Assessing the Quality of Large Language Models in Generating Mathematics Explanations

Allison Wang
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
awang9@wpi.edu

Ethan Prihar
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
ebprihar@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

The development and measurable improvements in performance of large language models on natural language tasks [12] opens up 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 [7]. In a prior 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 this replication study, we were interested in whether the measurable improvements in GPT-4 performance [6] 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, single-blinded 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. The implications of our significant results at Learning @ Scale are that digital platforms can start A/B testing the effects of GPT-4 generated explanations on student learning, implementing explanations at scale, and also prompt programming to test different education theories, e.g., social emotional learning factors [5].

CCS CONCEPTS

• **Applied computing** → **Education**; **Distance learning**; **Computer-assisted instruction**.

KEYWORDS

Large Language Models, Online Tutoring, A/B Testing

ACM Reference Format:
Allison Wang, Ethan Prihar, and Neil Heffernan. 2023. Assessing the Quality of Large Language Models in Generating Mathematics Explanations. In *Proceedings of the Fourth A/B Testing and Platform-Enabled Learning Research Workshop at Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3573051.3593376

L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3593376

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 a previous study [11], we compared explanations generated by GPT-3.5 with explanations written by math teachers for middle school math problems. Results from the analysis found that GPT-3.5 performed statistically worse than human-written explanations. 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]. In this work 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) Replication Study: How effective are GPT-4 explanations compared to GPT-3.5 explanations generated in prior work?
(2) 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.

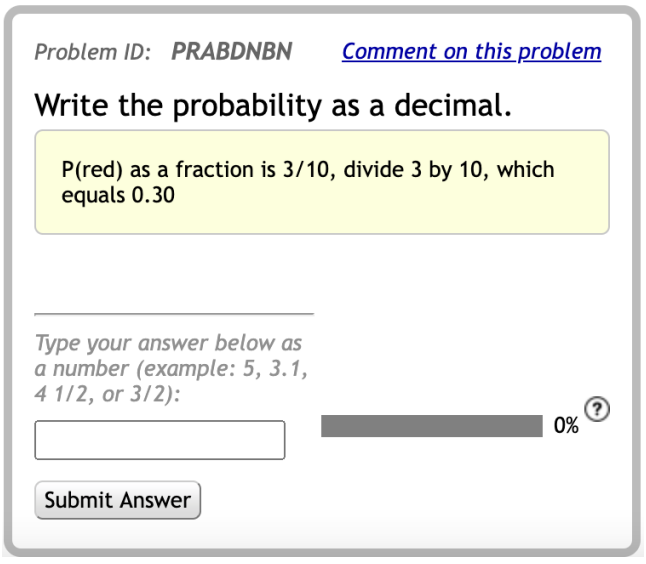

Figure 1: Example of a problem explanation in ASSISTments from the student perspective.

2.3 Prior Study

In a previous study [11] the text-davinci-003 model (OpenAI's GPT-3.5 [7]) was used in two different approaches to generate explanations for mathematics problems. In the few-shot learning approach, GPT-3.5 was trained on examples of math problems, their answer, and an existing explanation. The model then generated an explanation for a new problem with the answer provided. The summarization approach utilized the tutoring dialogue from chats between a student and tutor on the UPchieve online tutoring platform, which was integrated into ASSISTments [10]. Tutoring chat logs were given to GPT-3.5, and the model was asked through specific prompts to condense the advice from the tutor about solving the problem into a short and concise paragraph.

Results from the analysis concluded that both methods utilized to generate explanations from GPT-3.5 performed statistically worse than human-written explanations already in ASSISTments [11]. As such, the study that integrates GPT-4 explanations alongside GPT-3.5 and human-written explanations is a replication study that aims to answer whether there is measurable improvement between GPT-3.5 and GPT-4.

3 METHODOLOGY

3.1 Replication Study

3.1.1 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.

3.1.2 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 any math errors

present. If any of the two 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 prior work [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.

3.1.3 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.

3.2 Follow-Up Study

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

3.2.1 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]

3.2.2 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 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 2 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.

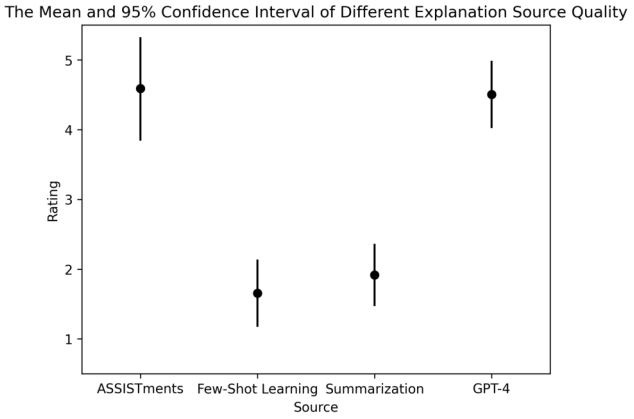

Figure 2: The mean and 95% confidence interval of explanation ratings for 4 sources of explanation generation.

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.

4.2 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 3. 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.

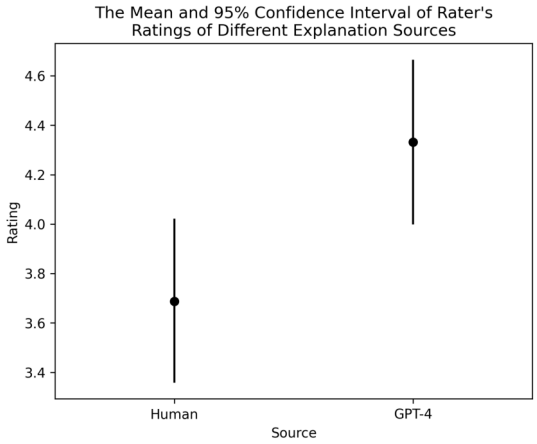

Figure 3: Distribution graph comparing ratings for GPT-4 and human created explanations.

5 CONCLUSION

This work concludes that GPT-4 explanations were preferred over both GPT-3.5 and humans. The implications of such findings opens the door for further research into GPT-4 as a viable alternative to human-written explanations in tutoring systems.

There are many confounding variables present in the experiment that could have affected outcomes. 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 1 High Schools. https://new.assistments.org/blog-posts/upchieve-free-on-demand-tutoring-for-title-1-high-schools. (2022). Accessed June 18, 2023.

[11] Ethan Prihar, Morgan Lee, Mia Hopman, Adam T Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil Heffernan. 2023. Comparing different approaches to generating mathematics explanations using large language models. *Journal of artificial intelligence in education* (2023).

[12] 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).