

Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot Learning

Dan Carpenter¹, Wookhee Min¹, Seung Lee¹,
Gamze Ozogul², Xiaoying Zheng², James Lester¹

{dcarpen2, wmin, sylee, lester}@ncsu.edu
gozogul@indiana.edu, zhengl2@iu.edu

¹ North Carolina State University

² Indiana University

Abstract

The practice of soliciting self-explanations from students is widely recognized for its pedagogical benefits. However, the labor-intensive effort required to manually assess students' explanations makes it impractical for classroom settings. As a result, many current solutions to gauge students' understanding during class are often limited to multiple choice or fill-in-the-blank questions, which are less effective at exposing misconceptions or helping students to understand and integrate new concepts. Recent advances in large language models (LLMs) present an opportunity to assess student explanations in real-time, making explanation-based classroom response systems feasible for implementation. In this work, we investigate LLM-based approaches for assessing the correctness of students' explanations in response to undergraduate computer science questions. We investigate alternative prompting approaches for multiple LLMs (i.e., Llama 2, GPT-3.5, and GPT-4) and compare their performance to FLAN-T5 models trained in a fine-tuning manner. The results suggest that the highest accuracy and weighted F1 score were achieved by fine-tuning FLAN-T5, while an in-context learning approach with GPT-4 attains the highest macro F1 score.

1 Introduction

Interactivity is critical to learning (Blasco-Arcas et al. 2013; Herppich et al. 2016). It has been widely demonstrated that by increasing interactivity in the classroom, we can significantly improve students' learning outcomes (Beauchamp and Kennewell

2010; Mayer et al. 2009). Student-teacher interaction is one of the most influential factors in learning (Beauchamp and Kennewell 2010), and when classrooms are interactive, students become more engaged, more participative, and are more motivated to learn (Bachman and Bachman 2011; Barnett 2006; Caldwell 2007). In addition, interactivity can improve comprehension and lead to improved learning (Freeman et al. 2014). Despite these benefits, many STEM classrooms use lectures as the primary method of instruction. The lack of interactivity poses serious issues in undergraduate education (Freeman et al. 2014), and large class sizes can inhibit meaningful exchanges between instructors and students in traditional classrooms (Caldwell 2007). The passive nature of lectures is particularly problematic in STEM courses, as research shows that undergraduate students in classes that use a traditional lecture format are much more likely to fail than students in classes that use a more active learning method (Freeman et al. 2014).

Classroom response systems have been touted as a potential solution to this problem. These systems capture and grade student responses to multiple choice questions posed by instructors during lectures. Each student submits a response using a handheld transmitter (a "clicker"), and software on the instructor's computer records, grades, and displays students' answers for the class to view. While research has shown that classroom response systems can promote student engagement and facilitate the learning of factual knowledge (Campbell and Mayer 2009; Hunsu et al. 2016), studies have also shown that "clickers" are less effective for promoting deep and meaningful learning. In fact, traditional classroom response

systems may actually obstruct students from developing a conceptual understanding of concepts and principles, particularly for novice students (Shapiro et al. 2017). Because students simply select an answer from a list of choices, “clickers” do not enable students to construct or generate their own responses to questions, which is a key component of active and constructive learning (Chi and Wylie 2014).

Decades of research have shown that self-explanation has a significant impact on student learning (Chi et al., 1994; Fonseca and Chi 2011). By explaining concepts and examples to themselves as they learn, students trigger the self-explanation effect, where they actively probe their own understanding and address gaps in their knowledge. Enabling students to generate short-answer textual explanations to prompts posed by instructors during lectures could open a rich communication channel between instructors and students. Eliciting self-explanations from students has the potential to yield substantial learning benefits for students in undergraduate STEM classrooms, and it has been widely demonstrated that self-explanation helps students learn much more effectively than students who do not engage in self-explanation (Chi et al., 1994; Fonseca and Chi 2011; Johnson and Mayer 2010; Roy and Chi 2005). Because self-explanation requires students to explain concepts to themselves in their own words, they learn much more deeply. However, despite the great potential offered by self-explanation for promoting learning, students in undergraduate STEM classrooms often have limited opportunities to engage in this type of active and constructive learning activity due to limited class time for discussion. Similarly, instructors have limited time to assess students’ self-explanation responses and provide formative and timely feedback during lectures.

