

Can Large Language Models Generate Middle School Mathematics Explanations Better than Human Teachers?

First Author's Name, Initials, and Last name*

First author's affiliation, an Institution with a very long name, xxxx@gmail.com

Second Author's Name, Initials, and Last Name

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Third Author's Name, Initials, and Last Name

Third author's affiliation, possibly the same institution, xxxx@gmail.com

The development and measurable improvements in performance of large language models on natural language tasks opens the opportunity to utilize large language models in an educational setting to replicate human tutoring, which is often costly and inaccessible. We are particularly interested in large language models from the GPT series, created by OpenAI. In the original study we found that the quality of explanations generated with GPT-3.5 was poor, where two different approaches to generating explanations resulted in a 43% and 10% success rate. In a replication study, we were interested in whether the measurable improvements in GPT-4 performance led to a higher rate of success for generating valid explanations compared to GPT-3.5. A replication of the original study was conducted by using GPT-4 to generate explanations for the same problems given to GPT-3.5. Using GPT-4, explanation correctness dramatically improved to a success rate of 94%. We were further interested in evaluating if GPT-4 explanations were positively perceived compared to human-written explanations. A preregistered, follow-up study was implemented where 10 evaluators were asked to rate the quality of randomized GPT-4 and teacher-created explanations. Even with 4% of problems containing some amount of incorrect content, GPT-4 explanations were preferred over human explanations.

CCS CONCEPTS • Applied Computing ~ Education • Distance learning • Computer-assisted instruction

Additional Keywords and Phrases: Large Language Models, Online Tutoring

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium

* Place the footnote text for the author (if applicable) here.

1 INTRODUCTION

In the past few years, significant advancements have been made in the field of natural language processing, where AI systems are becoming increasingly more capable of replicating human-like text [4]. Such abilities introduce the possibility of utilizing large language models (LLMs) to aid in the development of intelligent tutoring systems, particularly those that help students through text-based explanations of concepts [9]. One-on-one human tutoring can be expensive to expand to meet increasing student demands. Therefore, utilizing automated systems to mimic such interactions can greatly increase accessibility to academically struggling students.

In this work, we explore the effectiveness of using LLMs to create explanations of mathematics problems for students within the ASSISTments online learning platform [3]. Recent transformer-based LLMs have exhibited breakthrough performance on a number of domains [13, 14]. In this work, we perform experiments using some of the most powerful currently available LLMs, GPT-3.5 and GPT-4, accessed through OpenAI’s API.

In the original study, two different approaches that were used to generate content were explored. The first approach used few-shot learning [16] to generate new explanations from a handful of similar mathematics problems with answers and explanations, and the second approach attempted to generate new explanations by using the LM to summarize message logs between students and real human tutors. After each method was used to generate new explanations, these explanations were compared to existing explanations in the ASSISTments online learning platform through surveys given to mathematics teachers. Comparing teachers’ evaluations of the quality of the various explanations enabled an empirical evaluation of each LLM-based approach, as well as an evaluation of their applicability in a real-world setting.

With the continuing development of GPT models, it is highly plausible that newer iterations of the model perform better at the task of generating mathematical explanations, even though they have not been specifically fine-tuned for the task [1]. We investigated, through a replication study, whether GPT-4 can outperform its predecessor model, GPT-3.5, in the task of generating explanations for middle school math problems. Additionally, we conducted a second, larger follow-up study to compare explanations generated by GPT-4 and existing teacher explanations. The ability for GPT-4 to generate high quality explanations opens up the possibility of using LLMs to help expand tutoring systems to include more content without sacrificing teacher time, resources, and at a much lower cost. If analysis results conclude that GPT-4 explanations are favorable and accurate, this will open up numerous opportunities for large-scale utilization of GPT-4 in tutoring systems.

To reiterate, this work addresses the following research questions:

- 1) Original Study:
 - a. What is the most effective way to use GPT-3.5 to create explanations from existing mathematics problems and their answers and explanations?
 - b. What is the most effective way to use GPT-3 to create explanations from chat logs between students and tutors?
 - c. How effective are the methods employed with GPT-3.5 compared to human-written explanations?
- 2) Replication Study: How effective are GPT-4 explanations compared to explanations generated by GPT-3.5?
- 3) Follow-up Study: How effective are GPT-4 generated explanations compared to human-written explanations?

