Multi-Dimensional Performance Analysis of Large Language Models for Classroom Discussion Assessment

Nhat Tran
University of Pittsburgh
Pittsburgh, USA
nlt26@pitt.edu

Benjamin Pierce
University of Pittsburgh
Pittsburgh, USA
bep51@pitt.edu

Diane Litman
University of Pittsburgh
Pittsburgh, USA
dlitman@pitt.edu

Richard Correnti
University of Pittsburgh
Pittsburgh, USA
rcorrent@pitt.edu

Lindsay Clare Matsumura
University of Pittsburgh
Pittsburgh, USA
lclare@pitt.edu

Automatic scoring of classroom discussion quality is becoming increasingly feasible with the help of new natural language processing advancements such as large language models (LLMs). Whether scores produced by LLMs can be used to make valid inferences about discussion quality at scale remains less clear. In this work, we examine how the assessment performance of two LLMs interacts with three factors that may affect performance: task formulation, context length, and few-shot examples. We also explore the computational efficiency and predictive consistency of the two LLMs. Our results suggest that the three aforementioned factors do affect the performance of the tested LLMs and there is a relation between consistency and performance. Using these results in conjunction with data from a randomized controlled trial, we then examine whether LLM-based assessment approaches that have a practical balance of predictive performance, computational efficiency, and consistency can be used to identify growth in discussion quality. We find that the best-performing LLM methods partially replicate results derived from human scores.

Keywords: classroom discussion, large language models, scoring, reliability, validity

1. INTRODUCTION

Automatic assessment of classroom discussion quality has been a rising topic among educational researchers. Decades of research have shown that classroom discussion quality is central to learning (Jacobs et al., 2022; Wilkinson et al., 2015). However, assessing classroom discussions in large numbers of classrooms has been expensive and infeasible to carry out at scale. Automated scoring of classroom discussion quality will aid researchers in generating large-scale

data sets to identify mechanisms for how discussions influence student thinking and reasoning. In addition, automated scores could be used in formative assessments to aid teachers in improving their discussion quality.

The major advantage of modern generative large language models (LLMs) compared to the previous generation of encoder-only pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) for automated scoring is that the former do not require training and only need proper prompting to do the task. We attempt to test the capability of LLMs in automatically providing scores for different dimensions of classroom discussion quality, based on the *Instructional Quality Assessment (IQA)*, an established measure that has shown high levels of reliability and construct validity in prior learning research (Correnti et al., 2021). Prior work has largely used LLMs by designing a single prompt with fixed inputs and evaluating zero-shot performance (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024; Wang et al., 2023) or by finetuning, which requires a large amount of training data and does not take advantage of the zero-shot or few-shot capability of LLMs (Kupor et al., 2023). We instead analyze 3 factors that can potentially affect the predictive performance of the LLMs, as well as examine their impact on computational efficiency and consistency in providing the same answer given the same input. Specifically, we test the capability of LLMs to score 4 IQA dimensions with 3 prompt-based factors in consideration: *task formulations*, *context length*, and the presence of *few-shot examples*.

To evaluate LLM performance, we prompt two LLMs to score a set of classroom discussion transcripts (N=112) that have previously been human-scored as part of a randomized controlled trial of teacher professional development (Correnti et al., 2021). We experiment with different prompting approaches and run each approach three times to examine consistency among LLM scores, and compare the resulting sets of majority-vote scores with human-generated scores to establish an inter-rater reliability score between human and LLM scores. Such agreement between human and automated scores has long been considered the "gold standard" in automated scoring research (Powers et al., 2002; Powers et al., 2015). However, research on LLMs as assessment tools also needs to determine whether LLM-generated scores can be used to make valid inferences about what is being scored. We adopt an argument-based validity framework (Kane, 2013; American Educational Research Association et al., 2014), according to which the usefulness of scores should be judged based on sources of evidence appropriate to the particular purpose to which the scores are meant to be put. For example, a set of discussion quality ratings might be useful for the purpose of *formative* assessment if they can be shown to be useful for helping teachers reflect meaningfully on their own practice, but less useful for *summative* assessments of those teachers if they cannot be shown to yield useful information for evaluating discussion quality.

This work thus extends our previous work (Tran et al., 2024) by broadening our performance evaluation of these LLMs to include a focus on the validity of LLM-generated discussion scores, without assuming that LLM-generated scores with good reliability will always be able to lead to the same conclusions as human-generated ratings. While our previous approach focused on LLM performance in terms of reliability factors, including validity analysis in our performance evaluation allows us to examine how useful LLM-generated scores may be for making research inferences. Since, as it turns out, scores with high reliability cannot always be put to valid uses, this extension of our performance analysis provides important added context.

Our contributions are threefold. First, we show how three factors (task formulation, context length and few-shot examples) can influence LLM performance and computational efficiency

in the IQA score prediction task. Second, we examine the consistency between the LLMs' outputs and find correlations between reliability and consistency in certain high-performance approaches. Third, since our classroom data were collected in the context of a randomized controlled trial, we examine the validity of the LLMs' output scores by making inferences about treatment effects using scores generated by human raters, a pre-trained BERT-based model, and the best LLM approaches we identify in this work. Comparing the inferences derived from each set of scores allows us to understand how the LLM-generated scores might be put to valid use. To support reproducibility, we also make our source code available at https://github.com/nhattlm95/LLM_for_Classroom_Discussion.