In this paper, we present a large language model-based approach that automatically assesses students’ written responses. We investigate the performance of four Transformer-based large language models—Llama 2 (Touvron et al. 2022), GPT-3.5 (OpenAI 2023), GPT-4 (OpenAI 2023), and FLAN-T5 (Chung et al. 2022)—in assessing the correctness (i.e., fully correct, partially correct, and incorrect) of student self-explanations to undergraduate computer science questions. These explanations were collected from undergraduate students, including those who participated in an

undergraduate course using the EXPLAINIT system we have developed. Our findings suggest that FLAN-T5 demonstrates high performance in terms of accuracy and weighted F1, when fine-tuned using a prompt that includes information taken from a grading rubric in combination with an exemplar response provided by the instructor. However, we also find that the highest macro F1 score is achieved by GPT-4 in a few-shot learning setting, where examples of only ten students’ explanation responses are provided without any additional information from a rubric or an exemplar response. We discuss the tradeoffs between these models and the implications of our research for practical applications of LLM-based explanation assessment in classroom response systems.

2 Related Work

It has been found that students explaining concepts to themselves has a profound effect on learning. Known as the self-explanation effect (Chi et al. 1994; Fonseca and Chi 2011; Sidney et al. 2015), the result of self-explanation goes beyond simply rehearsing information: it requires students to express concepts in their own words, relate concepts to prior knowledge, make inferences, integrate information with prior knowledge, and monitor and repair faulty knowledge. Thus, self-explanation is a deeply constructive activity (Roy and Chi 2005). The significant learning gains associated with self-explanation have been demonstrated in a wide range of STEM disciplines including computer science (Pirulli and Recker, 1994), engineering (Johnson and Mayer 2010), chemistry (Crippen and Earl 2007), algebra (Atkinson et al. 2003), biology (McNamara 2004), physics (Chi et al. 1994), and physiology (Butcher 2006). Our EXPLAINIT classroom response system leverages the self-explanation effect to improve STEM classroom learning.

Widely known as “clickers,” classroom response systems have emerged as a tool to bridge the gap between students and instructors and to make lectures more interactive. Used by millions of students, classroom response systems allow students to anonymously respond to multiple choice questions presented during lectures. Research has shown that students appreciate the ability to compare their own answers to those of their peers, receive immediate feedback, and test their knowledge, and that “clickers” can increase student interactivity during lectures (Freeman et al.

2014; Hunsu et al. 2016; Kay and LeSage 2009). However, studies have also shown that clickers fail to promote deep and meaningful learning, which can be particularly problematic for students in STEM classes who are required to conceptually understand important concepts, relationships, and theories to effectively solve problems (Shapiro et al. 2017). Closely related to our work, commercial classroom response systems have been explored in various classroom settings. These systems typically support students through classroom discussions, questions, and assignments, and they support instructors with features for course material creation and assessment, which are incorporated with learning management systems. While they provide a range of functionalities required for a classroom response system, such as the ability to pose various types of questions (e.g., multiple choice, fill-in-the-blank, short answer questions), their automated assessment is typically limited to multiple choice and fill-in-the-blank types of questions that accept a predetermined set of answers, while they require a manual assessment process for other types of questions.

Deep learning-based language models such as BERT (e.g., Liu et al. 2019), FLAN-T5 (e.g., Chung et al. 2022), GPT (e.g., Brown et al. 2020), and Llama (Touvron et al. 2023) have been pivotal in the recent advancements in natural language processing (NLP; Torfi et al. 2020). In learning analytics, additional sources of training data, including data collected for free-response prompts (Rivera-Bergollo et al. 2022), text providing additional context for free-response prompts (Condor et al. 2021), response assessment rubrics (Condor et al. 2022), and synthetic data generated via data augmentation strategies (Lun et al. 2020), have effectively enhanced the training of NLP models, leading to improved predictive performance. NLP techniques have been used to accurately analyze student textual responses in the context of short-answer science assessment (Smith et al. 2019), student written reflections (Carpenter, Geden, et al. 2020), student-tutor dialogue (Carpenter, Emerson, et al. 2020), and student self-explanations (Chen and Wang 2022).

While previous work demonstrated considerable success with LLMs for short answer grading (Takano and Ichikawa 2022; Zhang et al. 2022) and short answer question generation (Moore et al. 2022), a research area that has seen limited exploration is assessing students' free-text

explanations (Nicula et al. 2023). Building on recent advances in NLP and deep learning-based language modeling techniques, our work makes a novel contribution by investigating an approach to assess students' self-explanations, collected from an undergraduate Artificial Intelligence course, utilizing large language models with fine-tuning and few-shot learning.