2 BACKGROUND

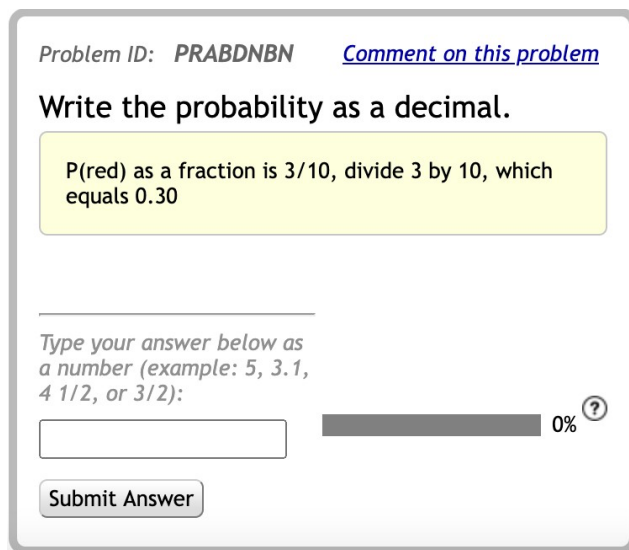
2.1 GPT-3.5 vs. GPT-4

Notable improvements have been researched in GPT-4 compared to GPT-3.5 for natural language tasks, such as completing the uniform bar exam. Compared to GPT-3.5, which scored in the 10th percentile, GPT-4 passed the test and scored in the 90th percentile [6]. The improvements to the GPT-4 model for solving math problems are particularly significant to the task of generating explanations. When comparing model performance on the GSM-8K data set, a common benchmark used to evaluate language models comprised of 8.5 thousand grade school math problems [2], GPT-4 performed significantly better than GPT-3.5. Using 5-shot chain-of-thought prompting, GPT-4 answered 92.0% correctly, compared to a 57.1% accuracy rate when a 5-shot approach was used with GPT-3.5 [6]. The notable increase in accuracy that GPT-4 has over GPT-3.5 provides solid reasoning that the model will show improvement in generating math explanations for middle school math problems.

Compared to a single prompt given to GPT-3.5, prompting GPT-4 requires an additional system prompt, which describes what GPT-4 is responding as, and an additional prompt written as a user interacting with the system [8], which GPT-4 responds to.

2.2 ASSISTments

ASSISTments is an online learning platform focusing on K-12 mathematics [3]. Within the platform, teachers assign problem sets, which their students complete. As students complete problems, they can request an explanation (shown in Figure 1, which reveal the correct answer and explain how to solve the mathematics problem. Currently, all explanations are written by expert teachers, which guarantees a high level of correctness but is also time-consuming and resource-heavy. Manual creation of explanations also limits the scalability of ASSISTments to more curricula. As such, this work aims to investigate whether the quality of GPT-4 explanations can supplement the process of teachers generating new explanations.



The screenshot shows a user interface for a math problem explanation. At the top, it displays "Problem ID: PRABDNBN" and a link "Comment on this problem". The main instruction is "Write the probability as a decimal." Below this, a yellow box contains the explanation: "P(red) as a fraction is 3/10, divide 3 by 10, which equals 0.30". Underneath the explanation, there is a text input field with the prompt "Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):". To the right of the input field is a progress indicator showing "0%" and a question mark icon. At the bottom left, there is a "Submit Answer" button.

Figure 1: Example of a problem explanation in ASSISTments from the student perspective.

2.3 Live Online Tutoring

The data used to generate explanations from summaries of tutoring chat logs comes from UPchieve, a provider of online tutoring, in partnership with ASSISTments. Organizations like Yup (yup.com), Tutor.com, and UPchieve (upchieve.com) offer online tutoring to students. Typically, students use these services by logging into the platform and requesting a session with a tutor, at which point they are connected to a volunteer or paid tutor and placed into a tutoring session. Within these tutoring sessions, students can chat with the tutor via text, or communicate via a virtual white board.

Recently, ASSISTments partnered with UPchieve and, for some students, replaced the ability to request an explanation with the ability to chat with a live tutor. When a live tutor was requested, a tutoring session was opened via UPchieve. This new feature provided the opportunity to compare explanations generated by an LM using both a few-shot learning approach with existing explanations and a summarization approach using the tutoring chat logs for the same mathematics problems.

3 METHODOLOGY

3.1 Original Study

The original study investigates the effectiveness of two methods in prompting GPT-3.5 to generate explanations to middle school math problems by comparing explanations generated to human-written explanations currently used in ASSISTments. GPT-3.5 was either prompted with examples of similar explanations written by teachers and their answers or a live tutoring chat log of a tutor assisting the student in a particular problem. This study used GPT-3.5 as the model of choice as it was the most powerful LLM available when conducting the study.

3.1.1 Data Processing

The style and format of the text generated by LLMs is highly dependent on subtle changes in the prompt used to generate it. There have been many studies of how to properly engineer a prompt for LLMs [17,5,18]. In order to examine the effects of changes to the prompts on the generate explanations in a way that would not bias the results of the analysis, all of the available data for generating explanations was split in half. Half of the data was used for prompt engineering (development set). This data was used iteratively to examine how small variations in the prompt effected the resulting explanations. Once the generated explanations reached a satisfactory level, the most effective prompts were used on the second half of the data (evaluation set). The analysis of the validity and quality of explanations discussed in the results was performed only on this second half of the data, eliminating any bias from the prompt engineering process.