2. Related Work

Researchers have measured classroom discussion at different grain sizes and with different foci. Human coding has often focused on either teaching moves or student moves, with some measures occurring at the utterance or turn level, while others focus on different dimensions of instructional quality using more holistic measures. Consequently, automated coding has followed similar directions (Alic et al., 2022; Jensen et al., 2020; Demszky et al., 2021; Demszky and Liu, 2023; Nazaretsky et al., 2023; Suresh et al., 2021; Jacobs et al., 2022; Lugini et al., 2018; Lugini and Litman, 2018; Lugini and Litman, 2020; Xu et al., 2024). In Appendix A, we provide the coding constructs each paper cited in this paragraph automates, the automation method, and the subject matter and grade level of students in the dataset, as well as the evidence by which each paper evaluates the performance of its automation approach. *Our work benefits from human coding that contains both holistic assessment of feature scores and fine-grained coding at the turn level.*

Automated discussion scoring research has emphasized the potential usefulness of automated scores for assessing the extent to which teachers exhibit discussion facilitation skills and providing teachers feedback to help them develop those skills (Jensen et al., 2020; Demszky and Liu, 2023; Kupor et al., 2023; Nazaretsky et al., 2023). These are implicit claims to score *validity*, i.e., claims that the measurements produced by automated scoring methods can be defensibly used for those evaluative and developmental purposes (Kane, 2013). Discussion facilitation scores based on coding face a variety of validation issues, though validity is rarely discussed in the discourse coding field as such (Mercer, 2010; Hennessy et al., 2020). Zechner and Loukina (2020) apply a consensus notion of construct validity (American Educational Research Association et al., 2014) to the case of automated speech scoring, noting that such scores are considered valid if they are based on item features that are relevant, appear in human scoring rubrics, and are related to human raters' cognitive processes. While accepting the importance of construct validity, our work also emphasizes *consequential* validity, the practical availability and intelligibility of scores for use in local contexts (Moss, 2016).

Prior to generative LLMs, encoder-only transformer models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) have been widely used for automatically predicting classroom discourses. BERT has been shown to outperform a traditional open-vocabulary approach using n-grams and Random Forest classifiers in predicting teacher discourse when there is enough data to fine-tune the model (Jensen et al., 2021). The fine-tuned BERT also works the best in identifying conversational "uptake" in school classrooms (i.e., moments when the teacher revoices, elaborates on, or asks a follow-up question to a student contribution) (Demszky et al., 2021). In classifying "TalkMoves" for both teacher and students (e.g. "keeping everyone

together", "restating", "press for reasoning", "providing evidence", ...), BERT also shows decent performance and outperforms LSTM (Suresh et al., 2021). In scoring 4 dimensions of IQA (Junker et al., 2005) based on turn-level Analyzing Teaching Move (ATM) (Correnti et al., 2015) prediction, utilizing BERT in 2 settings (Hierarchical Classification and Sequence Labeling) shows promising performance in 2 dimensions ("Teacher presses Student" and "Student supports their claims") but presents low results in the other 2 dimensions ("Teacher connects Students" and "Students build on other's ideas") (Tran et al., 2023).

Recently, as generative LLMs such as GPT-4, Llama, and Mistral have surpassed encoder-only language models such as BERT in many NLP tasks, more work has started to explore the use of these generative LLMs (which we will refer to as simply LLMs from now on) in assessing classroom discussion quality. For predicting accountable talk moves in classroom discussions, a finetuned LLM was shown to consistently outperform RoBERTa in terms of precision (Kupor et al., 2023). Since finetuning a LLM requires expertise and a large amount of data, and is computationally expensive, others have focused on zero-shot methods which do not require training. For example, researchers have tested the zero-shot capabilities of ChatGPT in three tasks: scoring transcript segments based on classroom observation instruments, identifying highlights and missed opportunities for good instructional strategies, and providing actionable suggestions for eliciting more student reasoning (Wang and Demszky, 2023). The results indicated that ChatGPT is bad at scoring classroom transcripts by following Classroom Assessment Scoring System (CLASS) or Mathematical Quality of Instruction (MQI) instruments, and the generated responses were relevant to improving instruction, but they were often not novel or insightful (e.g., repeating what the teacher already does in the transcript). Work utilizing LLMs on more fine-grained levels (e.g., sentence-level, utterance-level) have been also conducted. A comparison between zero-shot ChatGPT and BERT in classifying student talk moves showed that, although ChatGPT could provide detailed and clear explanations for its predictions on student utterances, it was significantly worse than the smaller BERT model in 3 out of 4 student talk moves by a large margin (Wang et al., 2023). Another work explored how LLMs can be used to predict Instructional Support domain scores of CLASS (Whitehill and LoCasale-Crouch, 2024). Specifically, for each utterance, they zero-shot prompted Llama2 (Touvron et al., 2023) to infer whether or not the utterance exhibited one of the behavioral indicators associated with CLASS Instructional Support. This resulted in an 11-dimensional vector representation for each utterance which was then used to predict the final holistic score via linear regression.

These prior works have shown the potential of using LLMs in a zero-shot setting for classroom discussion quality assessment, but all suggested room for improvement. Moreover, in these prior works, standard zero-shot approaches with fixed prompts were used and evaluated on pre-segmented excerpts of the transcripts, without further analyses of other factors that can potentially affect LMMs' performance in the context of classroom discussions. *Our work experiments with 3 such factors (i.e., different prompting strategies, different input lengths, zero versus few-shot examples) that have been shown to influence LLM performance in other domains.* First, different ways of formulating a task in the prompt may yield different outcomes (Zhou et al., 2023; Karmaker Santu and Feng, 2023; Jiang et al., 2021). Our study uses multiple prompting strategies reflecting different formulations of the holistic assessment task (e.g., end-to-end or via talk moves). Second, LLMs can struggle in processing very long text input (Sun et al., 2021). Since our transcripts are often long, we experiment with different ways of reducing the LLM input size. Third, providing few-shot examples is known to be an effective way to increase LLM performance (Karmaker Santu and Feng, 2023; Liu et al., 2022; Brown

et al., 2020; Kojima et al., 2024). Since few-shot examples have not yet been utilized in previous classroom discussion LLM studies (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024), we propose a method for constructing such examples.