3 EXPLAINIT Classroom Response System

The EXPLAINIT classroom response system leverages the self-explanation effect and active, constructive, and interactive learning, along with state-of-the-art natural language processing, to significantly improve STEM undergraduate education. With a specific focus on computer science, biology, and physics, it aims to create highly engaging classroom learning experiences. EXPLAINIT offers the opportunity to fundamentally improve classroom dynamics by supporting both students and instructors. The system is designed to support both students and instructors in undergraduate STEM courses by analyzing and providing feedback on students' explanations through an integrated five-step explanation feedback loop (Figure 1): (1) the instructor issues an explanation prompt, which appears in the EXPLAINIT app on students' computing devices (e.g., laptops, tablets, phones); (2) students write free-text explanations ranging from a sentence to a short paragraph in the EXPLAINIT app on their computing devices; (3) EXPLAINIT automatically analyzes students' explanations and provides real-time formative feedback to students individually in their apps; (4) EXPLAINIT provides a summary of correctness of student explanations to the instructor; and (5) the instructor makes "instructional pivots" by immediately tailoring pedagogy to respond to students' explanations to improve student learning and engagement by focusing the lecture and classroom discussion on the most important elements of the course material. Collectively, these interactive explanation-based activities are designed to synergistically lead to improved student learning and promote greater student engagement in undergraduate STEM classrooms.

Our initial prototype of the EXPLAINIT classroom response system was implemented using a web-based application architecture to support enhanced scalability, where instructors

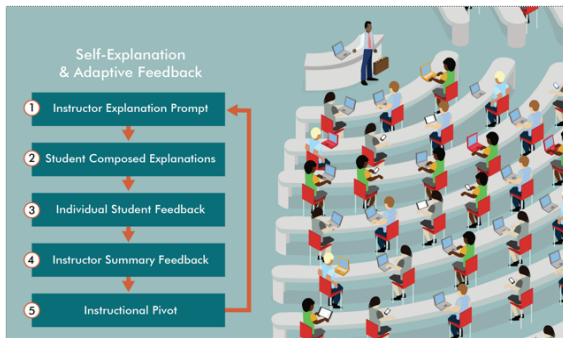


Figure 1: The EXPLAINIT explanation-based classroom response system.

and students can choose their platforms of choice such as laptops and handheld devices, while the software enables real-time interactions with the user interfaces. The EXPLAINIT user interfaces consist of an Instructor Authoring Tool, Instructor Dashboard, and Student Explanation App. The streams of communication data are uploaded into a cloud-based database by the server. For data synchronization and analysis purposes, all interaction data is timestamped. To support these functionalities, we implemented the software modules to include APIs using the HTTP protocol. We use Microsoft's Azure cloud computing service to host our cloud-based services.

The Instructor Authoring Tool enables instructors to create or edit questions and an exemplar correct response per question. All questions and responses are categorized by their subjects and topics in the tool. All authored content is stored and accessed from the cloud, allowing the original instructor to reference their own created questions for future courses. The Instructor Dashboard presents the pool of questions per subject and topic, and it allows instructors to select and send questions to the Student Explanation App, so that students can view and interact with the questions in real-time during lectures. The Instructor Dashboard is also designed to display student-written responses and NLP assessment results in visual analytics. The Student Explanation App enables students to receive questions posed by instructors and write self-explanation responses to instructor-posed questions. When students submit their responses, the Student Explanation App taps into the Explanation Analyzer, which performs NLP-driven assessment of student explanations, generates tailored feedback to students, and dispatches analytical summaries to instructors

through Instructor Dashboard. The Explanation Analyzer is in the development phase, and our findings about the Explanation Analyzer's NLP performance are presented in this paper.

4 Study and Data

This work uses data collected during a classroom pilot study of the EXPLAINIT system. The participants in the classroom study consisted of 36 consented undergraduate students enrolled in a Computer Science course focused on Artificial Intelligence. Thirty-two students completed a demographic pre-survey, and among them 8 indicated that they identified as female, 23 as male, and 1 preferred not to indicate gender identification. Participants ranged in age from 18 to 28 ($M = 21.1$, $SD = 1.64$). Of these participants, 40.6% were Asian, 50.0% were White, and 9.4% preferred not to answer.