3.1.1.1 Tutoring Chat Logs

During the live tutoring partnership period, there were 244 tutoring sessions across 93 students and 110 problems covering various middle-school mathematics skills. Of these tutoring sessions, 2 were excluded because they contained no interaction between student and tutor (the student never responded to a tutor's opening question) and 2 were excluded because they were longer than GPT-3's 4,000 token limit. The remaining 240 logs were randomly split into development and evaluation sets based on student IDs. While student IDs were used to split the data, we wanted to ensure the datasets would be similar, and verified that both sets contained problems of the same skills in similar proportions. Information on problems' skills was provided by ASSISTments, using the Common Core State Standards for Mathematics [12]. The split we generated was mostly balanced, except for examples where a particular student provided the only example of a particular skill. These instances were placed in the evaluation set. Overall, there were 121 chat logs from 45 students

answering 53 problems in the development set, and 119 chat logs from 48 students answering 61 problems in the evaluation set.

3.1.1.2 Problem-Level Explanations

To prepare ASSISTments data for few-shot learning, only problems and explanations from the EngageNY and Illustrative Mathematics curricula were considered because within ASSISTments, those two curricula are the most popular and have the most explanations. Only non-open response problems and explanations that were fully text-based could be used for few-shot learning because few-shot learning required the problem, answer, and explanation to be in the prompt. Additionally, some ASSISTments problems were excluded because they were follow-up problems to previous problems and did not provide all the context necessary to solve the problem. Of the 40,523 problems and 22,944 explanations available, only 9,200 problems and 11,345 explanations remained after removing problems that could not be used in the few-shot learning prompt. These problems and their explanations were evenly partitioned into a development and evaluation set as well, stratified by problem skills.

In order to compare few-shot learning based explanations to summarization-based explanations, the few-shot learning approach was used only to generate explanations for the problems that were discussed within the tutoring chat logs. Problems with skills different from the problems in the tutoring chat logs were removed, leaving 315 development problems and 599 evaluation problems for the few-shot learning approach.

3.1.2 Experimental Methodology

3.1.2.1 Tutoring Chat Log Summarization

Development data was used to engineer a four-step process for generating explanations from tutoring chat logs. The prompts are shown below, with the GPT-3.5 parameters shown in parentheses as (Frequency Penalty, Temperature, Max Tokens). The text-davinci-003 model was used for all prompts.

- 1) Does the tutor successfully help the student in the following chain of messages? [The tutoring chat log.] (0, 0.7, 128)
- 2) Explain the mathematical concepts the tutor used to help the student, including explanations the tutor gave of these concepts, and ignoring any names. [The tutoring chat log.] (0.25, 0.9, 750)
- 3) Reword the following explanation to not include references to a tutor or student, and to be in the present tense: [The previously generated explanation.] (0.25, 0.9, 750)
- 4) Summarize the following explanations, making sure to include the most generalizable math advice. [The previously generated explanations.] (0.25, 0.9, 500)

Step 1 asked GPT-3.5 to evaluate the initial tutoring chat log to determine if the tutor provided help to the student. This step was initially broken into two steps: one which asked if the tutor provided mathematical help to the student and one to ask if the mathematical help was useful to the student. This two-step evaluation of helpfulness proved to be too restrictive, as during the initial prompt engineering phase, about 60% of the original chat logs were excluded, even though some contained valid mathematics advice. The prompt was then changed to evaluate only if the tutor helped the student. This only filtered out message chains where little information was transferred between tutor and student, resulting in about 20% of chat logs being deemed unhelpful and discarded. Step 2 asked GPT-3.5 to summarize the mathematical help given by the tutor to the student. The outputs generated by GPT-3.5 at this stage contained mathematical content, but were worded as a past-tense summary of the interaction between tutor and student. To refine these summaries, Step 3 was added, which asked GPT-3.5 to reword the output of Step 2 into a present tense explanation by removing references to the interactions

between tutor and student. Finally, some problems had multiple tutoring chat logs discussing them. In order to generate a single explanation per problem, a final step was added to summarize all the previously generated explanations for a problem when more than one generated explanation existed.

3.1.2.2 *Problem-Level Explanation Few-Shot Learning*

Before generating explanations for the 53 problems in the summarization development set, problems that were open response or not text-based had to be removed. After this, 40 problems remained. This was deemed to be an acceptable level of loss, and no steps were taken to try an include problems with images in the few-shot learning approach. For each of the 40 remaining problems a prompt was constructed by randomly sampling problems of the same skill from the development set, and appending the phrase below, replacing the content in brackets with the problem content.

