Improving In-context Learning Example Retrieval for Classroom Discussion Assessment with Re-ranking and Label Ratio Regulation

Nhat Tran
Diane Litman
Benjamin Pierce
Richard Correnti
Lindsay Clare Matsumura
University of Pittsburgh
Pittsburgh, PA, USA
{nlt26,dlitman,bep51,rcorrent,lclare}@pitt.edu

Abstract

Recent advancements in natural language processing, particularly large language models (LLMs), are making the automated evaluation of classroom discussions more achievable. In this work, we propose a method to improve the performance of LLMs on classroom discussion quality assessment by utilizing in-context learning (ICL) example retrieval. Specifically, we leverage example re-ranking and label ratio regulation, which forces a specific ratio of different types of examples on the ICL examples. While a standard ICL example retrieval approach shows inferior performance compared to using a predetermined set of examples, our approach improves performance in all tested dimensions. We also conducted experiments to examine the ineffectiveness of the generic ICL example retrieval approach and found that the lack of positive and hard negative examples can be a potential cause. Our analyses emphasize the importance of maintaining a balanced distribution of classes (positive, non-hard negative, and hard negative examples) in creating a good set of ICL examples, especially when we can utilize educational knowledge to identify instances of hard negative examples.

1 Introduction

The automatic evaluation of classroom discussion quality has emerged as a significant area of interest within educational research. A wide range of studies have established that the quality of classroom discourse plays a pivotal role in facilitating student learning and cognitive development (Desimone and and, 2017; Wilkinson et al., 2015; Suresh et al., 2019; Jacobs et al., 2022). Nevertheless, large-scale assessment of classroom discussions remains prohibitively resource-intensive and logistically challenging. The development of automated scoring systems offers a promising solution, enabling the generation of extensive datasets to investigate the mechanisms through which discourse shapes student reasoning and understanding. Furthermore, such systems hold the potential for integration into formative assessment practices, providing educators with actionable feedback to enhance the effectiveness of classroom discussions.

Compared to pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), large language models (LLMs) have been shown to be more reliable in scoring different dimensions of classroom discussion quality, based on the *Instructional Quality Assessment (IQA)* (Tran et al., 2024a). Prior LLM approaches for classroom discussion assessment have ranged from using zero-shot prompts (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024) that do not exploit the few-shot learning capability of LLMs (Brown et al., 2020), to utilizing few-shot prompts but with a fixed set of examples for every input (Tran et al., 2024a,b). Inspired by the advancement of in-context learning (ICL) *example retrieval* (Wang et al., 2024; Zhang et al., 2023), we attempt to automatically select few-shot examples based on a given input.

Our work thus aims to improve the automated scoring of classroom discussion quality with ICL example retrieval. Utilizing LLMs for binary prediction with a 'target' label (e.g., if we are identifying if a label y is present in the current turn, the target now is y), we define the types of examples as follows. If an example has the same label as the target label, it is a *positive* example, otherwise, it is *a negative* example. A *hard negative* example is a negative example that we expect will be difficult for a model to distinguish from positive examples, i.e., positive and hard negative examples are semantically similar in the input space but represent different classes in the output space. From a retrieval perspective, the hard negative examples are often selected based on some quantitative metrics such as their distance in the embedding space or their ranking from a reward model (Wang et al.,

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 752–764
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

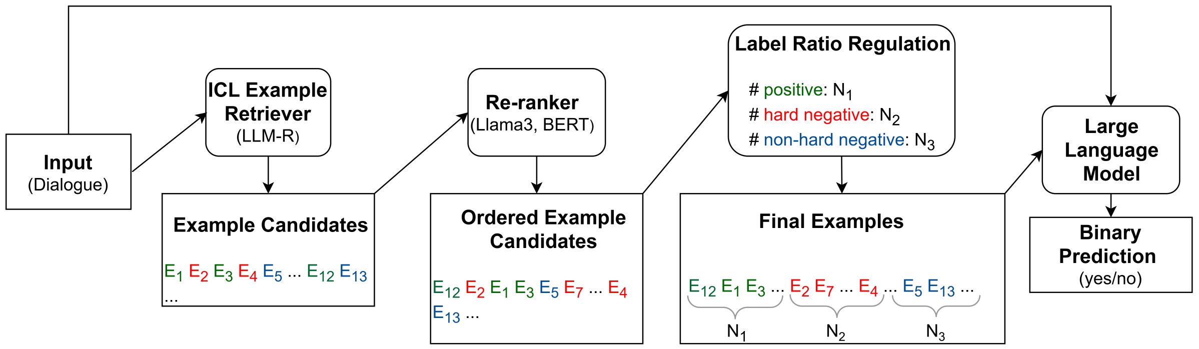

Figure 1: Overview of the proposed method.

2024; Zhang et al., 2023). However, in the context of classroom discussion, since we can leverage a qualitative metric (i.e., domain knowledge about the definition of the labels), we can identify hard negative examples for the target label more reliably. Specifically, based on the definitions of the labels, for a target label A, we know that label A' is closer to A compared to other labels from a human perspective. Therefore, when finding hard negative examples for A, we can quickly select instances with label A' as candidates without needing to calculate any kind of 'distance' between them.