In addition to testing the influence of task formulation, context length and few-shot examples on predictive performance, we also *evaluate the 3 factors' influence on computational efficiency* (an important consideration for real-time formative assessment). Finally, although aggregating multiple LLMs' results for the same input (i.e., majority vote) has achieved higher performance in various NLP tasks (Wang et al., 2023; Pan et al., 2023), the consistency of the predicted results has not been examined in the context of classroom discussion. *We explore result consistency at both the transcript and the score level and examine relationships with predictive performance*.

3. Dataset

Our corpus is created (with institutional review approval) from videos of English Language Arts classes in a Texas district. The videos were recorded during the course of a randomized controlled trial of the effectiveness of an online instructional coaching program (Correnti et al., 2021). The videos were collected from 18 fourth grade and 13 fifth grade classrooms, whose teachers on average had 13 years of teaching experience. Eight teachers were assigned to the treatment group and 23 to the control group. The majority of the student population was considered low income (61%), with students identifying as: Latinx (73%), Caucasian (15%), African American (7%), multiracial (4%), and Asian or Pacific Islander (1%).

The videos were manually scored 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 at the sentence level using the *Analyzing Teaching Moves (ATM)* discourse measure (Correnti et al., 2021). The final corpus consists of **112** discussion transcripts (44 from treatment teachers and 68 from controls) that have been scored using both the IQA and the ATM (see Appendix B for the statistics of the scores). Thirty-two videos (29 percent) were double-scored indicating good to excellent reliability for holistic scores on the IQA (the Intraclass Correlation Coefficients (ICC) range from .89-.98) and moderate to good reliability for fine-grained talk moves on the ATM (ICC range from .57 to .85).

We have used the IQA holistic assessment scores for one analytic purpose: to demonstrate a treatment effect aligned with the theoretical foundations of our coaching. Meanwhile, we have used the ATM fine-grained coding of talk moves for a different analytic purpose: to make research inferences about the process of change in relation to our theory of action (Correnti et al., 2021). In developing automated scores for our BERT-based model we used the ATM fine-grained coding for training the model in order to predict the rubric score holistic assessment of four IQA dimensions. The detail in the fine-grained codes enabled us to score several IQA dimensions with reliability. We then used the holistic assessment scores of IQA dimensions to examine growth models for validity evidence for whether 1) we observe a treatment effect over time for a composite for overall discussion quality using human scores from 3 IQA dimensions, and 2) if so, whether the treatment effect over time is replicated with automated scores (Matsumura et al., 2024). The work here focuses on automated holistic assessment of classroom discussion using LLMs to understand if scores derived from generative LLMs also replicate findings from human scores and from prior BERT-based models.

The university's IRB approved all protocols such as for consent and data management (e.g., data collection, storage, and sharing policies). Privacy measures include anonymizing teacher

names in the transcripts used for analysis. Additionally, we only used open-source LLMs which do not expose our data to external sources.

IQA dimension scores. The complete list of IQA dimensions can be found in Appendix C. For this initial analysis, we focused on four of the 11 IQA dimensions. We chose these dimensions because of their relevance to dialogic teaching principles that emphasize collaborative knowledge-building and active participation in meaning-making processes. Two of the dimensions focus on teaching moves and two focus on student contributions. 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 Table 1.

Overall discussion quality score. While all of these dimensions are relevant to dialogic teaching quality, three of these dimensions (T-Link, S-Link, and S-Evid) can be considered aspirational in the sense that higher scores on these dimensions indicate more dialogic teaching. T-Press, on the other hand, is transitional, in that an increase over a very low score on this dimension represents an improvement but a very high score might indicate that a teacher is over-relying on the simple talk move of T-Pressing students for more. Conversely, a decrease in T-Press might indicate that a teacher is replacing simple T-Presses with more dialogic talk moves. Thus we hypothesized that when combined, the three aspirational dimensions (T-Link, S-Link and S-Evid) would provide a theoretically grounded composite estimate of overall discussion quality (ODQ). We further hypothesized, based on our instructional coaching intervention, that T-Press scores would decrease as ODQ increases.

4. Prompting Methods for Scoring

Given a full classroom discussion transcript, our IQA score prediction task is to predict a score between 1 and 4 for each of the 4 targeted IQA dimensions. Because there are 3 factors that can affect the performance of LLMs, we use the same format to name the approaches. Specifically, each approach is named as ***tf-cl-fs*** depending on the combination of the 3 factors: task formulation (*tf*), context length (*cl*), and few-shot examples (*fs*). Figure 1 shows the final models and the combinations that create them. Example prompts are in Appendix D. In this section, we describe the 3 factors and how we experimented with them in the task.

4.1. Task Formulation Factor

LLMs receive instructions about the problem and how to achieve the desired results through prompts. Previous work has shown that different instructions can lead to different results for the same task (Zhou et al., 2023; Karmaker Santu and Feng, 2023). Additionally, although it is possible to prompt the LLM to do multiple tasks (Wang and Demszky, 2023), our preliminary experiments show that the LLM sometimes fails to complete some or all of the tasks. Therefore, we decided to use prompts that only require the LLM to do one task. We experimented with the following 4 ways to formulate the task:

Direct score (DS). We prompt the LLM to predict an IQA score for the transcript by giving it the description of each score for that dimension (1-4) (Figure 2a). {IQA description} informs

Table 1: IQA dimensions and their definitions. For each IQA dimension (i.e., T-Link-S-Evid), the italic line is {IQA description} and the remaining text is {Scoring instruction} used in the prompts in Section 4.

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 T-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 Students 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.