Problem: [The text of the problem.] Answer: [The answer to the problem.]

Explanation: [The explanation for the problem.]

A problem was considered to be of the same skill as another if the grade level and subject were the same. This was decided because if the entire Common Core Skill Code had to be identical, there would not have been enough problems for prompt generation. At the end of the prompt, the phrase above was used for the problem for which an explanation was being generated, but nothing was added for the explanation, allowing for GPT-3.5 to fill in the explanation. Due to the 4,000 token prompt limit, if including all the related problems in the prompt made the prompt over 11,523 characters long, related problems were randomly removed from the prompt until the prompt was less than 11,523 characters long, which was determined, using the development set, to approximate the 4,000 token limit. For these prompts, the Frequency Penalty was 0, the Temperature was 0.73, the Max Tokens was 256, and the code-davinci-003 model was used.

3.1.2.3 *Empirical Analysis of Generated Explanations*

After the summarization and few-shot learning processes were completed for the evaluation data using the processes developed with the development data. The explanations from both processes were manually evaluated by subject-matter experts for both structural and mathematical correctness. Structural correctness required that the explanation generated by GPT-3.5 be in the format of a mathematics explanation. For example, if GPT-3.5 generated the explanation “Go take a walk, then come back and try again.”, that would be structurally incorrect. Mathematical correctness refers to whether or not the explanation given by GPT-3.5 is mathematically correct. For example, if GPT-3.5 generated the explanation “To solve for x in the equation $x + 3 = 5$, subtract 3 from both sides of the equation, which gives you $x = 3$.”, that would be structurally correct because it is in the format of a mathematics explanation, but mathematically incorrect, because $x = 2$, not 3.

After structurally or mathematically incorrect explanations were removed by experts, the remaining explanations were mixed with any existing explanations already in ASSISTments written by teachers for the same problems. The source of the explanations from summarization, few-shot learning, and teachers was blinded, and mathematics teachers were given a picture of each mathematics problem and the text of the explanation and told to rate, on a scale from 1-5, how likely it is that the explanation would help a student. Mathematics teachers have proven in the past to be effective creators of explanations for ASSISTments [17], therefore, they are likely reliable evaluators of the quality of this content. After collecting all of the teachers’ survey results, the correlations between the teachers’ ratings were calculated to examine the inter-rater reliability of the survey results. A Pearson correlation matrix was used to examine the similarity of different

teachers' results. A correlation matrix was used because it allowed for the explanation ratings to be treated as continuous variables. By treating the ratings as continuous, as opposed to a categorical variable with five categories, teachers that were more or less strict with what they considered to be an excellent explanation would still have positive correlations as long as they generally agreed on how good the explanations were relative to other explanations. Additionally, only a small number of teachers were expected to participate in the survey, making the correlation matrix easily interpretable. Additionally, the correlation matrix had the potential to reveal different modes of thought among teachers, i.e., there could be clusters of teachers with similar opinions that differ from other clusters of teachers.

Once the survey results were deemed consistent, a multi-level model [7] was used to predict the rating of each explanation given random effects for the rater and the mathematics problem, and fixed effects for the source of the explanation. Random effects were used to compensate for low sample sizes while still taking into account the differences between raters, problems, and the effects that differences had on the ratings of explanations. The effects for the sources of explanations were used to determine if there were any statistically significant differences between the sources.

3.2 Replication Study

With the release of the more powerful GPT-4 model after the original study was conducted, the measurable increases in performance for GPT-4 in math benchmark tests likely indicates that using GPT-4 will produce better explanations for middle school math problems compared to its predecessor GPT-3.5. This hypothesis was tested by prompting GPT-4 to generate explanations for the same problem set used with GPT-3.5 in the original study, allowing for the comparison of both models in their explanation generation abilities.

3.2.1 Explanation Generation

The dataset of problems that were given to GPT-4 for problem generation included all problems from the summarization and few-shot learning approach used by GPT-3.5. Duplicate problems or problems that included images from the ASSISTments database that would be inaccessible to the text-based model were removed. After this, 33 problems remained to be given to GPT-4 for explanation generation. Explanations were generated using the system-level prompt below, which was found by attempting to generate explanations to the first three mathematics problems in the data set, which all tested different skills. The wording of the prompt was altered until the explanations generated by GPT-4 were of adequate quality for all three types of problems. For this prompt, the Temperature was 0.31 and Maximum Length was 256 tokens.

“The user will provide a middle school math problem that a student is currently struggling on. The student requests for an explanation to how to find an answer to the problem. Provide a step-by-step explanation as a middle school math teacher that is easy enough for a student to understand, and that they will learn from. Problem explanations must be under 170 words and very concise, and easy to follow. Respond in a direct and factual tone in third person. Value efficiency in finding the answer using the least number of steps rather than a single-step mathematical operation. Find a creative solution.”