After experimenting with a generic ICL example retrieval approach (LLM-R by Wang et al., 2024), we found it is ineffective for classroom discussion. We hypothesized that the problem is from 1) the imbalance of positive/negative examples and 2) the lack of hard negative examples, as we cannot control the retrieval process. The first hypothesis is well-known in ICL learning as we need both positive and negative examples to learn effectively (Min et al., 2022). The second hypothesis is from the observation that hard negative examples play a crucial role in getting good prediction performance (Tran et al., 2024a; Robinson et al., 2021). Moreover, although we have domain knowledge about hard negative examples based on the annotations of the labels, the ICL retriever only relies on quantitative metrics (e.g., higher-ranking examples) to identify them. To address these issues, we proposed a 2-step approach. First, we train a BERT-based re-ranker to re-order the retrieved examples from LLM-R (Wang et al., 2024). Second, we employ label ratio regulation (LRR), which selects examples from the sorted list while maintaining a specific ratio of positive, non-hard negative, and hard negative examples in the 10-example set used in the prompt (Figure 1).

Our goal is to answer these research questions:

RQ_1 Does the proposed method help improve performance?

RQ_2 Does ICL example retrieval have good coverage of the label space (type of examples)?

RQ_3 How does the ratio of the ICL examples used in the label ratio regulation influence the performance?

Our contributions are three-fold. First, we show that a standard ICL example retrieval approach, despite being useful for other natural language processing (NLP) tasks, is ineffective for classroom discussion assessment. Further analyses suggest that the lack of positive examples and hard negative examples can be causes for this poor performance. Second, we propose an approach utilizing re-ranking and label ratio regulation to complement the standard ICL example retrieval. It helps improve performance and yields comparable results to a finetuned retriever without finetuning the retriever. Third, we demonstrate that even with re-ranking, the retrieval process fails to effectively select hard negative examples, which emphasizes the importance of label ratio regulation when the domain knowledge of the classes (e.g., which class is a hard negative example) is available.

2 Related Work

2.1 LLMs for Classroom Discussion Scoring

As generative LLMs such as GPT-4 (OpenAI et al., 2024), Llama (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023) have outperformed PLMs in many NLP tasks, there has been growing interest in leveraging these LLMs for classroom discussions.

When predicting accountable talk moves in classroom discussions, a finetuned LLM such as GPT-3 has consistently surpassed RoBERTa (Liu et al., 2019) in precision (Kupor et al., 2023). However, since finetuning LLMs requires significant expertise, extensive data, and substantial computational resources, researchers have increasingly focused on zero-shot and few-shot approaches that do not require additional training. For instance, one study examined the zero-shot capabilities of ChatGPT in three tasks: scoring transcript segments using classroom observation instruments, identifying key strengths and missed opportunities in instructional strategies, and providing actionable suggestions for fostering student reasoning (Wang and Demszky, 2023). The findings revealed that ChatGPT struggled to score classroom transcripts using instruments like the Classroom Assessment Scoring System (CLASS) or the Mathematical Quality of Instruction (MQI) and offered repetitive feedback.

Research has also explored the application of LLMs at more granular levels, such as sentence-level or utterance-level analysis. While zero-shot ChatGPT provided clear and detailed explanations for its predictions, it performed significantly worse than the smaller BERT model in three out of four student talk move categories for the classification task (Wang et al., 2023). Tran et al. (2024b) analyzed three prompt-based factors: task formulations, context length, and the presence of few-shot examples and found that all of them can have impacts on the final performance. Although the importance of few-shot examples has been shown and some prior work utilized few-shot prompting, they had a fixed set of examples, which might not be representative enough and might not have examples relevant to the current input (Tran et al., 2024a,b). *Our work focuses on automatically retrieving a set of examples based on the input to cover dynamic scenarios in a classroom discussion.*

2.2 ICL Example Retrieval for LLMs

In-context learning, the emergent capability of LLMs that allows them to execute diverse tasks by conditioning on a limited set of input-output examples without requiring parameter updates or finetuning, has been demonstrated in many LLMs such as GPT-3 (Brown et al., 2020) or Llama (Touvron et al., 2023). Various approaches have been made to create better LLM prompts (Li and Liang, 2021; Le Scao and Rush, 2021; Hao et al., 2022). Different from the standard retrieval-augmented generation by using a dense retriever such as ColBERT (Khattab and Zaharia, 2020) to get additional information for LLMs, there is an area of research focused specifically on finding better ICL *examples* to boost LLMs' performance (Ye et al., 2023; Li and Qiu, 2023; Li et al., 2023; Zhang et al., 2023). Liu et al. (2022) demonstrated that ICL performance can be enhanced by either using the BM25 algorithm or by finetuning dense retrievers with feedback from LLMs to retrieve relevant examples from a training set. Wang et al. (2024) proposed an iterative training framework (LLM-R) to retrieve ICL examples in 3 steps: 1) rank an initial set of retrieved candidates based on the conditional LLM log probabilities of the ground-truth outputs; 2) train a cross-encoder reward model to capture the fine-grained ranking signals from LLMs; and 3) train a bi-encoder dense retriever using knowledge distillation. *Our work falls into this area by proposing a method to dynamically retrieve ICL examples. However, we focus on the label ratio of the example set, which has not been studied in prior work.*

3 Dataset

We use videos of English Language Arts classes in a Texas district to create our corpus. The videos were recorded during the course of an online instructional coaching program (Correnti et al., 2021). They were collected from 18 fourth-grade and 13 fifth-grade classes, whose teachers on average had 13 years of teaching experience. 61% of the student population was considered low income, with the following racial proportion: Latinx (73%), Caucasian (15%), African American (7%), multiracial (4%), and Asian or Pacific Islander (1%).

