Comparing Different Approaches to Generating Mathematics Explanations Using Large Language Models

BLINDED FOR REVIEW

No Institute Given

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 prompts, explanations, survey results, analysis code, and a dataset of tutoring chat logs are all available at BLINDED FOR REVIEW.

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

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 [8]. 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 [8]. Recent transformer-based LMs have exhibited breakthrough performance on a number of domains [3, 4]. In this work, we perform experiments

using one of the most powerful currently available LMs, GPT-3 [3], accessed through OpenAI's API.

Two different approaches to generate this content were explored. The first approach used few-shot learning [3] 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.

To summarize, in this work we evaluate the following research questions:

1. What is the most effective way to use an LM to create explanations from existing mathematics problems and their answers and explanations?
2. What is the most effective way to use an LM to create explanations from chat logs between students and tutors?
3. How effective are these methods compared to expert-written explanations?

While the explanations we generated using GPT-3 were not rated as highly as those generated by humans, our approach and evaluation methodology is general and could easily be applied to future LMs.

2 Background

2.1 Language models

Recent neural 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 [3, 4]. 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 [3].

Prior work has demonstrated the ability of LMs to summarize a wide variety of information. Most relevant is work specifically on summarizing chat logs between students [15]. In that work, chat logs between students as they discussed strategies during a collaborative game were summarized by an LM for teachers to review. The LM was able to summarize the strategies and opinions of students, as well as students' affect. LMs have also been able to summarize teachers responses in classroom simulations using few-shot learning [11]. Given that LMs

have been successful in a variety of fields and specifically with summarizing dialogue between students and teacher responses, it is likely that LMs would be able to summarize chat logs between a tutor and a student.

LMs have also been used to solve mathematics problems at a middle-school [20] and college [9, 6] level. Beyond solving problems, LMs have been used to generate explanations for problems, though most commonly computer science problems and code explanations [18, 12]. Due to LMs' wide applicability in these domains, it seems likely that LMs would be capable of writing explanations for existing mathematics problems when provided with the text of the problem, the correct answer, and examples of explanations written for similar mathematics problems.

2.2 ASSISTments

The data used to generate explanations using few-shot learning came from the ASSISTments online learning platform [8]. 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. Explanations in ASSISTments have been crowdsourced from graduate students, teachers, and researchers [13, 16], and therefore take on a wide variety of structures and formats.

2.3 Live Online Tutoring

The data used to generate explanations from summaries of tutoring chat logs comes from a provider of online tutoring in partnership with ASSISTments. Organizations like Yup[1], Tutor.com[2], and UPchieve[3] 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 payed 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 a live tutoring provider 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 the live tutoring provider. 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.

[1] https://yup.com/
[2] https://www.tutor.com/
[3] https://upchieve.org/

3 Data Processing

The style and format of the text generated by LMs 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 LMs [17, 5, 19]. 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 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 [2]. 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.

3.2 Problem-Level Explanations

To prepare ASSISTments data for few-shot learning, only problems and explanations from the Engage New York[4] and Illustrative Mathematics[5] 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

[4] http://www.nysed.gov/curriculum-instruction/engageny
[5] https://illustrativemathematics.org/

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.

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.

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)

Step 1 asked GPT-3 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 to summarize the mathematical help given by the tutor to

the student. The outputs generated by GPT-3 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 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.

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. 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 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.

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 correctness. Structural correctness required that the explanation generated by GPT-3 be in the format of a mathematics explanation. For example, if GPT-3 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 is mathematically correct. For example, if GPT-3 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[10], 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 [13], 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 [14] 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 those 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.

5 Results

5.1 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.

5.2 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.

5.3 Empirical Analysis of Generated Explanations

After both procedures for generating explanations using GPT-3 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 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 former middle-school or high-school mathematics teachers completed the survey. The correlation between all the teacher's ratings is shown in Figure 1, 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 [7] 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 2. 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

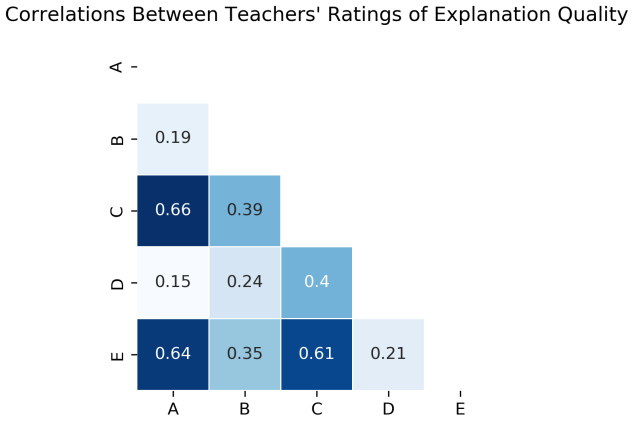

Fig. 1. The correlations between all teachers ratings of explanation quality, determined using the survey results.

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.