We gave all problems to GPT-4 in an HTML format to accurately account for math symbols, such as exponents, used in problem bodies.

3.2.2 Evaluation Method

After explanations were generated, they were manually evaluated based on whether the response was structured as an explanation, and also if there were no math errors present. If both conditions were true, the problem was considered an

invalid explanation and not included in the evaluation survey. There were 2 problems out of the 33 that were not added to the survey. The other 31 valid explanations generated by GPT-4 were appended to the survey from the original study that included the valid GPT-3.5 generated explanations from few-shot and tutor chat log summarization. The structure of the original survey was kept the same. Our new survey included the 43 old explanations from the original study, generated in July 2022, and 31 newly generated valid explanations from GPT-4 process generated in June 2023, for a total of 74 problems.

After the GPT-4 problem explanations were generated and an evaluation survey was created, explanations were manually evaluated by three undergraduate students familiar with the content present on the ASSISTments platform with a high level of mathematics understanding. Verbal instructions were given the evaluators to rate problems based on if they were correct, and if they would help students in similar problems in the future. Explanations were evaluated on a scale of 1-5 (1 = Very Bad, 5 = Very Good), and each evaluator rated the problems independent of any influence from the ratings of others.

All ratings were aggregated, and a mixed-effects model was fit with random effects for the problem and rater, and fixed effects for the source of the explanation. This model was identical to the model used in the original study above [11]. The fixed effects for the source of the explanations were used to measure the difference in quality of explanations, as these effects can be interpreted as the average rating of an explanation generated by the corresponding source after factoring out confounding from different raters' strictness and the difficulty of explaining specific problems.

3.2.3 Limitations

The methodology of the replication study allows for an opportunistic comparison between GPT-4 and human written explanations, but the conclusions drawn would not be strong. Explanations created by GPT-4 were not for the same set of problems with explanations written by teachers, potentially introducing bias if the problems with human-written explanations were inherently easier to write explanations for or vice versa. The content utilized was also limited to only problems that generated valid explanations from the summarization of tutoring chat logs, limiting the data available for the study. As such, the methodology of the follow-up study removes constraints on problems that explanations are generated for and ensures each problem has a GPT-4-created and human-written explanation available, decreasing bias and increasing statistical power.

3.3 Follow-Up Study

The analysis plan has been preregistered on OSF prior to data access and can be found at <https://osf.io/x3qrh>.

The follow-up study addresses many of the limitations that were present in the replication study. Mathematics problems from the problem set that GPT-4 generated explanations for were not random, and as such, conclusions cannot be drawn about GPT-4's standalone capabilities of explanation generation. As the end goal of using LLMs is to streamline the process of writing explanations for all types of problems in ASSISTments, the follow-up study evaluates GPT-4's ability to generate explanations for a random set of problems picked from the ASSISTments database.

3.3.1 Explanation Generation

In the follow-up study, the set of problems given to GPT-4 to generate explanations were randomly sampled from the ASSISTments problem database. 100 problems were randomly selected from all text-based problems that contained a teacher-written text-based explanation. Additionally, only the problems that came first in a multi-part problem were sampled from so that each problem required no prior context to solve. Explanations were generated using the following

zero-shot process. Problems were provided to GPT-4 in HTML format to account for special symbols, the temperature was 0.5, and the responses had no maximum token length. First, a system prompt was given to GPT-4:

“You are a middle-school math teacher. A student is completing an online math assignment. Provide the student with a very concise explanation that teaches them, step-by-step, how to solve for the answer to the following problem. The explanation should be easy for a middle-school student to understand and learn from. If there are efficient shortcuts or rules of thumb that can be used to solve the problem, include them in the explanation. Return only the explanation formatted as HTML starting with <p>.”

Then, the following user prompt was given to GPT-4:

```
Problem HTML:
[Problem HTML]
Acceptable Answer(s):
[First Acceptable Answer]
[Second Acceptable Answer]
...
[Last Acceptable Answer]
```

3.3.2 Evaluation

The existing explanations in the ASSISTments database were combined with the 100 explanations created by GPT-4 for the same problems, resulting in a total of 200 explanations. We combined both sets of 100 explanations into a spreadsheet-based survey that included a column for the problem, a column for the explanation, and a column for raters' ratings. Explanations were evaluated using the same 1-5 scale in the replication study, and raters were given the same verbal instructions for how to rate explanation quality. The source of the explanation (GPT-4 or human) was blinded and the order of the explanations was randomized for each rater to reduce ordering effects. The 200 explanations were each evaluated by ten raters with the same qualifications as raters from the replication study.