Annotators manually scored videos holistically, on a scale from 1 to 4, using the *IQA* on 11 dimensions (Matsumura et al., 2008) for both teacher and student contributions. They were also scored using more fine-grained talk moves annotated at the sentence level using the *Analyzing Teaching Moves (ATM)* discourse measure (Correnti et al., 2021). The final corpus consists of **112** discussion transcripts that have been scored using both the IQA and the ATM (see Appendix A for the statistics of the scores). Thirty-two videos (29%) were double-scored, showing good to excellent reliability for the IQA (the Intraclass Correlation Coefficients (ICC) range from .89-.98) and moderate to good reliability for the ATM (ICC range from .57 to .85). Below is a excerpt with annotated *ATM* codes:

> **Teacher**: [Justin.]$_{\text{Repeat}}$ [Tell me who's Justin?]$_{\text{Press}}$
>
> **Student**: [Justin is... Well, Via's boyfriend who stands up for August and

is very nice to him. Even though he saw him for the first time, he was kind of shocked, but he kind of got used to him.]Strong Explanation

IQA dimension scores. To compare with prior work (Tran et al., 2024a), we focused on **4** of the 11 IQA dimensions, in which 2 of them focus on teaching moves and 2 focus on student contributions. They were previously chosen because of their relevance to dialogic teaching principles that emphasize collaboration and active participation in meaning-making. Furthermore, all four scores are calculated based on the frequency of their related ATM codes. The four dimensions include: *Teacher links Student's contributions* (T-Link), *Teacher presses for information* (T-Press), *Student links other's contributions* (S-Link), *Student supports claims with evidence and explanation* (S-Evid). We define S-Evid as the higher score of *Student provides text-based evidence* and *Student provides explanation*. Descriptions of these dimensions can be found in Appendix B.

Based on the definitions of the ATM codes (Appendix C), 2 IQA dimensions have hard negative examples (e.g., have examples that are semantically similar to the positive examples but have a different label based on a notion of strength). For S-Link, a positive example has *Strong Link* as the ATM label while a hard negative example has *Weak Link*. A similar rule applies to S-Evid (*Strong Text-based Evidence* vs *Weak Text-based Evidence*; *Strong Explanation* vs *Weak Explanation*).

Due to the small size of the data, we follow Tran et al. (2024a) and use 2-fold cross-validation. In each fold, half of the data (56 transcripts) is considered as training data and the remaining data (56 transcripts) is used for evaluation. We also make sure that transcripts of the same teacher are in the same fold to prevent data leakage.

4 Methods

4.1 ICL Example Retrieval for LLMs

We adopt the prompts from prior work (Tran et al., 2024a) for our LLM. We utilize the predictive approach, which is the approach that yields the best results in all 4 IQA dimensions (Predictive-llm). It is the BC-5turns-10s strategy described by Tran et al. (2024a), utilizing the LLM as a binary classifier by prompting it to determine whether an observation related to an IQA dimension is present in a single turn (yes or no) (see Appendix D).

For our ICL example retriever, we use LLM-R (Wang et al., 2024)[1]. It uses LLMs to rank the candidates based on the log-likelihood of the ground-truth output, then trains a cross-encoder as a reward model to mimic the preferences of LLMs, and finally distills that knowledge to a bi-encoder for efficient inference. For a given input (a 5-turn dialogue excerpt), we retrieve the top 10 examples from the training data and use them as few-shot examples in the LLM prompt. We use separate retrievers (LLM-R) for teachers' and students' turns. In other words, when predicting a teacher or student's turn, we will only try to retrieve examples from a pool consisting of examples from the same speaker role (student or teacher). For example, if we are predicting if the last turn (given its 4 previous turns) is T-Press, the retriever will only try to find examples (5-turn dialogue windows) by looking at ones that end with a teacher's turn.

Although LLM-R specializes in ICL example retrieval, it was trained on tasks different from classroom discussions (e.g., sentiment, reading comprehension, closed-domain QA). Besides using off-the-shelf LLM-R, we also fine-tune it on classroom discussions. However, because our dataset is small, finetuning an ICL example retriever on the training set is ineffective. We instead use another classroom discussion dataset, TalkMoves (Suresh et al., 2022), to finetune LLM-R.

The TalkMoves dataset contains K-12 math classroom transcripts, annotated for talk moves based on accountable talk theory and dialog acts. The dataset includes 567 transcripts, comprising 174,186 annotated teacher utterances, 59,874 annotated student utterances, and 1.8 million words (15,830 unique). All of the transcripts are annotated for 6 teacher talk moves (Keeping everyone together, Getting students to relate to another's ideas, Restating, Pressing for accuracy, Revoicing, and Pressing for reasoning) and 4 student talk moves (Relating to another student, Asking for more info, Making a claim, and Providing evidence or reasoning). For finetuning the retriever, we use the same binary prediction task as Predictive-llm. However, we perform multiple binary predictions (yes/no) for all possible talk moves in each turn and use the definitions of these talk moves from the dataset (Suresh et al., 2022). While these moves differ from ATM codes, they share similarities and reflect

[1] https://github.com/microsoft/LMOps/tree/main/llm_retriever

a theoretical approach closely related to the one behind ATM.

4.2 Re-ranking and Label Ratio Regulation

In this section, we propose a method that uses re-ranking and forces a specific label ratio in the example set to improve ICL performance for classroom discussion quality assessment.

