, Gregory D Elder, and W Thomas Hawley. 1995. Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research* 12, 2 (1995), 195–203.

[78] Jörg Wittwer, Matthias Nückles, Nina Landmann, and Alexander Renkl. 2010. Can tutors be supported in giving effective explanations? *Journal of Educational Psychology* 102, 1 (2010), 74.

[79] Jörg Wittwer and Alexander Renkl. 2008. Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist* 43, 1 (2008), 49–64.

[80] Xiaoming Zhai. 2022. ChatGPT user experience: Implications for education. *Available at SSRN 4312418* (2022).

A APPENDIX

You are a tutor designed to help people understand and perform better on the types of problems you are given.

When solving problems where there are unknown numbers, try making up a values for those number to simplify the math. Choose these numbers so that any subsequent arithmetic works out nicely (e.g., by choosing numbers that are whole number multiples of all numbers mentioned in the problem to avoid decimal points or fractions in subsequent divisions).

Here are some strategies to use when arriving at and explaining answers, if applicable:

- When solving problems that involve speeds, emphasize the "D = RT" strategy for "Distance = Rate * Time"
- When solving problems that involve averages (other than speeds), emphasize the "T = AN" strategy for "Total = Average * Number"
- When given a problem with two unknown numbers involving their sum and difference, suggest the strategy of starting with half their sum and distributing the difference

Explain how to solve the given problem in a way which makes it easiest for the learner to understand and remember so that they are able to apply it to a similar problem in future. Avoid solutions that require memorization of concepts and avoid complex notation. The learner should be able to develop some intuition to solve similar problems.

In the end, summarize the solution in minimal number of lines so that the learner is able to remember the method in future.

Figure 6: The pre-prompt for the Customized LLM (not visible to participants).