Comparing Different Approaches to Generating Mathematics Explanations Using Large Language Models *

Ethan Prihar[1 [0000−0002−5216−9815]], Morgan Lee[1 [0000−0002−9839−9608]], Mia Hopman[1 [0000−0001−8267−5455]], Adam Tauman Kalai[2 [0000−0002−4559−8574]], Sofia Vempala[1 [0000−0002−5855−4503]], Allison Wang[1 [0000−0002−4646−7621]], Gabriel Wickline[1], Aly Murray[3], and Neil Heffernan[1 [0000−0002−3280−288X]]

[1] Worcester Polytechnic Institute, Worcester, MA 01609, USA
{ebprihar,mplee,mahopman,svempala,awang9,gwickline,nth}@wpi.edu
[2] Microsoft Research, 1 Memorial Dr, Cambridge, MA 02142, USA
adam.kalai@microsoft.com
[3] UPchieve, Inc.
aly.murray@upchieve.org

Abstract. Large language models have recently been able to perform well in a wide variety of circumstances. In this work, we explore the possibility of large language models, specifically GPT-3, to write explanations for middle-school mathematics problems, with the goal of eventually using this process to rapidly generate explanations for the mathematics problems of new curricula as they emerge, shortening the time to integrate new curricula into online learning platforms. To generate explanations, two approaches were taken. The first approach attempted to summarize the salient advice in tutoring chat logs between students and live tutors. The second approach attempted to generate explanations using few-shot learning from explanations written by teachers for similar mathematics problems. After explanations were generated, a survey was used to compare their quality to that of explanations written by teachers. We test our methodology using the GPT-3 language model. Ultimately, the synthetic explanations were unable to outperform teacher written explanations. In the future more powerful large language models may be employed, and GPT-3 may still be effective as a tool to augment teachers' process for writing explanations, rather than as a tool to replace them. The explanations, survey results, analysis code, and a dataset of tutoring chat logs are all available at https://osf.io/wh5n9/.

Keywords: Large Language Models · GPT-3 · Online Learning · Tutoring

* We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NHI (R44GM146483), and Schmidt Futures. None of the opinions expressed here are that of the funders.

1 Introduction

Online learning platforms offer students tutoring in a variety of forms, such as one-on-one messaging with real human tutors [1] or providing expert-written messages for each question that students are required to answer [5]. These methods, while effective, can be costly and time consuming to scale. However, recent advances in Language Models (LMs) may provide an opportunity to offset the cost of providing effective tutoring to students.

In this work, we explore the effectiveness of using LMs to create explanations of mathematics problems for students within the ASSISTments online learning platform [5]. Recent transformer-based LMs have exhibited breakthrough performance on a number of domains [2,3]. In this work, we perform experiments using one of the most powerful currently available LMs, GPT-3 [2], accessed through OpenAI's API.

Two different approaches to generate this content were explored. The first approach used few-shot learning [2] 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 LM-based approach, as well as an evaluation of their applicability in a real-world setting.

2 Background

2.1 Language models

LMs are a type of deep learning model trained to generate human-like text. They are trained on a massive dataset of millions of web pages, books, and other written documents, and are capable of generating text that is often indistinguishable from human-written text [2,3]. In this work, we focus on GPT-3 since it is a powerful LM that is publicly accessible through a paid API. When using GPT-3, one can specify parameters for the text generation such as Frequency Penalty, which penalizes GPT-3 for repeating phrases in its response, Temperature, which increases the frequency of picking a less-that-most-likely word to include in the response, and Max Tokens, which specifies the maximum length of the response [2].

2.2 Data Sources

The data used to generate explanations using few-shot learning came from the ASSISTments online learning platform [5]. Within ASSISTments, middle-school mathematics students complete mathematics problem sets assigned to them by

their teacher. If students are struggling with their assignment, ASSISTments will provide them with an explanation upon their request. When a student requests an explanation, a message that explains how to solve the mathematics problem they are currently struggling with and the solution to the problem is provided to them.

The data used to generate explanations from summaries of tutoring chat logs comes from UPchieve, a provider of online tutoring. UPchieve[4] offers live online tutoring with volunteers through an interface that facilitates sharing of text and images. In ASSISTments, students had the ability to request a chat with a live tutor. When a live tutor was requested, a tutoring session was opened via UPchieve.