Re-ranking. Re-ranking is a popular approach in retrieval tasks. The initial retrieval process is generally designed to be fast, often prioritizing speed over perfect accuracy. As a result, in ICL example retrieval, the first batch of examples retrieved can be broad, including both highly relevant and somewhat irrelevant information. Re-ranking addresses this by filtering and reordering these examples according to refined relevance scores, reducing noise and irrelevant information. In the first step, we re-rank the top-100 retrieved examples to get a set of examples ordered by their usefulness. We experiment with 2 re-ranking methods.

LLM as a re-ranker: We use a Llama3 model as the scorer. Specifically, for a given input and a retrieved example, we ask a yes/no question if the example can help answer the given question and use the probability of "yes" as the score.

BERT as a re-ranker: We train a BERT-based model as a cross-encoder reward model that gives higher scores to good ICL examples. We first create the necessary training data to train the BERT model. To do this, from our available training data (Section 3), for each instance (a turn), we retrieve the top-K using the LLM-R retriever (either trained or not trained). We then employ Llama3 to obtain the rankings. The ranking score is calculated as $\log p(y|x, x_i, y_i)$ where x is the given input, y is the gold answer, x_i and y_i are an in-context learning example retrieved and its label. For a training example (x, y), we first sample one positive example (x^+, y^+) from the top-ranked candidates and N_{neg} negative examples $(x_i^-, y_i^-)_{i=1}^{N_{neg}}$ from the bottom-ranked candidates. The reward model takes as input the concatenation of (x, y, x^+, y^+) and produces a score $s(x, y, x^+, y^+)$ for the positive example, and $s(x, y, x_i^-, y_i^-)$ for the negatives. The training objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{\text{reward}} = -\log \frac{e^{s(x,y,x^+,y^+)}}{e^{s(x,y,x^+,y^+)} + \sum_{i=1}^{N_{\text{neg}}} e^{s(x,y,x_i^-,y_i^-)}}$$

Label Ratio Regulation (LRR). Thinking that the lack of hard negative examples and the imbalance of positive/negative examples can be potential causes for the poor performance of the off-the-shelf retrieval setting, we want to ensure that this will not happen. To do so, we make sure the 10-example set follows a specific label ratio of positive, negative, and hard negative (if applicable) examples. For a fair comparison, we force the ratio to mimic the ratio from the fixed setting (defined in Tran et al. (2024a)). Although the ordering of few-shot examples is also a non-trivial factor (Ye et al., 2022), it is not what we focus on. Therefore, we fix the order of the chosen examples. For T-Press and T-Link, from top to bottom, we want 5 positive and then 5 negative examples. Similarly, for S-Link and S-Evid, we will see 4 positive, 4 easy negative, and 3 hard negative examples, respectively, from top to bottom of the example set. To do so, given the ranked list of examples, we pick from top to bottom until the predetermined label ratio is satisfied and skip examples that violate the label ratio if added. For instance, if we already have 5 positive examples for T-press, we will ignore the remaining positive examples in the list and only pick an example if it is a negative one as we go down the list.

5 Experimental Setup

To make it comparable to prior work without ICL example retrieval from Tran et al. (2024a), we use LLama3-8B (Grattafiori et al., 2024) as the LLM for classroom discussion assessment [2].

We use 3 baselines to test the effectiveness of the proposed method:

fixed: In this setting, we use a set of 10 fixed examples for each fold in the cross-validation. We follow prior work to pick those 10 examples for the LLM prompts (Tran et al., 2024a). This setting is also used as a baseline for comparison with approaches that utilize ICL example retrieval. One thing to note is that using this sampling method, we will have a fixed ratio of positive, easy negative, and hard negative examples in the 10-shot example set.

retrieved: In this setting, we use LLM-R (Wang et al., 2024) to find the top-10 examples from the training data. Then, we use those 10 examples for few-shot prompting.

mixed: In this setting, we construct a set of top-5 *retrieved* examples and 5 examples from the *fixed* set. For the 5 examples from the fixed set, we pri-

[2] https://huggingface.co/meta-llama/Llama-3.1-8B

oritize hard-negative examples first. In addition, prior work has shown that lacking hard negative examples is detrimental to the performance (Tran et al., 2024b). Therefore, we decide to select harder negative examples from the fixed set, as we cannot guarantee that they exist in the retrieved set. Specifically, for S-Link and S-Evid, we pick 3 hard negative examples, 1 positive example, and 1 non-hard negative example. For T-Press and T-Link, we pick 3 positive and 2 negative examples. Then, we choose the remaining 5 examples from the retrieved ones based on the descending order of cosine similarity between them and the input embedded by LLM-R.

To test the performance of the proposed method, we experimented with 2 re-ranking methods: using LLama3 as the *LLM re-ranker* and using a finetuned *BERT re-ranker*. To highlight the importance of each component (LRR and re-ranking), we report the performance from utilizing both re-ranking and label ratio regulation in combination and from using each component separately.

To compare the performances between non-finetuned and finetuned retrievers, we finetune a new LLM-R on another classroom discussion dataset (TalkMoves from Suresh et al., 2022) and repeat the experiments.

Quadratic Weighted Kappa (QWK) is used as the main evaluation metric. It is a common metric for quantifying inter-rater reliability that penalizes disagreements proportional to the degree of disagreement, which is vital in contexts where a greater distance between scores is meaningful.

6 Results and Discussion

RQ$_1$: Effectiveness of the proposed method. Table 1 shows the macro average over 2-fold cross-validation of QWK scores in various settings, including the 3 baselines and the proposed method for both non-finetuned and finetuned retrievers.