To compare the quality of GPT-4 generated explanations to human-written explanations, the same model from the replication study was used, except random effects for the interactions between raters and problems were also included to help remove any additional confounding from the interactions. The fixed effects of GPT-4 generated explanations and human-written explanations were again measures of the average quality of each sources explanations.

4 RESULTS

4.1 Original Study

4.1.1 Summarization

Performing the summarization process on the evaluation data resulted in 26 acceptable explanations. Initially, there were 119 chat logs across 61 problems. Step 1 of the summarization process removed 14 tutoring logs, resulting in 105 explanations across 57 problems. The complete process generated 57 explanations. Expert review of the final explanations found 14 invalid explanations due to bad structure and 17 due to incorrect mathematics, which is only about a 43% success rate. The 26 valid summarization based explanations were included in the explanation quality survey.

4.1.2 Few-Shot Learning

Performing the few-shot learning process on the evaluation data resulted in only 6 acceptable explanations. 28 of the initial 61 evaluation problems were removed because they were not solely text-based. Of the 33 remaining problems, 1 was invalid due to bad structure, and 26 due to incorrect mathematics, which is only about a 10% generation success rate. The 6 valid few-shot learning based explanations were included in the explanation quality survey.

4.1.3 Empirical Analysis of Generated Explanations

After both procedures for generating explanations using GPT-3.5 were complete, and the structurally or mathematically incorrect explanations were removed, any explanations for 61 problems in the evaluation set that were already written by teachers for ASSISTments were combined with the remaining GPT-3.5 generated explanations. In total, 26 summarization, 6 few-shot learning, and 10 ASSISTments explanations were included for a total of 42 survey questions. Five current or

Correlations Between Teachers' Ratings of Explanation Quality

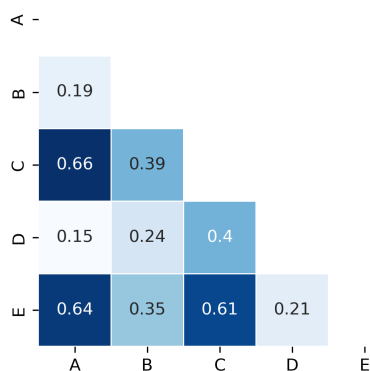


Figure 2: The correlation between all teachers' ratings of explanation quality, determined using the survey results.

former middle-school or high-school mathematics teachers completed the survey. The correlation between all the teacher's ratings is shown in Figure 2, where each teacher is anonymized as a letter of the alphabet, and the value in the cell shows the correlation between the row and column teachers' ratings. Although some teachers had a low correlation between their ratings, no teachers had a negative correlation between their ratings. Teachers A, C, and E have the highest correlation with each other, while Teachers B and D were less correlated with other teachers. Although some teachers were more or less strict than others, which lowered their correlations, in general, teachers agreed on what makes an explanation good or bad.

Once the inter-rater reliability was deemed sufficient, a multi-level model [15] was fit with random effects for the rater and the mathematics problem, and fixed effects for the source of the explanation. Two different models were fit, one that only included teachers' ratings of valid explanations, and one that included all the generated explanations, with a rating of 1 for explanations that were invalid. The effects and 95% confidence intervals of the different sources of explanations are shown in Figure 3. ASSISTments explanations are rated the highest, with an average rating of about 4.2. It is unsurprising that the explanations written by teachers were the most highly rated. Summarization based explanations were statistically significantly worse than ASSISTments explanations, with an average rating of about 2.6 for the valid explanations and 1.7

for all explanations. Qualitatively, teachers reported that the summarization-based explanations used terms that the students did not necessarily know, and tended to give advice that was too general. Few-shot learning based explanations received an average rating of about 3.6 for valid explanations, which was not statistically significantly worse than ASSISTments explanations, but only 6 of the few-shot learning based explanations were valid. If the invalid explanations are included in the analysis, then few-shot learning based explanations received an average rating of about 1.6, which is statistically significantly worse than ASSISTments explanations.

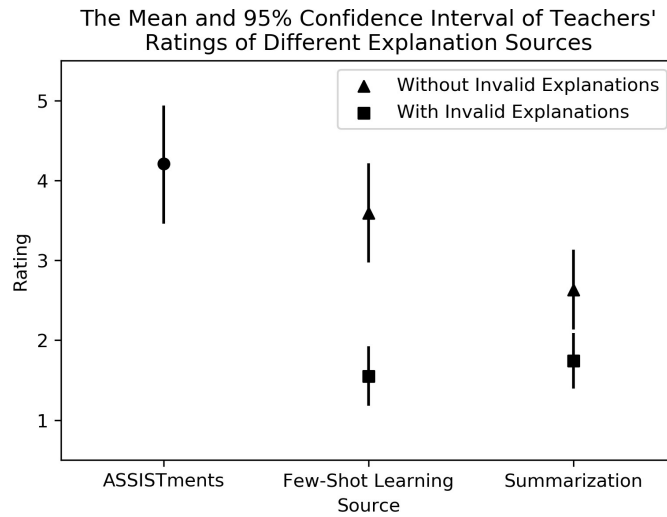


Figure 3: The mean and 95% confidence interval of teachers' ratings of explanation quality by source, determined using survey results. Invalid explanations, when included in the model, are assumed to have a rating of 1 for quality.