3 Data Processing

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.

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 and 2 were excluded because they were longer than GPT-3's 4,000 token limit. Of the 40,523 problems available, only 914 problems remained after removing the problems that were not of the same skills as the problems in the tutoring sessions and the problems that could not be used in the few-shot learning prompt due to the presence of images within the problems, answers, or explanations. Both the chat logs and the problems and their explanations were evenly partitioned into a development and evaluation set stratified by subject matter.

4 Methodology

4.1 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 parameters shown in parentheses as (Frequency Penalty, Temperature, Max Tokens). The text-davinci-003 model was used for all prompts.

[4] https://upchieve.org/

1. Does the 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)

4.2 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. 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, until there were no more same-skill problems or the max token length was reached. 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.

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

4.3 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 validity. Structural validity required that the explanation be in the format of an explanation. Mathematical validity required that the explanation be mathematically correct.

The valid generated explanations and teacher-written explanations for the same problems were compiled into a survey. The source of the explanations was blinded, and mathematics teachers were given a picture of each mathematics problem and the text of the explanation and told to rate the explanations on a scale from 1-5. A multi-level model [4] 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. The effects for the sources of explanations were used to determine if there were any statistically significant differences between the sources.

5 Results

Performing the summarization process on the evaluation data resulted in 57 explanations. Expert review found 14 structurally and 17 mathematically invalid explanations. For the 61 problems available in the summarization evaluation set, only 33 explanations could be generated using the few-shot learning approach due to the presence of images in the problems. Expert review found 1 structurally and 26 mathematically invalid 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 former middle-school or high-school mathematics teachers completed the survey. Once survey results were collected, 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 1. ASSISTments explanations are rated the highest, with an average rating of about 4.2. 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. Few-shot learning based explanations received an average rating of about 1.6 when invalid explanations were included in the model, which is statistically significantly worse than ASSISTments explanations.

6 Conclusion

Overall it seems that GPT-3 based explanations do not compare in quality to those created by teachers. Fundamentally, GPT-3 was trained to understand language, but not mathematics, and while the structure of what GPT-3 generated made proper use of the English language, it often generated incorrect mathematical content, or simply failed to generate content in the proper format. Summarizing tutors' advice to students created explanations that were significantly worse than teacher-written explanations, and while the valid explanations generated through few-shot learning were not significantly worse than teacher-written explanations, only 10% of the generated explanations were mathematically valid. Ultimately, GPT-3 does not seem to have the grasp of mathematics necessary to generate high-quality explanations.

References

1. Upchieve's mission, `https://upchieve.org/mission`

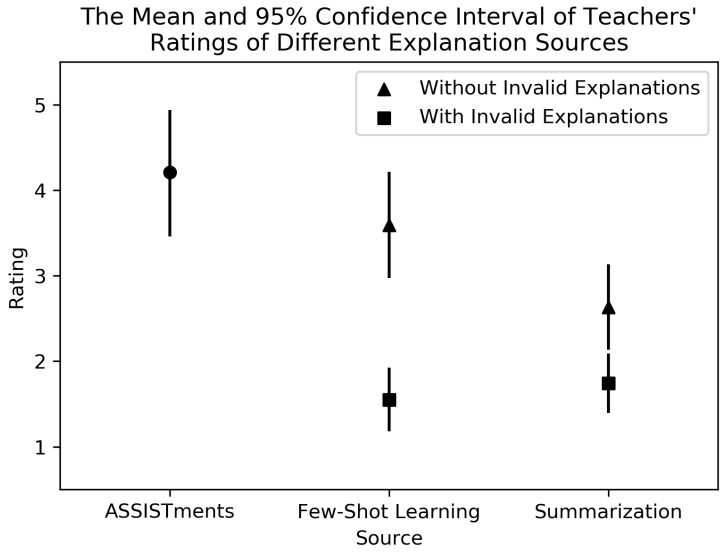

Fig. 1. The mean and 95% confidence interval of teachers' ratings of explanation quality by source, determined using the survey results. Invalid explanations, when included in the model, are assumed to have the lowest rating for quality.

2. 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)
3. 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`
4. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press (2006)
5. Heffernan, N.T., Heffernan, C.L.: 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**(4), 470–497 (2014)