The standard ICL example retrieval is not effective. When using a non-finetuned LLM-R, we observe that relying solely on retrieved examples (row 2) is worse than the *fixed* baseline (row 1). This implies that using ICL retrieval is ineffective in this case, despite helping to improve performance in previous work on other domains (Wang et al., 2024; Zhang et al., 2023). On the other hand, the *mixed* settings (row 3), where we combine examples from the retriever with the fixed set, are the baselines that achieve the best performance in all IQA dimensions. This suggests that the retrieved examples are still useful to some extent.

Our proposed method with BERT as the re-ranker achieves the best performance in all 4 IQA dimensions (row 7) for both non-finetuned and finetuned retrievers. Although finetuning the LLM-R boosts the performance of the *retrieved* setting (row 2), the proposed method performs comparably for both non-finetuned and finetuned settings of the LLM-R retriever (row 7), suggesting that finetuning the retriever on a new domain, which is computationally expensive, is not necessary. Our hypothesis for this minimal gain is that the TalkMoves data consists of math discussions, which contain math-specific lexicons not present in English Language Art discussions from our dataset. Additionally, the TalkMoves dataset is skewed towards sixth-grade to eighth-grade students, while our data only has discussions from fourth-grade and fifth-grade students.

As a re-ranker, although LLama3 shows equal or better performance over the *retrieved* setting in T-Link and T-Press (row 5 vs 2), it is inferior to the *fixed* setting in S-Link and S-Evid (row 5 vs row 1). On the other hand, using BERT as a re-ranker with label ratio regulation achieved the best results in all dimensions. With this combination, we are now able to outperform the mixed setting despite using only retrieved examples. This implies that for this task, using an LLM such as Llama3 as a judge for re-ranking is not a reliable method in comparison with finetuning a PLM such as BERT.

The LRR is shown to be essential for improved performance as removing it leads to decreases in QWK (rows 6 and 8 compared to the previous rows). The drop in performance in S-Link and S-Evid is larger than the drop in T-Link and T-Press. The former 2 dimensions (S-Link and S-Evid) have hard negative examples based on the coding manual, which suggests that LRR is more important when hard negative examples are available for the target dimension. With only re-ranking, we can perform similarly or worse than the *retrieved* setting. For instance, using a Llama3 re-ranker without LRR is worse than vanilla retrieval (row 6 versus 2). On the other hand, with LRR, we consistently outperform the *retrieved* setting, with or without using a re-ranker (row 4, 5, 7 versus 2)[4]. Moreover, when the retriever is finetuned, if we have to pick

[3] Two-tailed t-test on 2-fold cross-validation.

[4] Except for S-Link with finetuned retriever.

ID	Setting	Non-finetuned Retriever				Finetuned Retriever			
		T-Link	T-Press	S-Link	S-Evid	T-Link	T-Press	S-Link	S-Evid
1	fixed	0.65	0.73	0.64	0.79	0.65	0.73	0.64	0.79
2	retrieved	0.62	0.71	0.62	0.75	0.66	0.73	0.66	0.80
3	mixed	*0.68*	*0.76*	*0.67*	*0.81*	*0.72*	*0.77*	*0.71*	*0.82*
4	LRR only	0.63	0.72	0.65	0.79	0.68	0.76	0.65	0.81
5	Llama3 + LRR	0.65	0.72	0.62	0.76	0.68	0.75	0.63	0.77
6	w/o LRR	0.61	0.70	0.56	0.68	0.65	0.72	0.60	0.70
7	BERT + LRR	**0.72**	**0.80**	**0.73**	**0.83**	**0.73**	**0.81**	**0.73**	**0.83**
8	w/o LRR	0.66	0.78	0.64	0.77	0.66	0.78	0.65	0.77

Table 1: Quadratic Weighted Kappa (QWK) scores of the two retrievers. For each IQA dimension (T-Link, T-Press, S-Link, S-Evid), italic numbers represent the best baseline results. Bold numbers highlight the best retriever results. All numbers are statistically significant compared to their counterparts in the *mixed* baseline ($p < 0.05$).[3]

IQA	Non-finetuned LLM-R		Finetuned LLM-R	
	Avg	% w/o hard negative	Avg	% w/o hard negative
S-Link	3 / 1.2 / 3.2 / 1.5	0 / 27.2 / 0 / 24.3	3 / 1.5 / 3.3 / 1.7	0 / 23.3 / 0 / 20.7
S-Evid	3 / 1.9 / 3.1 / 2.1	0 / 22.7 / 0 / 20.8	3 / 1.7 / 3.4 / 2.0	0 / 20.5 / 0 / 19.1

Table 2: Presence of hard negative examples in the fixed, retrieved, mixed setting and an approach utilizing BERT re-ranking without LRR. We report the average number of hard negative examples included in the 10 examples (Avg) and the percentage of test instances where the few-shot examples in the prompt do not have any hard negative example. In each cell, from left to right, the 4 numbers represent the statistics for *fixed*, *retrieved*, *mixed* settings, and from an approach utilizing BERT re-ranking without LRR.

only one component, using LRR is usually better than using a re-ranker (row 4 versus rows 6 and 8). This suggests that we should always enforce the label ratio in the example set.

RQ$_2$: Issues in the label ratio of retrieved examples from automatic ICL example retrieval.