4.2 Replication Study

Out of the 33 problems given to GPT-4 in the replication study, 2 were manually evaluated to have math errors, equal to an approximately 94% success rate. Comparatively, the original study that used GPT-3.5 to generate explanations had a success rate of 43% for the summarization approach and a 10% generation success rate for the few-shot learning approach, which shows a dramatic increase in the ability for GPT to accurately generate explanations. Figure 4 shows the graph of the mean ratings and 95% confidence interval for explanations created by humans, both methods from GPT-3.5, and GPT-4. Compared to the previous study with GPT-3.5, the average ratings for the 3 categories that already existed were almost identical, which gives us confidence that this study is a valid replication and results are comparable.

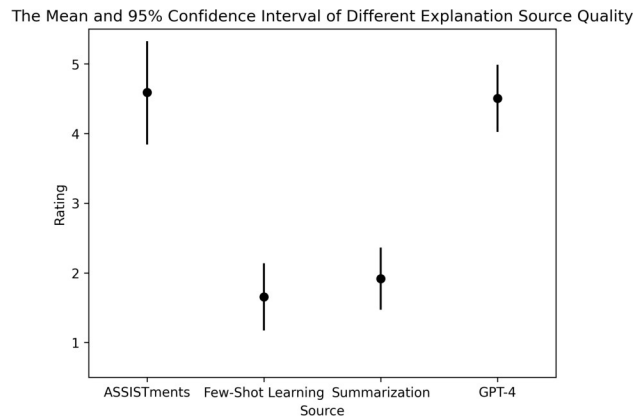


Figure 4: The mean and 95% confidence interval of explanation ratings for 4 sources of explanation generation.

The GPT-4 ratings scored much higher than two methods of GPT-3.5 generation. Ratings were also approximately equal to existing ASSISTments ratings; however, such conclusion cannot be drawn given the possibility for bias in the problems chosen from the ASSISTments database, and the small sample size of 10 problems in the evaluation. The follow-up study has a larger sample size and problems are randomly sampled.

4.3 Follow-Up Study

The 100 problems sampled and given to GPT-4 to generate explanations for were manually evaluated for correct explanation structure and correctness. Of the 100, 4 of them contained errors in the wording or utilized a problem approach different to the one specified in the problem, which is a 96% success rate. While there are still instances of explanations that were classified as invalid, all 4 errors still led to the correct answer and the error rate is much smaller than the GPT-3.5 explanations, so we are confident the explanations could be implemented into the ASSISTments system without harming student learning. The distribution of ratings is shown in Figure 5. The GPT-4 explanations were rated higher than the human created explanations, with an average rating of 4.3. Comparatively, human-created explanations were rated with an average of 3.7.

The lack of overlap in the 95% confidence intervals indicates that it is highly likely the GPT-4 explanations were preferred with higher ratings. Such results are quite surprising, as we assumed the human-created explanations that are currently used in ASSISTments would be the most preferred. It is important to note that ratings signify the perception of an explanation's quality and whether the evaluator preferred it, not the inherent quality of the explanation itself. After the survey was completed and the purpose of the survey was revealed, evaluators noted that they preferred explanations generated by GPT-4 because it took on a clearer step-by-step approach to solve the problem, and the explanations also explained the steps more with relevant concepts compared to the ASSISTments explanations.

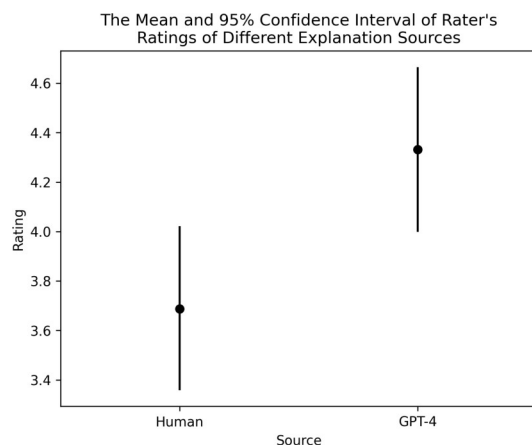


Figure 5: Distribution graph comparing ratings for GPT-4 and human created explanations.