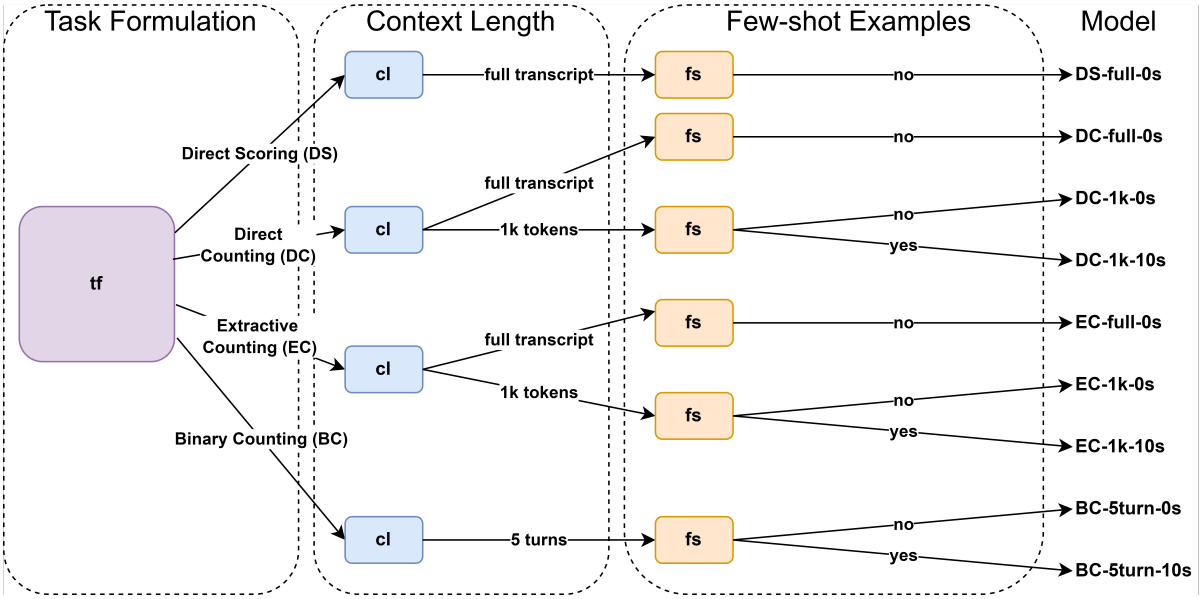

Figure 1: The experimented LLM approaches and how they are constructed based on the 3 factors: Task Formulation (tf), Context Length (cl) and Few-shot Examples (fs).

the LLM about the definition of the focused IQA score and {Scoring instruction} provides the criteria of each score from 1 to 4 for that IQA dimension. This is similar to end-to-end approaches that directly output the final score, either through transformer (Nazaretsky et al., 2023) or LLMs (Wang and Demszky, 2023; Kupor et al., 2023).

Direct counting (DC). For each IQA dimension, the description of each score from 1 to 4 is based on the count of relevant observations (i.e., a count of associated ATM codes at the turn level). Therefore, the {Scoring instruction} in DC can be formulated as a counting task. We ask the LLM to count how many times a certain observation that represents an IQA dimension appears in the transcript by giving the IQA description (Figure 2b). This can be treated as an alternative way to prompt the LLM with more direct and specific instructions (i.e., the LLM does not have to infer that the Scoring instruction is indeed a counting task).

Extractive counting (EC). We prompt the LLM to extract turns from the transcript that satisfy certain observations that contribute to an IQA dimension (Figure 2c). The final IQA score can be inferred by counting the number of turns found. This task formulation gives some explainability to the final score. Since a count higher or equal to 3 results in the maximum IQA score (4), we limit the number of extracted examples to 3 in the prompt.

Binary counting (BC). We use the LLM as a binary classifier by prompting it to predict if an observation that represents an IQA dimension appears in one turn (yes/no) (Figure 2d). Based on the performance in preliminary tests, we chose 4 previous turns for the dialogue history. Unlike the other 3 approaches which process the entire transcript in one go, this approach uses LLM on the turn level. We then add the binary counts of each IQA dimension to get the final counts and infer the IQA scores. This is similar to approaches identifying turn-level talk moves to predict holistic scores (Tran et al., 2023; Nazaretsky et al., 2023), except a LLM is the classifier instead of a transformer and there is no training/finetuning. This is also the most specific instruction as the output only has 2 labels (yes/no).

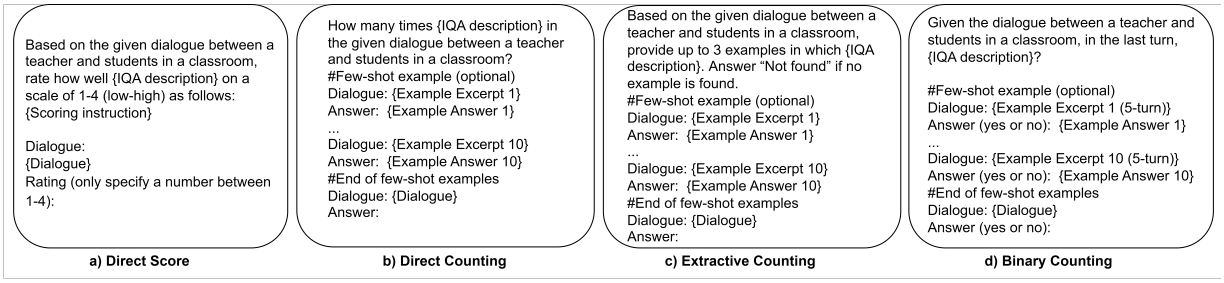

Figure 2: Prompts used in this work. Lines starting with # are comments and are not part of the prompts. {IQA descriptions} and {Scoring instruction} can be found in Table 1.