The lack of hard negative examples and skew in the ratio of positive and negative examples can be potential causes for the low performance of example retrieval. Noticing that directly using the retrieved examples is not an effective way to improve the performance of LLM-based classroom discussion quality assessment, we hypothesize the potential causes and do analyses to test them. Compared to the *fixed* and *mixed* settings, one thing we could not control in the *retrieved* setting is the distribution of the examples. We can think of two causes for the poor performance using the *retrieved* setting: 1) the lack of hard negative examples and 2) the lack of positive examples.

Missing hard negative examples in the few-shot example set will have a negative influence (Tran et al., 2024a). Table 2 shows the presence of hard negative examples in the *fixed*, *retrieved*, *mixed* settings, and an approach using BERT ranking without LRR for S-Link and S-Evid (the only two dimensions that have hard negative examples according to the definitions in the coding manual). We can see that the *retrieved* setting has fewer hard negative examples on average compared to the *fixed* and *mixed* settings. We also only witness cases in which the example set has no hard negative examples in the retrieved setting. With only a BERT re-ranker, these numbers barely change as we only see small increases in the average number of hard negative examples and decreases in the number of cases without any hard negative example compared to the retrieved setting (4th number versus 2nd number) in each cell. This aligns with one of our previous observations from Section 6 that removing LRR results in bigger decreases in QWK for S-Link and S-Evid compared to the other two dimensions. This implies that re-ranking alone still does not guarantee the presence of hard negative examples in the set of 10 few-shot examples for prompting. However, with domain knowledge of

IQA	Non-finetuned LLM-R		Finetuned LLM-R	
	Avg	% without positive	Avg	% without positive
T-Link	5 / 3.7 / 4.7 / 4.2	0 / 6.8 / 0 / 1.2	5 / 3.3 / 3.5 / 3.5	0 / 5.3 / 0 / 2.3
T-Press	5 / 7.3 / 5.4 / 6.1	0 / 0.0 / 0 / 0.0	5 / 6.8 / 5.2 / 5.5	0 / 0.0 / 0 / 0.0
S-Link	4 / 3.2 / 3.8 / 3.5	0 / 6.2 / 0 / 0.0	4 / 3.8 / 4.3 / 4.1	0 / 0.0 / 0 / 0.0
S-Evid	4 / 5.9 / 5.1 / 5.5	0 / 2.3 / 0 / 0.0	4 / 5.6 / 4.9 / 4.3	0 / 3.1 / 0 / 1.2

Table 3: Presence of positive examples in the fixed, retrieved, mixed setting, and an approach with only BERT re-ranking (no LRR), with the same notations as Table 2.

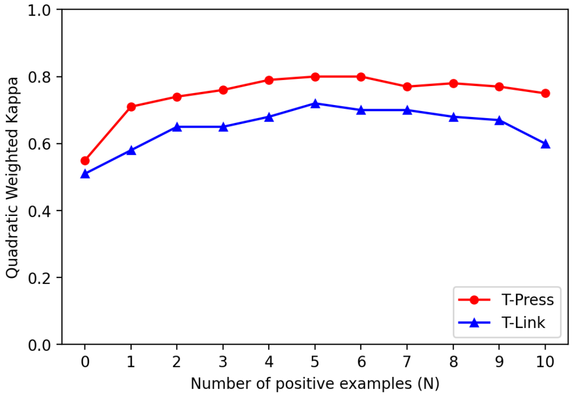

Figure 2: Results of T-Press and T-Link from different label ratios with N positive examples for BERT+LRR.

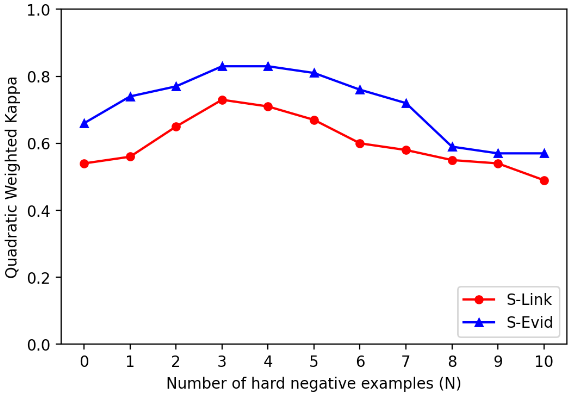

Figure 3: Results of S-Link and S-Evid from different label ratios with N hard negative examples for BERT+LRR.

hard negative examples (i.e., knowing that Weak Link is a hard negative example label for Strong Link, which represents S-Link), even with automated retrieval, we can ensure that hard negative examples are in the set.

Looking at the presence of positive examples (x_i, $y_i = y$) in Table 3, we see that the *retrieved* settings (2nd numbers of the cell) tend to include more positive examples in the 10-example set for T-Press and S-Evid while having fewer positive examples for T-Link and S-Link. Although they are rare, there are still cases in which we have no positive examples in the 10-example set for the retrieved setting, which never happens for the fixed and mixed settings. The BERT re-ranking (4th number) helps decrease the number of cases without any positive examples, and it makes the average number of positive examples retrieved in each IQA dimension closer to the *fixed* and *mixed* settings.

RQ3: The ratio of different types of examples does matter. We conduct additional experiments on our best model (BERT+LRR) by varying the ratios by changing the number of positive examples, negative examples, and hard negative examples (if they are available) to see if certain label ratios yield better results. Specifically, for T-Press and T-Link, because there is no hard negative example for these two dimensions, we record the performance with the N positive examples (N = 0 to 10) and 10-N negative examples. For S-Link and S=Evid, we pick N (N = 0 to 10) hard negative examples and equally split the remaining 10-N examples into positive examples and non-hard negative examples (if 10-N is odd, we have 1 more positive example).

