Augmenting LLM Generated Feedback with Data Mining

Eamon Worden(✉) and Neil Heffernan

Worcester Polytechnic Institute, Worcester, MA 01609, USA
{elworden,nth}@wpi.edu

Abstract. Automated feedback systems are important in mathematics education for providing timely and scalable support to students. While pretrained Large Language Models (LLMs) such as GPT have shown promise in generating feedback, fine-tuning LLMs to improve their performance is costly and resource intensive. In this work, we explore cost-efficient alternatives, focusing on data mining to enhance LLMs for feedback generation for open-ended math answers. We evaluate the effectiveness and practicality of data mining for few-shot prompting in generating both descriptive feedback and numerical scores for middle school math open responses. Our results show that data mining significantly improves the result. Further, we explore why we believe the feedback is better by delving into teacher written reasonings for why they liked or disliked certain feedbacks.

Keywords: Automated feedback · Large Language Models · Prompt Engineering · Mathematics Education · Open-ended Problems

1 Introduction

Automated feedback systems are crucial in online learning, particularly for mathematics education, where open-ended problems pose challenges due to their reliance on subjective reasoning [3]. Effective feedback enhances learning, but poorly designed feedback can cause confusion [2,5]. Recent advancements in LLMs offer promising solutions for improving automated feedback, with techniques like retrieval-augmented generation (RAG) and few-shot prompting providing cost-efficient alternatives to fine-tuning [1,6]. This study explores how data mining can be used to enhance feedback for middle school math responses, comparing few-shot and zero-shot prompting and evaluating the effectiveness of each prompting technique. The main aims of this study are:

1. How can data mining be utilized to generate effective automated feedback for open-ended math problems?
2. How do few-shot and zero-shot, compare in aligning LLM feedback with teacher-authored feedback?
3. How effective and practical are these prompting methods from the perspective of teachers?

A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2590, pp. 340–345, 2025.
https://doi.org/10.1007/978-3-031-99261-2_34

2 Dataset

For our study, we selected open-response problems Illustrative Mathematics Curriculum[1], a widely used middle school math curriculum in the United States, as they were implemented in ASSISTments. We focused exclusively on middle school problems (grades 6–8). We included only problems that had received at least 100 unique student responses and had been both scored and given written feedback by a teacher. A total of 67 problems met these criteria, each containing between 100 and 600 student responses. From this set, we randomly selected 50 problems and 100 student responses per problem, resulting in a dataset of 5,000 responses. Each selected problem was either a stand-alone open-response question or a follow-up prompt requiring students to explain their reasoning based on a prior answer. For our prompt, we split our dataset into an 80-20 retrieval-evaluation split.

3 Methodology

Our study three prompting methods for generating automated feedback for student responses to open-ended math problems using the state-of-the-art GPT-4o model. Specifically, we utilize an intelligent few-shot prompt to enhance feedback generated by the GPT-4o model and then evaluate it against traditional zero-shot and random few-shot prompting. This section details our proposed intelligent few-shot-based approach, along with the baseline prompting methods used for comparison.

Zero-Shot Prompting. is the simplest approach, where a LLM is asked to generate feedback without any examples or context. The model relies entirely on its pre-trained knowledge and the provided student response to generate feedback. While this method requires no additional setup, we consider this as a baseline for comparison with the proposed method. In our zero-shot prompt we assigned GPT-4o a persona of a math teacher, and asked it to provide feedback and a score to a students answer.

Few-Shot Prompting. Few-shot prompting improves upon zero-shot prompting by providing the model with a small set of example responses and their corresponding teacher-written feedback. In the random few-shot approach, we randomly select three example responses and feedback from our retrieval set to include in the prompt. While this method improves feedback quality compared to zero-shot prompting, the random selection of examples may lead to suboptimal performance due to mismatches between the provided examples and the student response being evaluated. For our random few-shot prompt we utilize both a persona as well as three shots in our prompt.

Intelligent Few-Shot Prompting. To further enhance the quality of the generated feedback, we utilize an intelligent few-shot prompting method that dynamically

[1] https://illustrativemathematics.org/.

selects the three most similar student answers from our retrieval set for each student response, along with the associated feedback and scores. This resembles RAG, which selects text from a corpus when prompting, however differs in that we select items from our retrieval set rather than a textbook. We used Meta's Faiss library [4], an efficient tool for similarity search for relevant responses and feedback examples. For each math problem in the dataset, we first vectorized each student's responses from the training set using the state-of-the-art SFR-embedding-2 model [7], which generated semantically rich embeddings for student responses. Once the training set responses were embedded, we calculated the semantic similarity between the embedded test set response and each response in the training set for the same math problem using Euclidean distance. We then passed the top three most similar responses along with the feedback and score assigned by a teacher to our prompt, and continued to use a persona.

4 Results

We conducted evaluations with seven teachers who have 3 to 25 years of experience teaching middle school mathematics to assess the quality and effectiveness of the generated feedback. Each teacher received six unique problems, with five unique student responses per problem for a total of 42 problems and 210 unique responses. They were tasked with ranking the feedback from different methods (teacher-authored, zero-shot, random few-shot, and intelligent few-shot) and providing feedback on the quality and effectiveness of the responses.

4.1 Ranking

We first asked teachers to rank the feedback from their favorite to least favorite, and asked whether they felt any were similar. For evaluation, we removed rows where two feedbacks were considered similar.

To determine whether there were significant differences in the rankings of feedback methods, analyzed the mean reciprocal rank of each feedback type and performed a pairwise Wilcoxon signed-rank test to determine whether any feedbacks were different from one another. Table 1 shows the mean reciprocal rank for each model and 2 shows the pairwise test to determine which were different.