4.2. Context Length Factor

While previous work experimenting with LLMs on classroom discussion used short transcripts (e.g., several turns, 15-min passage) (Wang and Demszky, 2023; Nazaretsky et al., 2023), our transcripts are generally much longer (35 minutes on average). Specifically, 32 out of 112 transcripts have more than 4000 tokens, which exceed the token limit of many modern LLMs. Furthermore, although LLMs are claimed to be able to process long input text, their capabilities in dealing with long-range context are still questionable (Sun et al., 2021). Therefore, we test whether giving the LLM a shorter context such as an excerpt instead of the entire transcript ({Dialogue} in Figure 2) leads to a change in performance. For DC and EC, we split the transcripts into smaller excerpts of 1k tokens (best performance based on preliminary results) and aggregate the counts predicted by LLMs of each split to get the final counts of a transcript. For DS, we only use the full transcripts because we want to test the capability of LLMs to understand and follow the scoring instructions directly (i.e., no manual counting). We call these approaches **DC-1k** and **EC-1k**. BC approach is a turn-level prediction, thus we need to call the LLMs for every turn. Adding a large context (e.g., 1k tokens) will significantly increase the computational cost as it is part of the prompt. In addition, when predicting for only one turn, we do not need that many previous turns in the conversation to see the whole picture. Therefore, we decide to use 4 previous turns as the context for BC (**BC-5turns**).

4.3. Few-shot Examples Factor

Providing examples is a simple yet effective way to improve a LLM's performance (Brown et al., 2020; Karmaker Santu and Feng, 2023). For DS, since it is not possible to fit 10 full transcripts in the prompt, we only experiment with zero-shot (DS-full-0s). For approaches that have free spaces in the prompt, we try few-shot prompting by adding 10 more examples to the prompts. For BC, since each example is short (5 turns), we can freely provide any 10 5-turn excerpts with answers (yes/no) as few-shot examples for a selected IQA dimension. For DC-1k and EC-1k, we select 10 excerpts (700 tokens max) and create their gold answers. The gold answer is the count of relevant ATM codes for DC-1K and a list of turns containing relevant ATM codes for EC-1K in the excerpt. We end up with 3 approaches using 10-shot examples: **DC-1k-10s**, **EC-1k-10s** and **BC-5turn-10s**.

For consistency, we have a fixed set of 10 examples for each approach. To not expose test instances in these examples, we split the data into 2 segments A and B and for transcripts in one segment, we only draw examples from the other segment. In other words, for each

IQA dimension of DC-1k-10s, EC-1k-10s, and BC-5turn-10s, we create 2 10-example fixed sets (from segment A and B). When working on a transcript, only 1 of those 2 sets are used depending on the segment the transcript belongs to. We also make sure that every possible label is covered in the 10 examples: 0-3 for DC, 0-3 extracted turns for EC, and yes/no for BC.

To create those 10-example sets, instead of hand-picking the examples from the data, we use sampling. We use the word *sample* from now on to describe the process of randomly selecting a text unit (several consecutive turns in the conversation) from the dataset until a certain condition is satisfied. In BC-5turn-10s, for T-Link and T-Press, we first sample 5 positive (yes) and then sample 5 negative (no) few-shot examples (5-turn each). Previous work has shown that presenting hard-negative examples yields better prediction results (Robinson et al., 2021). In our case, for a certain ATM code x, there is another ATM code y that is more similar to x compared to other ATM codes. We consider y as a hard-negative example of x. Specifically, the hard-negative example ATM codes of Strong Link (S-Link), Strong Text-based Evidence (S-evid(a)), and Strong Explanation (S-evid(b)) are Weak Link, Weak Text-based Evidence, and Weak Explanation, respectively. We decide to sample 4 positive, 3 hard-negative and 3 easy-negative examples when predicting S-Link, S-Evid(a) and S-Evid(b) for BC-5turn-10s. For DC-1k-10s, we sample 2 text excerpts with the count of the IQA observation as k (0 to 3), respectively, creating 8 examples. Similarly, for EC-1k-10s, we sample 2 dialogue excerpts in which k (0 to 3) examples that satisfy the {IQA description} are extracted. The last two examples of DC-1k-10s and EC-1k-10s do not have any restrictions.

5. Experimental Setup

5.1. The BERT Baseline

We use a BERT-based model from a previous work (Tran et al., 2023) as our baseline. It was trained to predict sentence-level ATM codes and the final IQA scores were inferred based on the counts of predicted ATM codes through a linear layer. There are two configurations used.

For **Hierarchical Classification**, we perform a 2-step hierarchical classification at sentence level as follows: step 1, binary classification, is to classify *Other* versus *5 focal ATM Codes*; if the ATM code is not *Others*, step 2 is to perform another 5-way classification to identify the final label. The BERT-based classifiers for each step are trained separately. Because each *ATM* code except *Others* can be only from one speaker, either Teacher or Student, we train two classifiers for the 5-way classification of Step 2, one classifier used to predict teacher codes (*Recap or Synthesize S Ideas* and *Press*) and one classifier specialized in student codes (*Strong Link*, *Strong Text-Based Evidence* and *Strong Explanation*).

The other configuration is **Sequence Labeling**. This approach assigns a label (1 out of the 6 *ATM* codes) to each sentence in a conversation sequentially. Different from the hierarchical classification approach, which predicts the label of each sentence independently, this approach makes the label of a given sentence more dependent on the labels of nearby sentences. We use BERT-BiLSTM-CRF for this, which leverages BERT for its powerful performance for sentence representation and BiLSTM-CRF for its state-of-the-art performance in sequence labeling tasks (Huang et al., 2015; Panchendrarajan and Amaresan, 2018).

We follow the original work (Tran et al., 2023) and use down-sampling to mitigate imbalanced data and merge consecutive sentences with the same predicted ATM code as one ATM to make it more consistent with the annotation. Since the results show that each configura-

tion works better for the prediction of certain ATM codes (Tran et al., 2023), we use the best configuration for each IQA dimensions based on their related ATM code. Specifically, we use Hierarchical Classification for T-Link or Sequence Labeling for T-Press, S-Link and S-Evid.

5.2. Testing for Reliability

Commercial LLMs are costly and do not always guarantee data privacy, so we use open-source ones. To make a fair comparison with end-to-end scoring (DS), we want a LLM that can fit long classroom discussion transcripts (as 32 out of 112 transcripts have more than 4000 tokens). Also, we want to test more than one LLM to make the findings more generalizable. Among the open-source LLMs, Mistral (Jiang et al., 2023) and Vicuna (Zheng et al., 2023) have a token limit of at least 8000, which is enough to cover any of our transcripts. Specifically, we use Mistral-7B-Instruct-v0.1[1] and Vicuna-7b-v1.5-16K[2] from huggingface with default parameters. We do not train or fine-tune the LLMs and use them as is.