Figure 2 shows the results for T-Press and T-Link with various positive/negative ratios. We observe that increasing the number of positive examples helps improve the performance to a certain point. Specifically, we see noticeable improvements until N=3, then it starts to slow down. However, after N=5 positive examples, the QWK begins to go down as N increases. This suggests that we should have a balance between positive and negative examples, which is reasonable, as having too many examples of a certain perspective (positive or negative) can create biases in that direction for the LLMs. Similarly, for S-Link and S-Evid, the performances are boosted until N=3 hard negative examples are selected and then they quickly drop (Figure 3). This time, the downward trend after N=4 is more noticeable compared to Figure 2, implying that having too many hard negative examples causes more harm than good. In other words, although hard negative examples are essential for high performance, we should keep room for positive and non-hard negative examples. Overall, these observations suggest that we should be cautious when selecting a label ratio for the example set, and having a balanced split between the possible labels (positive, non-hard negative, and hard negative) is a safer choice, as the dominance of a label tends to result in decreased performance.

7 Conclusions

This work proposes a simple but effective ICL example retrieval method that utilizes example reranking and label ratio regulation (LRR) to improve few-shot LLM performance in automated classroom discussion assessment. The results show that our fully automated example retrieval and selection approach outperforms the baselines in all tested IQA dimensions. Additionally, the performance of a non-finetuned example retriever (LLM-R) is comparable to that of a retriever finetuned on a similar domain dataset, suggesting that skipping the finetuning process of the retriever is viable. Further analyses show that the lack of positive and hard negative examples can be the reason for the poor performance of the traditional ICL example retrieval approaches. We also observe that even with re-ranking, both finetuned and non-finetuned retrievers fail to select enough hard negative examples to make the few-shot prompting effective, which highlights the importance of label ratio regulation in maintaining the presence of hard negative examples. Finally, we investigate the influence of the ratio of positive, non-hard negative, and hard negative examples, demonstrating that having an excessive number of any category hurts performance. We would like to explore the proposed method with a more advanced prompting method, such as Chain-of-thought (Wei et al., 2022), in the future.

Limitations

While our method uses Label Ratio Regulation (LRR) to maintain a specific ratio, it treats each potential example separately when selecting the top-10. This independent selection might not be ideal, as the chosen examples can interact with each other. Exploring combinatorial optimization and sequential decision-making techniques could lead to improvements.

Another limitation of our study is the lack of analysis on the influence of the size of the example pool on the performance. Because our training set is small, the relevance and diversity of the candidate examples can be a hindrance to the generic ICL example retrieval baseline. If we have a bigger example pool with more diversity, the LRR might become unnecessary.

The proposed approach involves several components, which people might find too complex and counterintuitive, potentially hindering the ease of LLM usage for downstream tasks. Additionally, the experiments are conducted on a single and small dataset. As a result, the generalizability of the findings is weakened.

Furthermore, we only use the smallest version of LLama3. Utilizing a bigger LLM (e.g., LLama3-70B) might yield higher results and undermine the effectiveness of ICL example retrieval.

Last but not least, our experiments are grounded on a binary classification task and the assumption that hard negative examples can be identified based on the definitions of the labels.

Acknowledgments

We thank the Learning Research and Development Center for their student financial support, as well as the Pitt PETAL group and the anonymous reviewers for their useful comments for improving this work.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Richard Correnti, Lindsay Clare Matsumura, Marguerite Walsh, Dena Zook-Howell, Donna DiPrima Bickel, and Baeksan Yu. 2021. Effects of online content-focused coaching on discussion quality and reading achievement: Building theory for how coaching develops teachers' adaptive expertise. *Reading Research Quarterly*, 56(3):519–558.

Laura M. Desimone and Katie Pak and. 2017. Instructional coaching as high-quality professional development. *Theory Into Practice*, 56(1):3–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured prompting: Scaling in-context learning to 1,000 examples. *Preprint*, arXiv:2212.06713.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lindsay Clare Matsumura, Helen E. Garnier, Sharon Cadman Slater, and Melissa D. Boston. 2008. Toward measuring instructional interactions “at-scale”. *Educational Assessment*, 13(4):267–300.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9721–9728.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024a. Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 500–510, Atlanta, Georgia, USA. International Educational Data Mining Society.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024b. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2):304–335.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519, Bengaluru, India. International Educational Data Mining Society.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*, 16(1):34–60.

Ian A. G. Wilkinson, P. Karen Murphy, and Sevda Binici. 2015. *Dialogue-Intensive Pedagogies for Promoting Reading Comprehension: What We Know, What We Need to Know*, pages 37–50. American Educational Research Association.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *Preprint*, arXiv:2310.07554.

A Data statistics

The transcripts have an average length of 3,421 tokens, with a median length of 3,537 tokens. The shortest transcript contains 1,986 tokens, while the longest reaches 6,393 tokens. Table 4 presents the statistics for the four key IQA dimensions highlighted in this work.

B Description of IQA Dimensions

Descriptions of the 4 focused IQA dimensions can be found in Table 5.

C Description of ATM codes

Descriptions of the relevant ATM codes can be found in Table 6.

D Prompts

Figure 4 contains the prompt used for the binary prediction of a target IQA adopted from Tran et al. (2024a), where {IQA description} is from the second column in Table 5.