Prior to using EXPLAINIT in the class, the instructor used the Instructor Authoring Tool to prepare a set of questions, each accompanied by an exemplar correct answer. These answers were presented to students immediately after they submitted their responses to the respective questions. The classroom implementation unfolded over 6 weeks within a single semester. Throughout this period, a total of 13 questions were sent to the class, eliciting 356 responses from 36 participants, which were utilized in our evaluation (Table 1).

Students' responses to the questions were labeled by two of the researchers, who are experts in computer science. First, a rubric item was constructed for each question that described the qualities of a correct, partially correct, or incorrect answer to the question. For example, the rubric for the question "In a neural network, what function is responsible for introducing non-linearity to the model?" indicated that a correct response should mention the term "activation function", a partially correct response might present an example of an activation function (e.g., "sigmoid") without explicitly mentioning the term "activation function", and that an incorrect response would not include any of this information. We also referenced instructor-provided exemplar answers to further refine the rubrics for each question. These were comprehensive and well-reasoned responses, serving as a representative correct answer to each question.

Then, based on the developed rubric, both researchers labeled twenty percent of the student

Question	Topic	Number of Questions Sent	Number of Student Responses
What does the term "deep" in deep learning refer to?	Deep Learning	1	28
What is the basic building block of a neural network called?	Deep Learning	1	27
In a neural network, what function is responsible for introducing non-linearity to the model?	Deep Learning	1	24
What is clustering in the context of machine learning?	Clustering	2	37
Name a commonly used algorithm for clustering and briefly describe how it works.	Clustering	2	34
What is the main difference between K-means and hierarchical clustering?	Clustering	2	37
The K-means algorithm may end up with different clustering results when the initial clustering centers are chosen differently. Yes or No?	Clustering	2	36
What is the "purity" of an external measure for cluster quality?	Clustering	1	22
What are support vectors in the context of SVMs	SVM	1	23
How does a soft-margin SVM differ from a hard-margin SVM?	SVM	1	23
Is it always better to use a soft-margin SVM to ensure model flexibility? Why?	SVM	1	22
Is an SVM more suitable for small datasets than large datasets? Why?	SVM	1	23
Can SVMs be used for both classification and regression tasks? Example?	SVM	1	20

Table 1: Descriptive statistics of questions sent during the classroom study.

responses. After one cycle of rubric refinement, a Cohen’s Kappa of 0.702 was achieved, indicating substantial agreement (McHugh 2012). All labels that the annotators did not agree on were discussed and agreement on a single label was reached. Across all questions, 73% of explanations were labeled as correct, 22% were labeled as partially correct, and 5% were labeled as incorrect.

5 Method

We evaluated the performance of Llama 2 (Touvron et al. 2022), GPT-3.5 (OpenAI 2023), GPT-4 (OpenAI 2023), and FLAN-T5 (Chung et al. 2022) on the self-explanation assessment task. Large language models (LLMs) have been demonstrated to achieve state-of-the-art performance on many natural language processing tasks, with GPT-4 particularly excelling with few-shot prompting where training examples are integrated into the task description (OpenAI 2023). This enables GPT-4 to readily adapt to new tasks without re-training, avoiding the prohibitive cost of

updating its extensive parameters. However, GPT models’ proprietary nature and associated costs pose barriers to its educational adoption, such as EXPLAINIT.

To address this challenge, we also evaluated the performance of open-source models, FLAN-T5 and Llama 2. FLAN-T5 is an instruction-fine-tuned language model that has demonstrated competitive performance with other state-of-the-art models across a range of tasks when it was released (Chung et al. 2022). Llama 2 is an open-source pre-trained large language model that has demonstrated leading performance compared to other open-source models and performs similarly to GPT-3.5 on several tasks (Touvron et al. 2023). In this work, we investigate the performance of the base FLAN-T5 model (250M parameters) and Llama 2-7B, the smallest version of the model. These versions of FLAN-T5 and Llama 2 were selected due to their computational efficiency. For all models, default hyperparameters were used.

We investigated several different zero-shot and few-shot prompting approaches to evaluate the

performance of Llama 2, GPT-3.5, and GPT-4 for automated assessment of students' self-explanation. As a baseline, these models were provided with instructions that described the task (i.e., "Please evaluate a student's explanation response to the following question.") in addition to the question and student response. Then, we systematically evaluated the impact on model performance of including the following information in the prompt: (a) rubric items for the current question, (b) an exemplar correct response provided by the instructor, and (c) other students' labeled responses to the current question. Prompts were constructed with all possible combinations of the different information elements, and model performance was evaluated for each combination.