To test the influence of the 3 aforementioned factors on score reliability, we report the average Quadratic Weighted Kappa (QWK[3]) of the LLM approaches mentioned in Section 4. QWK is a standard metric for quantifying inter-rater reliability that penalizes disagreements proportional to the degree of disagreement, which is important in contexts such as this where a greater distance between scores is meaningful. For BC-5turns-10s, we also report the performances on S-Link and S-Evid without using hard-negative examples (i.e., 5 positives and 5 non-restricted negatives) to further test the effectiveness of having harder examples. Due to our small dataset, we use 5-fold cross-validation for our BERT baseline, even though this makes the baseline not directly comparable to the results of our zero-shot and few-shot approaches. For each prompt, we run the LLM 3 times and aggregate the final predictions. Since LLMs' outputs can be inconsistent, we use majority voting[4] as previous work has shown that this is a simple yet effective technique (Wang et al., 2023).

To compare the computational efficiency, we record the average inference time (i.e., time to produce the set of 4 IQA scores for 1 transcript). We do not include the training time of BERT and the time spent on prompt engineering for LLMs. All experiments were done on a computer with a single RTX 3090 Nvidia GPU.

To measure the per-transcript consistency of LLMs, for each transcript, we record the number of times 2 out of 3 runs (2/3) and all 3 runs (3/3) have the same predictions per IQA dimension. The frequency that none of the 3 runs have the same predictions can be self-inferred. We also report the per-score consistency to see if the LLMs are more/less consistent in certain scores.

5.3. Testing for Validity

As in prior work evaluating automated essay scoring (Correnti et al., 2021) we develop a validity argument for our automated discussion quality measures. We view this as an important, often over-looked research activity because the gold standard for assessing automated measures typically begins and ends with inter-rater reliability. Given our study investigates the use of automated discussion quality scores for research purposes, and more specifically, for testing the

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[2] https://huggingface.co/lmsys/vicuna-7b-v1.5-16k
[3] We use scikit-learn, a python package, to compute QWK
[4] We calculate the mean and round it to the closest integer if all 3 runs have different predictions

effects of professional development interventions on classroom discussion quality, we sought to collect evidence about what research inferences we could make with our automated scores. In order to do this, we relied on an argument approach to evaluating construct validity (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989). Utilizing this approach, we used theory and evidence from multiple sources to draw conclusions about the trustworthiness of information derived from our automated scores. One important tenet of this approach is related to the interpretation/use argument. That is, conclusions about validity need to be based on the interpretation and usefulness of scores for serving a particular purpose. In our case, we are seeking research inferences about whether automated scores replicate research inferences for effects of a coaching intervention on discussion quality using human scores. Although it is possible to conceive of different uses for our automated measures – e.g., as formative assessments that inform future professional learning – our validity argument, at this time, is limited only to replicating our prior research inferences because that is where we have data to generate evidence. Engaging in the validation process first involves articulating claims around what is important to know with respect to whether the inferences that can be made from automated scores are appropriate for research purposes, and then selecting evidence to test those claims.

Following this argument approach, we claim that automated scores used for the purpose of research should satisfy three main conditions related to: 1) meaning-making about the construct being evaluated; 2) agreement with human raters and 3) the inferences researchers can make when automated measures replace human measures. Specifically, in our case, automated discussion quality scores need to be constructed using theoretically important dimensions of discussion quality and should preferably include a focus on both teacher and student talk moves. Second, automated scores for different dimensions of discussion quality should show a reasonable level of agreement with human-generated scores (i.e., as discussed above they should demonstrate reasonable inter-rater agreement and also be consistently reproducible). Reliability is important for advancing transparency, that is, for helping ensure that automated scores capture features of quality that align with human judgment of the target domain (Kane et al., 1999). Third, automated scores of discussion quality need to be sensitive to intervention main effects, the bottom-line criteria used most often to determine whether an intervention 'worked.' Given our longitudinal data we explore whether automated scores are sensitive to differences between treated and control teachers in their growth in discussion quality over the same time period.

In relation to the meaning-making from our dimension scores, we focused on a subset of four IQA dimensions (Tran et al., 2023) for our automation work. As our original analyses (Correnti et al., 2021) relied on having human-scored all eleven IQA dimensions for analysis, we engaged in data reduction activities for both our human and automated dimension scores. To reduce the data, we first conducted factor analyses to develop a composite measure of discussion quality based on human scoring which were then replicated in a confirmatory factor analysis with the automated scores. Based on these analyses we determined the best factor consisted of three aspirational dimensions, one focused on teacher facilitation and the second and third focused on the quality of students' contributions. The teacher facilitation dimension *Teacher links students' contributions* assesses the extent to which a teacher shows how the ideas expressed by students in discussions relate to each other. The student focused dimensions *Students link to each other's contributions* and *Students support their contributions with evidence/explanation* assess the extent to which students build on other students' ideas and support and explain their ideas with evidence. All three of these dimensions are seen to be foundational for dialogic instruction in the

literature on discussion quality because each dimension supports collective knowledge building and students' active participation in meaning-making processes during text-based discussions.