6 Limitations and Future Work

While this 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 writes explanations with the assumption that fundamental mathematics concepts are already known. Based on the success that other studies have had using GPT-3 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

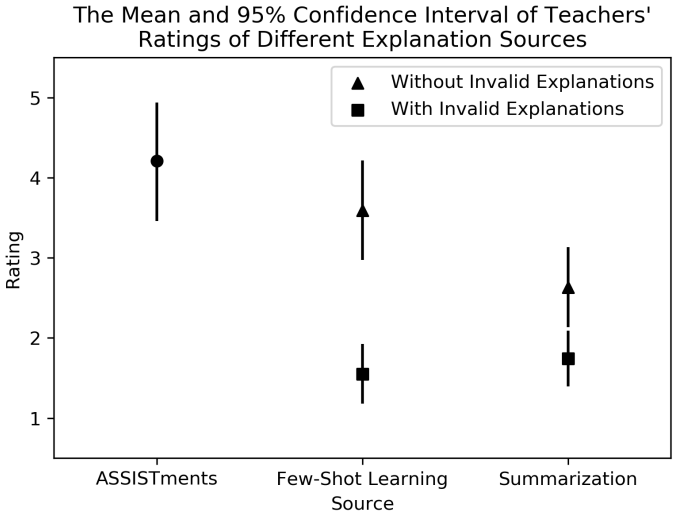

Fig. 2. The mean and 95% confidence interval of teachers' ratings of explanation quality by source, determined using the survey results. Invalid explanations, when included in the model, are assumed to have the lowest rating for quality.

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.

7 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. When summarizing tutors' advice to students, four different prompts had to be used

before the content began to resemble an explanation suitable for integration into an online learning platform, and even after removing invalid explanations, the explanations that were both structurally and mathematically valid were statistically significantly worse in quality than teacher-written explanations. The valid explanations generated through few-shot learning were not statistically significantly worse than teacher-written explanations, but only 10% of the generated explanations were valid. Almost all the other explanations were mathematically invalid.

Ultimately, the latest version of GPT-3 does not seem to have the grasp of mathematics necessary to generate high-quality explanations, but it has other strengths that should be taken advantage of. There are likely much more effective use cases where GPT-3 can improve online learning. Interpreting student affect or identifying the sentiment, emotional, or motivational content in tutoring chat logs all seem like tasks that GPT-3 is more applicable to than explanation generation, and all of these tasks could be used to study and improve students' experiences within online learning platforms. Additionally, Larger language models which perform better on tasks, including mathematical tasks, are being released with increasing frequency. For instance, the PaLM LM, with 540B parameters, is reported to outperform GPT-3 with 175B parameters on a number of tasks [4]. While we do not have access to this LM, our methodology can be applied to future more powerful LMs.

References

1. Upchieve's mission, https://upchieve.org/mission
2. Akkus, M.: The common core state standards for mathematics. International Journal of Research in Education and Science **2**(1), 49–54 (2016)
3. 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)
4. 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
5. Denny, P., Kumar, V., Giacaman, N.: Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. arXiv preprint arXiv:2210.15157 (2022)
6. Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, K., Chen, L., Tran, S., Cheng, N., et al.: A neural network solves, explains, and generates uni-

versity math problems by program synthesis and few-shot learning at human level. Proceedings of the National Academy of Sciences **119**(32), e2123433119 (2022)
7. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press (2006)
8. 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)
9. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., Misra, V.: Solving quantitative reasoning problems with language models (2022). https://doi.org/10.48550/ARXIV.2206.14858, https://arxiv.org/abs/2206.14858
10. Likert, R.: A technique for the measurement of attitudes. Archives of psychology (1932)
11. Littenberg-Tobias, J., Marvez, G., Hillaire, G., Reich, J.: Comparing few-shot learning with gpt-3 to traditional machine learning approaches for classifying teacher simulation responses. In: International Conference on Artificial Intelligence in Education. pp. 471–474. Springer (2022)
12. MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., Huang, Z.: Generating diverse code explanations using the gpt-3 large language model. In: Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2. pp. 37–39 (2022)
13. 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)
14. Pearson, K.: Vii. note on regression and inheritance in the case of two parents. proceedings of the royal society of London **58**(347-352), 240–242 (1895)
15. Phillips, T., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Mott, B., Lester, J.C.: Exploring the use of gpt-3 as a tool for evaluating text-based collaborative discourse. Companion Proceedings of the 12th International Conference on Learning Analytics Knowledge (LAK22) p. 54 (2022)
16. Prihar, E., Syed, M., Ostrow, K., Shaw, S., Sales, A., Heffernan, N.: Exploring common trends in online educational experiments. In: Proceedings of the 15th International Conference on Educational Data Mining. p. 27 (2022)
17. Reynolds, L., McDonell, 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. Sarsa, S., Denny, P., Hellas, A., Leinonen, J.: Automatic generation of programming exercises and code explanations using large language models. In: Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1. pp. 27–43 (2022)
19. 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)
20. Zong, M., Krishnamachari, B.: Solving math word problems concerning systems of equations with gpt-3. In: Proceedings of the Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence (2022)