For the prompts incorporating student self-explanation responses, 10-fold student-level cross-validation was used to prevent bias from students' individual writing styles and to ensure generalizability, avoiding data leakage in model evaluation. Additionally, this approach accurately represents the real-world scenario that will be faced when deploying EXPLAINIT in future classroom implementations, as the students interacting with the system will be new but the models will have access to past student's responses to each question. Due to LLM token limits and the per-token cost of proprietary models like GPT-4, we sampled ten responses from the training set for each cross-validation fold to include in the prompts rather than including the entire training set.

In comparison to Llama 2, GPT-3.5, and GPT-4, the FLAN-T5 base model's smaller parameter count facilitates easier and more cost-effective training. Given its sufficient size for fine-tuning using our available resources, we chose this approach over few-shot prompting. We applied LoRA for efficient fine-tuning, changing only a subset of the model's parameters to conserve time and computational resources, while achieving similar performance to full fine-tuning (Hu et al. 2021). The evaluation of fine-tuned FLAN-T5 models is also based on 10-fold student-level cross-validation using the same data split as was used for in-context learning with the other models. However, rather than including example explanations and their assigned labels in the prompt, they were used as training examples in a supervised learning approach. As with the in-context learning approach, we explored variants of prompts including the rubric item for each question

and/or the exemplar correct response created by the instructor. A separate FLAN-T5 model was fine-tuned for each prompt variant.

6 Results

Results from all experiments are presented in Table 2. Our task involves multi-class classification, where each student response is categorized into *correct*, *partially correct*, or *incorrect*. We evaluated the explanation assessment models in terms of accuracy, macro F1, and weighted F1. As noted above, all combinations of the three different information elements (i.e., rubric, exemplar response, and student example responses) were explored for each LLM. Due to length constraints, Table 2 reports only the results of including one element at a time as well as including all types of information, while the findings from all combinations are discussed in this paper.

Across all experiments, FLAN-T5 models that were fine-tuned with rubric information and the instructor's exemplar response achieved the highest accuracy (acc.=0.824). This was a substantial improvement over the majority baseline, which always predicts the most common class (acc.=0.730), as well as the next-highest performing approach, which was GPT-4 with ten student examples included in the prompt (acc.=0.775). In terms of macro F1 score, GPT-4 with ten labeled student explanation responses included in the prompt achieved the highest performance (F1=0.664). This was a significant improvement over the majority baseline (F1=0.281) and the next-highest performing approach, which was GPT-4 with all three information elements included in the prompt (F1=0.641). In terms of weighted F1 score, FLAN-T5 models that were fine-tuned with rubric information and the instructor's exemplar response achieved the highest performance (F1=0.798). This was an improvement over the majority baseline (F1=0.616) and a small improvement over the next-highest performing approach, which was GPT-4 with ten labeled student explanation responses included in the prompt (F1=0.792).

In general, our results demonstrate that including rubric information in the prompt improved model performance. For FLAN-T5, Llama 2, and GPT-4, both accuracy and F1 score were improved relative to the prompting approach that only provided high-level instructions for the explanation assessment task. We observed the

largest improvement in model performance when the sole additional information was a set of ten labeled explanation responses from other students. With this prompt, Llama 2 and GPT-4 demonstrated improved accuracy over the instruction-only approach, while Llama 2, GPT-3.5, and GPT-4 exhibited improved macro F1 scores. However, we found that including the instructor’s exemplar response into the prompt led to reduced model performance across all models except for Llama 2, compared to the instruction-only approach. This reduction may stem from the exemplar responses often containing comprehensive details that exceed the question’s scope, leading the models to apply a very strict standard in assessing student responses. Consequently, responses were more frequently categorized as partially correct or incorrect, even though they should be labeled correct within the question’s intended scope.

Next, we looked at the effects of including two information elements in the prompt. Note that these results are omitted from Table 2 to save space. We observed that the highest accuracy and F1 score for

FLAN-T5 were achieved when the models had access to both rubric information and the instructor’s exemplar response. That is, we found that there was an additive effect of including multiple information elements for FLAN-T5 models. In comparison, the general trend across the prompting approaches for Llama 2, GPT-3.5, and GPT-4 that utilized two information elements was that there was not an additive benefit of including multiple information elements. For example, GPT-3.5 and GPT-4 including either rubric information or the exemplar response in addition to labeled student responses led to reduced performance compared to models that only had access to ten student example responses. In addition, Llama 2 generally demonstrated a decrease in performance when using two information elements compared to only one; however, the combination of the exemplar response and ten student responses without the rubric led to improved performance over all approaches that incorporated only one information element.