Our findings suggest that both few-shot prompts were ranked higher than the zero-shot prompt or teacher-authored feedback in ASSISTments, and while intelligent few-shot was ranked higher than average, we do not have evidence that its mean reciprocal rank was different than random few-shot. Similarly, zero-shot was higher than teacher on average but we do not have evidence it was significantly better than teacher authored feedback.

Table 1. Mean Reciprocal Rank for each prompt.

Prompt	MRR
Teacher	0.421
Zero	0.449
Random	0.588
Intelligent	0.625

Table 2. Pairwise Wilcoxon signed-rank test results between feedback types.

Prompt 1	Prompt 2	p-value
Teacher	Zero	0.199
Teacher	Random	¡ 0.001
Teacher	Intelligent	¡ 0.001
Zero	Random	¡ 0.001
Zero	Intelligent	¡0.001
Random	Intelligent	0.278

Acceptability, Learning Impact, and Performance. Following this, we asked teachers whether they thought each feedback was'acceptable','would cause students to learn', and'would result in students getting a higher score'. Each of these was binary. Table 3 shows the portion of feedback deemed acceptable, would cause learning, and would improve a students score.

Table 3. Proportion of each feedback that was deemed acceptable, to cause learning and improved performance.

Feedback	Acceptable	Learning	Performance
Teacher	0.39	0.27	0.26
Zero	0.50	0.52	0.65
Random	0.71	0.70	0.67
Intelligent	0.68	0.61	0.61

4.2 Acceptability, Learning Impact, and Performance Discussion

Our results show a surprising trend. We found that real teacher-written feedback in ASSISTments was rated significantly worse than any LLM feedback across all three metrics. Further, neither few-shot method performed much better than the other. The zero-shot prompt, notably, was rated as acceptable less often than both few-shot prompts, but would result in a higher score equally as often as

either few-shot prompt. In order to explain this, we dive into the last questions we asked teachers of'Why did you rank your top choice as the best?' and'Why did you rank you bottom choice as the worst?'

4.3 Reasoning

We reviewed the reasons for why teachers rated a feedback as the best or worst. We developed a codebook based on our finds. The categories we found were excellent, actionable, clarifying, concise, personalized, encouraging, and default. Table 4 shows the number of times each feedback was ranked as the best and how often teachers reasons related to one of the above categories.

Table 4. Distribution of reasons for positive feedback across feedback types.

Reason	Teacher	Zero-shot	Random	Intelligent
Excellent	86%	91%	84%	91%
Actionable	33%	59%	43%	59%
Clarifying	20%	51%	29%	37%
Concise	60%	9%	28%	36%
Personalized	23%	19%	15%	22%
Encouraging	17%	19%	20%	9%
Default	13%	21%	23%	11%
Total	30	37	65	78

We developed a similar codebook for the worst feedbacks. The identified categories were incorrect, no help, too long, irrelevant, confusing, and rude. Table 5 shows the results.

Table 5. Distribution of reasons for negative feedback across feedback types.

Reason	Teacher	Zero-shot	Random	Intelligent
Incorrect	37%	35%	52%	87%
No help	72%	21%	24%	44%
Too long	1%	55%	8%	0%
Irrelevant	14%	16%	36%	25%
Confusing	9%	15%	8%	13%
Rude	14%	10%	8%	0%
Total	101	68	25	16

4.4 Reasoning Discussion

When teacher feedback was rated highly, it was almost always because it did not say a lot. When teacher feedback was rated poorly, it was almost always because it did not say a lot. As many teachers have quite a few students and a lot of work to do they do not have time to write lengthy feedback the way LLMs do. As a result, their messages are much shorter and often are just "Great work!". Teachers felt those were effective on correct student answers, but not so effective for learning or higher score, especially for incorrect student answers. The zero-shot prompt suffered the opposite problem–it was too long. There were occasions where it included every step needed to solve a problem. As a result, teachers felt student who received this feedback would often get a higher score and any misconceptions clarified, but also it may give away too much information quite frequently.

The few-shot prompts had tradeoffs. Random few-shot suffered a risk of provided irrelevant information, possibly due to the randomness of the shots provided. However, it was often seen as effective, and particularly for causing student learning as it could tie in feedback from a variety of mistakes since it had more diverse shots. Intelligent few-shot feedback was rarely rated the worst, but when it was often it was due to some math error. However, it typically provided high-quality, actionable feedback to students. As many OLPs have been around for a large number of years and have a significant amount of data stored, we believe few-shot is viable for many platforms aiming to automate effective feedback.

References

1. Baral, S., et al.: Automated feedback in math education: a comparative analysis of llms for open-ended responses. arXiv preprint arXiv:2411.08910 (2024)
2. Brown, G.T., Peterson, E.R., Yao, E.S.: Student conceptions of feedback: impact on self-regulation, self-efficacy, and academic achievement. Br. J. Educ. Psychol. **86**(4), 606–629 (2016)
3. Cavalcanti, A.P., et al.: Automatic feedback in online learning environments: a systematic literature review. Comput. Educ. Artif. Intell. **2**, 100027 (2021)
4. Douze, M., et al.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
5. Hattie, J.: Visible learning: a synthesis of over 800 meta-analyses relating to achievement. Routledge (2008)
6. Nguyen, H.A., Stec, H., Hou, X., Di, S., McLaren, B.M.: Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In: Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T. (eds.) Responsive and Sustainable Educational Futures, pp. 278–293. Springer Nature Switzerland, Cham (2023)
7. Rui, M., Ye, L., S.R.J.C.X.Y.Z.S.Y.: Sfr-embedding-2: advanced text embedding with multi-stage training (2024). https://huggingface.co/Salesforce/SFR-Embedding-2_R