Finally, to test our research inferences we examined IQA scores from the four LLM approaches that we found outperformed the BERT baseline model on reliability (Table 2). Since these measures were derived from transcripts collected during a randomized controlled trial we use these scores to test for a treatment effect on each IQA dimension as well as overall discussion quality (ODQ). The goal is to determine whether any of these approaches would lead us to identify a treatment effect in the same cases as human scores would. To test for an effect, we use mixed effects models (Raudenbush and Bryk, 2002) to produce growth estimates using both human- and LLM-generated scores. Since our data is longitudinal and we are interested in growth over time, we adopt a mixed effects repeated measures design, treating individual discussions as nested in time points and time points as nested in teachers. In our models we adjust for the number of turns at the discussion level as well as classroom features such as test scores and demographic features at the teacher level. Our unconditional two-level model for each outcome is:

$$\text{IQA}_{ti} = \pi_{0i} + \pi_{1i}(Time_{ti}) + e_{ti}$$
$$\pi_{0i} = \beta_{00} + \beta_{01}(CFC_i) + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}(CFC_i) + r_{1i}$$

IQA_{ti} is the IQA dimension being scored or ODQ, π_{0i} is the baseline IQA score for teacher i; Time_{ti} is the change in time since baseline; π_{1i} is the linear growth slope for teacher i; β_{00} is the average baseline IQA score across teachers; CFC_i is a dichotomous indicator for teacher i for participation in Online Content Focused Coaching (1 = treated; 0 = control); β_{01} is the difference in baseline IQA score for treated versus control teachers; β_{10} is the average linear growth slope for teachers; β_{11} is the difference in the growth slope of IQA score for treated versus control teachers; e_{ti} is the within-person residual; r_{0i} is the between-teacher variance estimate for the intercept at baseline, and r_{1i} is the between-teacher variance for the linear growth slope. Our primary interest is in the treatment effect (β_{11}) on each IQA dimension's (and ODQ's) growth over time.

In order to account for distributional features of the scores, we examined continuous models for our outcomes with normal distributions and Poisson models for our outcomes with skewed count distributions. Poisson models were most appropriate for T-Link, T-Press, and S-Link, and continuous models were most appropriate for S-Evid and ODQ.

6. Results and Discussion

Since our goal is to explore the appropriateness of using LLMs to rate discussion quality, we examine inter-rater reliability between LLMs and human raters, as well as internal reliability or consistency among multiple LLM runs. Table 2 (reliability) shows the Quadratic Weighted Kappa (QWK) between LLM and human scores for each approach along with their computational time. Figures 3 and 4 present the consistency results for each approach. We also consider whether, internal and inter-rater reliability aside, LLM-generated scores can be used to make the same research inferences as human-generated scores. Table 4 (validity) compares growth coefficients derived from human scores, baseline BERT scores, and scores from the most reliable LLM approaches for the 4 individual IQA dimensions as well as for ODQ (based on T-Link,

Table 2: Quadratic Weighted Kappa (QWK) and Inference Time (ITime) from Mistral and Vicuna. The best numbers are **bolded**. *Italic* numbers mean they are equal or better than the BERT baseline. Inference time is the average of 3 runs. IQA dimensions have been abbreviated (TL = T-Link, TP = T-Press, SL = S-Link, SE = S-Evid).

ID	Method	Mistral					Vicuna				
		IQA Dimension				ITime	IQA Dimension				ITime
		TL	TP	SL	SE		TL	TP	SL	SE	
1	BERT	0.59	0.71	0.56	0.72	82.0s	0.59	0.71	0.56	0.72	82.0s
2	DS-full-0s	0.42	0.50	0.43	0.49	10.7s	0.42	0.48	0.45	0.48	12.3s
3	DC-full-0s	0.44	0.54	0.45	0.55	9.6s	0.45	0.53	0.45	0.56	9.7s
4	DC-1k-0s	0.46	0.57	0.47	0.61	10.7s	0.47	0.57	0.49	0.60	12.5s
5	DC-1k-10s	0.46	0.56	0.48	0.61	12.3s	0.46	0.58	0.49	0.63	12.4s
6	EC-full-0s	0.43	0.58	0.45	0.60	17.1s	0.41	0.54	0.42	0.60	18.2s
7	EC-1k-0s	0.45	0.59	0.49	0.63	24.7s	0.45	0.57	0.47	0.64	26.8s
8	EC-1k-10s	*0.61*	*0.71*	*0.60*	*0.74*	27.3s	*0.59*	0.70	*0.56*	*0.72*	30.5s
9	BC-5turn-0s	0.49	0.62	0.50	0.65	223.4s	0.49	0.63	0.50	0.63	234.1s
10	BC-5turn-10s	***0.63***	***0.75***	***0.64***	***0.77***	232.5s	***0.62***	***0.73***	***0.60***	***0.74***	237.9s
11	*w/o* hard-negative	-	-	0.55	0.69	-	-		*0.56*	0.69	-

S-Link, and S-Evid). We present results for reliability[5] and validity separately, and discuss decisions that influenced the results.

6.1. Reliability

Task formulation is an important consideration as there were differences in performance on the IQA score assessment tasks. DS underperforms other approaches, including the baseline BERT model, with QWK scores of no more than 0.50 in all dimensions. This is consistent with previous work which showed poor correlations between the scores predicted by a LLM and human raters on classroom transcripts (Wang and Demszky, 2023). DC's variants (rows 3-5) outperform DS-full-0s, suggesting that the LLM cannot fully infer the relation between the counts of IQA observation and the final scores. EC-based approaches generally achieve higher QWK than DC's counterpart, except in some zero-shot instances (T-Link of EC-full-0s and EC-1k-0s for Mistral; T-Link, S-Link of EC-full-0s and EC-1k-0s for Vicuna). This implies that the LLM is generally better at extracting the IQA observations than counting them directly. The BC approaches obtain the highest performance, with BC-5turn-0s and BC-5turn-10s beating their counterparts (i.e., same few-shot settings) in all IQA dimensions, except for EC-1k-0s in S-Evid with Vicuna.

Context length also affects performance. With the same task formulation, reducing the context length to 1K always increases the QWK. BC-5turn-0s can be considered a zero-shot approach with a very short context length (5 turns) and it outperforms all other zero-shot approaches. These observations suggest that breaking a long transcript into smaller chunks of text is the recommended way when using LLMs for our task because it not only yields higher QWK but also enables usage of a wider variety of LLMs with lower token limits (e.g., LLama2 with a token limit of 4k).