A distinct trend emerged when all three information elements were included in the prompt.

Model	Prompt Variation	Accuracy	F1 (macro)	F1 (weighted)
Majority Baseline	--	0.730	0.281	0.616
FLAN-T5-base (250M)	Fine-tuned with instructions	0.803	0.476	0.764
	Fine-tuned with instructions + Rubric	0.820	0.506	0.789
	Fine-tuned with instructions + Exemplar response	0.792	0.465	0.754
	Fine-tuned with instructions + Rubric + Exemplar response	0.824	0.550	0.798
Llama 2-7B	Instructions only	0.509	0.184	0.538
	Instructions + Rubric	0.664	0.400	0.698
	Instructions + Exemplar response	0.526	0.234	0.579
	Instructions + 10 student examples	0.706	0.443	0.717
	Instructions + Rubric + Exemplar response + 10 student examples	0.744	0.444	0.751
GPT-3.5	Instructions only	0.664	0.545	0.684
	Instructions + Rubric	0.564	0.449	0.586
	Instructions + Exemplar response	0.519	0.425	0.539
	Instructions + 10 student examples	0.612	0.591	0.644
	Instructions + Rubric + Exemplar response + 10 student examples	0.533	0.537	0.560
GPT-4	Instructions only	0.685	0.574	0.708
	Instructions + Rubric	0.709	0.606	0.732
	Instructions + Exemplar response	0.651	0.422	0.686
	Instructions + 10 student examples	0.775	0.664	0.792
	Instructions + Rubric + Exemplar response + 10 student examples	0.754	0.641	0.779

Table 2: Student explanation assessment results across models and prompt variations.

GPT-3.5 and GPT-4 models with access to all three information elements performed worse than models provided with only ten labeled student example responses, both in terms of accuracy and F1 scores. However, for Llama 2 models, incorporating all three information elements in the prompt resulted in the highest accuracy and F1 scores compared to any other combinations of information.

These results suggest that the best results are not necessarily guaranteed by providing the model with the maximum amount of task-related information. Models consistently performed well when the prompt included labeled examples of other students' responses, but including the instructor-created exemplar response tended to reduce model performance as discussed. Adjusting the exemplar response provided to the models, by adding clarification or simplifying its content, could potentially lead to improved performance when this information element is included. This underscores an important area for future research.

Overall, these results demonstrate that fine-tuning FLAN-T5 and utilizing few-shot learning with GPT-4 are both viable approaches to this explanation assessment task. Although FLAN-T5 requires more training data than GPT-4 to reach high performance levels (our preliminary analysis indicated that the predictive accuracy of a FLAN-T5 model, fine-tuned with only the data from five focus group students, was 60%), this tradeoff may be acceptable considering that FLAN-T5 is open-source and GPT-4 is proprietary. This consideration becomes more critical as our classroom implementation scales, especially in large classroom settings with multiple sessions where deployment costs become a significant factor. Conversely, if the EXPLAINIT system is implemented in a course where FLAN-T5 models have not been trained with student data from that course, GPT-4 with one-shot learning (with rubric information) might significantly outperform FLAN-T5, making GPT-4 potentially more suitable for the classroom response system. It will be crucial to weigh practical benefits, scalability, and cost considerations when deploying a runtime version of the explanation assessment system during the classroom use of EXPLAINIT. In practice, these results suggest that a hybrid system may be a viable approach. When a new question is deployed using the system, zero-shot learning with GPT-4 can be used based on a pre-defined rubric that was

created for assessing responses to the question. Since this information can be created at the same time as the question, it can be provided to the system when the new question is first deployed. Then, as student responses to the question are collected, they can be used to fine-tune a FLAN-T5 model, which can then replace the GPT-4 model once it starts showing superior performance.