IQA	**Distribution**	**Avg Score**	**Relevant *ATM* code**	**Hard negative ATM Code**
T-Link	[69, 23, 9, 11]	1.66	Recap or Synthesize S Ideas	n/a
T-Press	[8, 13, 11, 80]	3.46	Press	n/a
S-Link	[84, 7, 10, 11]	1.54	Strong Link	Weak Link
S-Evid	[38, 17, 9, 48]	2.60	Strong Text-based Evidence	Weak Text-based Evidence
			Strong Explanation	Weak Explanation

Table 4: Data distribution and mean (**Avg**) of 4 focused *IQA* rubrics for Teacher (T) and Student (S) with their relevant *ATM* codes and hard negative ATM code (if available). An *IQA* rubric's distribution is represented as the counts of each score (1 to 4 from left to right) (n=112 discussions).

IQA Dimension	IQA Dimension's Description
T-Link: Teacher links Student's contribution	*Did Teacher support Students in connecting ideas and positions to build coherence in the discussion about a text*? 4: 3+ times during the lesson, Teacher connects Students' contributions to each other and shows how ideas/ positions shared during the discussion relate to each other. 3: Twice… 2: Once… OR The Teacher links contributions to each other, but does not show how ideas/positions relate to each other (re-stating). 1: The Teacher does not make any effort to link or revoice contributions.
T-Press: Teacher presses Students	*Did Teacher press Students to support their contributions with evidence and/or reasoning*? 4: 3+ times, Teacher asks Students academically relevant Questions, which may include asking Students to provide evidence for their contributions, pressing Students for accuracy, or to explain their reasoning. 3: Twice… 2: Once…OR There are superficial, trivial, or formulaic efforts to ask Students to provide evidence for their contributions or to explain their reasoning. 1: There are no efforts to ask Students to provide evidence for their contributions or to ask Ss to explain their reasoning.
S-Link: Student links other's contributions	*Did Students' contributions link to and build on each other during the discussion about a text*? 4: 3+ times during the lesson, Students connect their contributions to each other and show how ideas/positions shared during the discussion relate to each other. 3: Twice… 2: Once… OR the Students link contributions to each other, but do not show how ideas/positions relate to each other (re-stating). 1: The Students do not make any effort to link or revoice contributions.
S-Evid(a): Student provides text-based evidence	*Did Students support their contributions with text-based evidence*? 4: 3+ times, Students provide specific, accurate, and appropriate evidence for their claims in the form of references to the text. 3: Twice… 2: Once…OR There are superficial or trivial efforts to provide evidence. 1: Students do not back up their claims.
S-Evid(b): Student provides explanation	*Did Students support their contributions with reasoning*? 4: 3+ times, Students offer extended and clear explanation of their thinking. 3: Twice… 2: Once…OR There are superficial or trivial efforts to provide explanation. 1: Students do not explain their thinking or reasoning.

Table 5: IQA dimensions and their definitions. For each IQA dimension, the italic line is {IQA description} used in the prompt in Appendix D.

Code	**Definition**	**Example**
Press	T asks the same S follow-up Questions (i.e., uptake/pushback Q's, request for text-based evidence and explanation).	Why did you say that? Where is the evidence? How else might Salva feel?
Recap or Synthesize S Ideas	T links multiple Ss' ideas or positions. T synthesizes multiple responses.	What I hear you saying is that the character has changed from the beginning of the book which is similar to Ana's idea that the character has matured.
Weak Link	Ss attempt to link contributions to each other, but do not show how ideas/positions relate to each other. The S might simply be revoicing or repeating another S's contribution.	"I disagree with Ana"... without explaining why or which aspect of Ana's statement S disagrees with.
Strong Link	Ss connect their contributions to each other and show how ideas/positions shared during the discussion relate to each other. Ss elaborate, challenge, or build on each other's ideas.	I'm not sure what Ana says is right because I don't see where in the text it says that. . .
Weak Text-Based Evidence	Ss provide inaccurate, incomplete, inappropriate, vague, or trivial evidence from/reference to text	Naya lived a hard life because in the chapters about her, we learn that she has to do a lot of things for her family.
Strong Text-Based Evidence	Ss provide accurate, appropriate, specific evidence from/reference to text that supports claim	On page 59, in the last paragraph it says, "I have talked to the others here,' uncle Jake said. 'We believe that the village of Loun-Ariik was attacked and probably burned your family.' Uncle paused and looked away."
Weak Explanation	S provides a brief or circular explanation that basically repeats or restates the response or relies on evidence to speak for itself.	I think that they didn't catch the fish because, , Tim hasn't caught any fish and Tim and Tom haven't caught any fish lately.
Strong Explanation	Ss provide an elaboration/justification of their answer or of the evidence they selected to support their answer.	Yeah, it is. The cause is, he didn't get the little girl's advice so, the effect of that is the calabash broke.

Table 6: ATM codes and their definitions

Prompt for binary prediction

Given a dialogue between a teacher and students in a classroom, in the last turn, {IQA description}?
# Example 1
Dialogue: {Example Excerpt 1(5-turn)}
Answer (yes or no): {Example Answer 1}
...
# Example 10
Dialogue: {Example Excerpt 10 (5-turn)}
Answer (yes or no): {Example Answer 10}
# Input
Dialogue: {Dialogue}
Answer (yes or no):

Figure 4: Prompt templates for binary prediction.