5 CONCLUSION

This work concludes that GPT-4 explanations were preferred over both GPT-3.5 and teachers. As the original study determined that GPT-3.5 simply does not have the mathematics training necessary to generate high-quality explanations, the replication and follow-up study leads to the conclusion that the GPT-4 model is far more successful at interpreting

mathematics and generating high-quality explanations. The implications of such findings open the door for further research into GPT-4 as a viable alternative to human-written explanations in tutoring systems.

While the original study makes it apparent that GPT-3's explanations are worse than teacher-written explanations, it is limited to just middle school mathematics problems. A difficult part of generating explanations for simple mathematics problems is that often GPT-3.5 writes explanations with the assumption that fundamental mathematics concepts are already known. Based on the success that other studies have had using GPT-3.5 to interpret college level mathematics [6], it may be, for example, easier for an LM to understand integrals than scientific notation because there is far more language used in the descriptions and use cases of integrals than there is in the description of scientific notation, which is just a simple mathematics operation.

Additionally, there is no closed-form solution for prompt engineering. To avoid bias, a development set was used to create prompts via trial and error, but there is no guarantee that this work constructed the best prompts for generating explanations from tutoring chat-logs or from similar problems and their answers and explanations. Even with a four-stage process for summarizing tutoring chat logs, a better, more concise prompt might be achievable when approaching the generation process differently. More work to explore and refine the prompts used to generate explanations could be done to better understand how to get the best content from an LM.

Even if one assumes that there exists a prompt that would generate effective explanations of mathematics problems, the entire process is still limited to purely text-based content. Many mathematics problems use diagrams or equations to represent information. Without the ability to interpret this information, the capacity to use an LM to create explanations will be limited to a small subset of mathematics curricula. In the short term, efforts should be made to algorithmically generate text alternatives to mathematics diagrams and equations. Then, these text alternatives could be substituted for the diagrams and equations they represent. In the long term, a large LM capable of interpreting mathematics, or even logic, could have a tremendous impact on the quality of the content generated from the model.

There are many confounding variables present in the experiment that could have affected outcomes in the replication and follow-up study, where explanations generated were rated by humans. One such variable is length of explanation. When evaluators were asked, many said that they also considered the length of the problem as a sign of whether it was high-quality: the longer the problem, the more detailed the steps and the higher chance it would have to help a student learn. If an explanation was shorter, some evaluators also marked the explanation as better because the explanation given was more concise, brief, and easy to understand. Evaluators also mentioned following the evaluation period that not all explanations were rigorously checked for potential math errors. This introduces a potential bias that favors GPT-4 because the explanations generated from the model are often convincing enough in argument and structure that math errors are overlooked. While experienced raters preferred GPT-4 generated explanations, there is no guarantee that the explanations cause more learning. As a next step, we will conduct a randomized controlled experiment where GPT-4 explanations are integrated into ASSISTments. Students will be randomly assigned to receive GPT-4-created or human-written support in order to compare GPT-4 explanations' true effect on student learning.

ACKNOWLEDGMENTS

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, & 1903304), IES (e.g., R305N210049, R305D210031, R305A1-70137, R305A170243, R305A180401, & R305A1-20125), EIR (U411B190024 & S411B210024), NHI (R44GM146483), and Schmidt Futures. None of the opinions expressed here are that of the funders.

REFERENCES

- [1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- [3] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24 (2014), 470–497.
- [4] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.
- [5] Ming-Chi Liu and Yueh-Min Huang. 2017. The use of data science for education: The case of social-emotional learning. *Smart Learning Environments* 4, 1 (2017), 1–13.
- [6] OpenAI. 2023. GPT-4 Technical Report. (2023). arXiv:2303.08774 [cs.CL]
- [7] OpenAI. 2023. Model Index for Researchers. <https://platform.openai.com/docs/model-index-for-researchers>. Accessed June 18, 2023.
- [8] OpenAI. 2023. OpenAI Playground. <https://platform.openai.com/playground?mode=chat>. Accessed June 18, 2023.
- [9] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).
- [10] Dawn Peterson. 2022. Upchieve: Free On-Demand Tutoring for Title I High Schools. <https://new.assistments.org/blog-posts/upchieve-free-on-demand-tutoring-for-title-1-high-schools>. (2022). Accessed June 18, 2023
- [11] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [12] Akkus, M.: The common core state standards for mathematics. *International Journal of Research in Education and Science* 2(1), 49–54 (2016)
- [13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901 (2020)
- [14] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J. Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E. Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022), <https://arxiv.org/abs/2204.02311>
- [15] Gelman, A., Hill, J.: *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press (2006)
- [16] Patikorn, T., Heffernan, N.T.: Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In: *Proceedings of the Seventh ACM Conference on Learning@ Scale*. pp. 115–124 (2020)
- [17] Reynolds, L., McDonnell, K.: Prompt programming for large language models: Beyond the few-shot paradigm. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. pp. 1–7 (2021)
- [18] Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022)