7 Conclusion

Prompting students to craft self-explanations has demonstrated to offer numerous educational advantages. However, it often requires substantial time and effort necessary for instructors to manually assess student responses and provide feedback for students, which renders them unsuitable in large classroom environments. To address this challenge, we present EXPLAINIT, a self-explanation-based classroom response system specifically designed to encourage students in formulating written self-explanations during undergraduate STEM lectures. Our NLP framework builds on Transformer-based large language models, such as FLAN-T5 and GPT-4, in assessing the correctness of student explanations, and it is evaluated using our dataset collected from classroom interactions with the EXPLAINIT system. Results demonstrate that fine-tuned FLAN-T5 models using prompts with rubric information and an exemplar response achieved the highest accuracy and weighted F1 score, while few-shot prompting that provided GPT-4 with ten labeled student response examples achieved the highest macro F1 score. These results indicate the potential to use large language models for automated explanation assessment, which can be leveraged to provide adaptive support for students' self-explanations in classroom environments.

Moving forward, there are several promising directions for future work. First, it will be important to implement the full suite of EXPLAINIT system functionalities, including NLP assessment models, in a classroom environment and investigate their impact on students' learning outcomes. It would also be interesting to incorporate AI capabilities to support question and rubric generation, thereby reducing the amount of work required by instructors to use EXPLAINIT in their classes. Additionally, the explanation assessment system could be expanded to support a finer-grained assessment of students' self-explanations. For example, concept-level assessment of students'

self-explanations could provide more insightful feedback for both students and instructors. Also, it will be important to investigate this explanation assessment approach in disciplines other than computer science to evaluate its performance in other domains. Finally, it will be important to explore how different types of exemplar responses and rubric items impact model performance. If we are able to identify characteristics of exemplar responses and rubric items that most improve the predictive accuracy of our LLM-based framework for self-explanation assessment, that will enable our classroom response system to more effectively support student learning in new settings where there is limited student data that can be used to inform the assessment models.

8 Limitations

One limitation of our work is the challenge associated with evenly comparing fine-tuned models (i.e., FLAN-T5) with models that are evaluated based on few-shot in-context learning (i.e., Llama 2, GPT-3.5, and GPT-4). In our work, FLAN-T5 had access to 90% of the dataset as training data because of the 10-fold student-level cross-validation setup. In contrast, while the models that used in-context learning used the same cross-validation setup, they had access to only ten student responses that were sampled from the training set for each cross-validation fold. This limitation was a result of the practical consideration that LLMs have limited context lengths and that proprietary LLMs have monetary costs on a per-token basis. As a result, it is not feasible to provide an unlimited number of labeled student explanation responses in the prompt to an LLM, and the limit of ten student responses was chosen because it seemed reasonable. To overcome this limitation, future work could systematically investigate whether there is a more optimal number of example student responses that balances between model performance and costs. Another limitation of this work is the generalizability of the result suggesting that including an exemplar response created by the instructor in the prompt led to reduced model performance. It may be the case that certain characteristics of the exemplar responses used in this work were suboptimal for providing an LLM with guidance on how to correctly assess students' explanation responses. Further investigation into the impacts of various characteristics of exemplar

responses would be helpful for addressing this limitation.

Acknowledgments

This research was supported by funding from the National Science Foundation under Grants DUE-2111473 and DUE-2111216. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Leonard Bachman and Christine Bachman. 2011. A study of classroom response system clickers: Increasing student engagement and performance in a large undergraduate lecture class on architectural research. *Journal of Interactive Learning Research*, 22(1):5-21.
- John Barnett. 2006. Implementation of personal response units in very large lecture classes: Student perceptions. *Australasian Journal of Educational Technology*, 22(4):474-494. <https://doi.org/10.14742/ajet.1281>.
- Gary Beauchamp and Steve Kennewell. 2010. Interactivity in the classroom and its impact on learning. *Computers & Education*, 54(3):759-766.
- Lorena Blasco-Arcas, Isabel Buil, Blanca Hernández-Ortega, and F. Javier Sese. 2013. Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers & Education*, 62:102-110.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877-1901.
- Kirsten Butcher. 2006. Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, 98(1):182.
- Jane Caldwell. 2007. Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education*, 6(1):9-20.
- Julie Campbell and Richard Mayer. 2009. Questioning as an instructional method: Does it affect learning from lectures?. *Applied Cognitive Psychology*, 23(6):747-759.
- Dan Carpenter, Andrew Emerson, Bradford Mott, Asmalina Saleh, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Proceedings of the*

- Dan Carpenter, Michael Geden, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. Automated analysis of middle school students' written reflections during game-based learning. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*.
- Xinyue Chen and Xu Wang. 2022. Scaling Mixed-Methods Formative Assessments (mixFA) in Classrooms: A Clustering Pipeline to Identify Student Knowledge. In *Proceedings of the International Conference on Artificial Intelligence in Education*.
- Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219-243.
- Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439-477.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, ... and Jason Wei. 2022. Scaling instruction-finetuned language models. arXiv:2210.11416.
- Aubrey Condor, Max Litster, and Zachary Pardos. 2021. Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of the International Conference for Educational Data Mining*.
- Aubrey Condor, Zachary Pardos, and Marcia Linn. 2022. Representing scoring rubrics as graphs for automatic short answer grading. In *Proceedings of the International Conference on Artificial Intelligence in Education*.
- Kent J. Crippen, and Boyd L. Earl. 2007. The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy. *Computers & Education*, 49(3):809-821.
- Brenda A. Fonseca, and Michelene TH Chi. 2011. Instruction based on self-explanation. In *Handbook of Research on Learning and Instruction*, edited by R. E. Mayer and P. A. Alexander, pages 296-321. New York, NY: Routledge.
- Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. In *Proceedings of the National Academy of Sciences*.
- Stephanie Herppich, Jörg Wittwer, Matthias Nückles, and Alexander Renkl. 2016. Expertise amiss: interactivity fosters learning but expert tutors are less interactive than novice tutors. *Instructional Science*, 44(3):205-219.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv:2106.09685.
- Nathaniel Hunsu, Olusola Adesope, and Dan Bayly. 2016. A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, 94:102-119.
- Cheryl I. Johnson and Richard E. Mayer. 2010. Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26(6):1246-1252.
- Robin H. Kay and Ann LeSage. 2009. Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education*, 53(3):819-827.
- Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Automatic short answer grading via multiway attention networks. In *Proceedings of the International Conference on Artificial Intelligence in Education*.
- Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Richard E. Mayer, Andrew Stull, Krista DeLeeuw, Kevin Almeroth, Bruce Bimber, Dorothy Chun, Monica Bulger, Julie Campbell, Allan Knight, and Hangjin Zhang. 2009. Clickers in college classrooms: fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, 34:51-57.
- Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276-282.
- Danielle S. McNamara. 2004. SERT: Self-explanation reading training. *Discourse Processes*, 38(1):1-30.
- Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *Proceedings of the European Conference on Technology Enhanced Learning*.
- Bogdan Nicula, Mihai Dascalu, Tracy Arner, Renu Balyan, and Danielle S. McNamara. 2023. Automated Assessment of Comprehension Strategies from Self-Explanations Using LLMs. *Information*, 14(10), 567.

- OpenAI. 2023. *GPT-4 Technical Report*.
- Peter Pirolli and Margaret Recker. 1994. Learning strategies and transfer in the domain of programming. *Cognition and Instruction*, 12(3):235-275.
- Raysa Rivera-Bergollo, Sami Baral, Anthony Botelho, and Neil Heffernan. 2022. Leveraging auxiliary data from similar problems to improve automatic open response scoring. In *Proceedings of the International Conference for Educational Data Mining*.
- Marguerite Roy, and Michelene TH Chi. 2005. The self-explanation principle in multimedia learning. In *The Cambridge Handbook of Multimedia Learning*, edited by R.E. Mayer, pages 271-286. Cambridge University Press.
- Amy M. Shapiro, Judith Sims-Knight, Grant V. O’Rielly, Paul Capaldo, Teal Pedlow, Leamarie Gordon, and Kristina Monteiro. 2017. Clickers can promote fact retention but impede conceptual understanding: The effect of the interaction between clicker use and pedagogy on learning. *Computers & Education*, 111:44-59.
- Pooja G. Sidney, Shanta Hattikudur, and Martha W. Alibali. 2015. How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, 40:29-38.
- Andy Smith, Osman Aksit, Wookhee Min, Eric Wiebe, Bradford Mott, and James Lester. 2016. Integrating real-time drawing and writing diagnostic models: An evidence-centered design framework for multimodal science assessment. In *Proceedings of the Thirteenth International Conference on Intelligent Tutoring Systems*.
- Andy Smith, Samuel Leeman-Munk, Angi Shelton, Bradford Mott, Eric Wiebe, and James Lester. 2019. A multimodal assessment framework for integrating student writing and drawing in elementary science learning. *IEEE Transactions on Learning Technologies*, 12(1):3-15.
- Shunya Takano and Osamu Ichikawa. 2022. Automatic scoring of short answers using justification cues estimated by BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv:2003.01200*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, ... and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic short math answer grading via in-context meta-learning. *arXiv:2205.15219*.