[5]The reliability results were originally presented in EDM 2024 (Tran et al., 2024).

Few-shot examples do matter. The only two approaches that can outperform the baseline BERT model are both few-shot attempts (EC-1k-10s and BC-5turn-10s for both LLMs, except T-Press in EC-1k-10s with Vicuna). The biggest gain in terms of performance is found when the Binary Counting approach is provided with 10 additional examples since BC-5turn-10s yields at least 0.10 points of QWK improvement over BC-5turn-0s, making it the best approach in all 4 IQA dimensions. While few-shot demonstration boosts the performances of Extractive Counting and Binary Counting, it does not help Direct Counting since DC-1K-10s performs similarly to DC-1K-0s, and worse in T-Press with Mistral and T-Link with Vicuna. We hypothesize that few-shot examples only help if they enhance the reasoning capability of LLM through those examples. For EC, the provided answers increase performance because the examples help the LLM better identify similar turns for scoring the IQA. For BC, the direct guidance from examples (yes/no) provides patterns (positive/negative) that the LLMs can absorb and generalize. In the case of DC, even with the correct counts given, the LLMs still need an intermediate reasoning step to identify the relevant IQA observations. In other words, the LLMs have to infer the characteristics of IQA observations from the counts - a task they struggle with. For DC, although the main task relies on counting, the bottleneck is likely from the capability of identifying related IQA observations, which the few-shot examples do not directly inject. The last two rows (10 and 11) also show BC-5turn-10s benefited from hard-negative examples, suggesting that having examples that are harder to distinguish from the focused labels when possible boosts classification performance of LLMs.

Computational efficiency. The BERT approach runs slower than most of the LLM-based approaches (except BC approaches) because it processes on the sentence level. EC-based approaches run slower than DC-based approaches as the former require generating more tokens (generate a turn versus a single number). BC approaches have superior performance in QWK compared to their counterparts but require excessive inference time. The best approach, BC-5turn-10s, needs around 8 times the amount of time to process a transcript on average compared to the second best approach, EC-1k-10s (232.5 seconds versus 27.3 seconds). Although running slower, EC-based and BC-based approaches can be more useful if we want to go beyond summative to formative assessment for coaching or feedback as they present examples to justify the decision. Therefore, if we want a balance between performance and inference time, EC-1k-10s is our recommended approach.

Figure 3 shows the **transcript-level consistency** across 3 runs for each approach. Each colored block represents the percentage (out of 112 transcripts) of different agreement rates. Specifically, a green block (3/3) with a number x in an IQA dimension S (x-axis) shows that in x% of 112 transcripts, all 3 runs have the same predicted score (i.e., between 1 and 4) in that IQA dimension. Similarly, a blue block (2/3) shows the percentage in which exactly 2 out of 3 runs have the same predictions for that dimension, and a red block (0/3) shows the percentage in which 3 runs predicted different scores for that IQA dimension. Although there are discrepancies among Mistral and Vicuna in different levels of agreement (2/3 and 3/3), most of the time, when majority voting is applied (i.e., at least 2 out of 3 agree on the final prediction; summation of green and blue blocks), they are within 5% of each other. The results also indicate that reaching total agreement (3/3) is hard for LLMs since red blocks are the smallest (the highest number is less than 37%). DS-full-0s is not only the worst approach performance-wise but also is very inconsistent as it has the lowest numbers overall (top 3 lowest agreement rates according to majority voting in all dimensions). On the other hand, the two approaches with the highest QWK, EC-1k-10s and BC-5turn-10s, obtain better consistency compared to the rest,

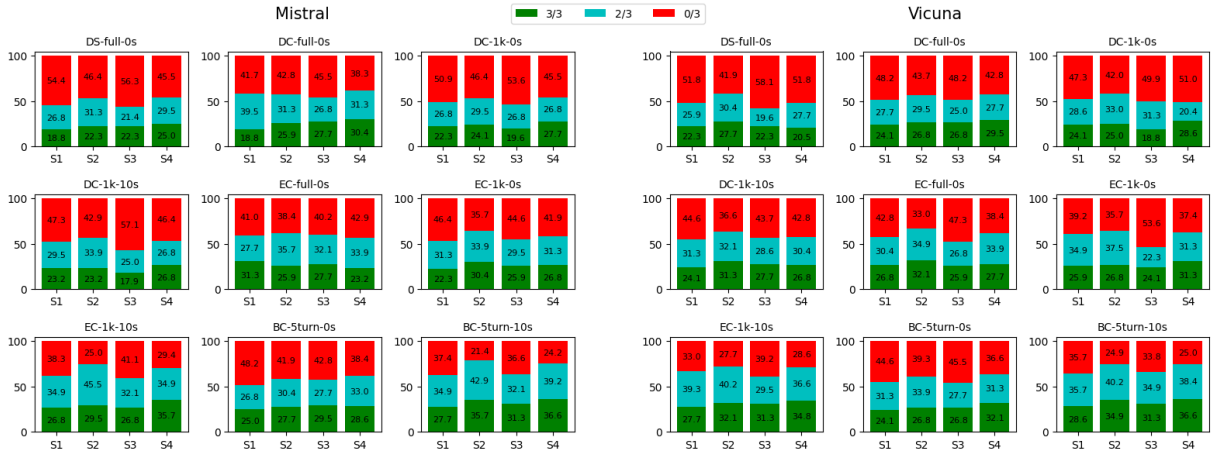

Figure 3: Per-transcript statistics of the agreement from 3 runs for each approach. Each number is a percentage of the number of transcripts (out of 112). For each IQA dimension (S1: T-Link, S2: T-Press, S3: S-Link, S4: S-Evid), 2/3 means exactly 2 out of 3 runs have the same predictions, 3/3 means all 3 runs have the same predictions, and 0/3 means no agreement between 3 runs.

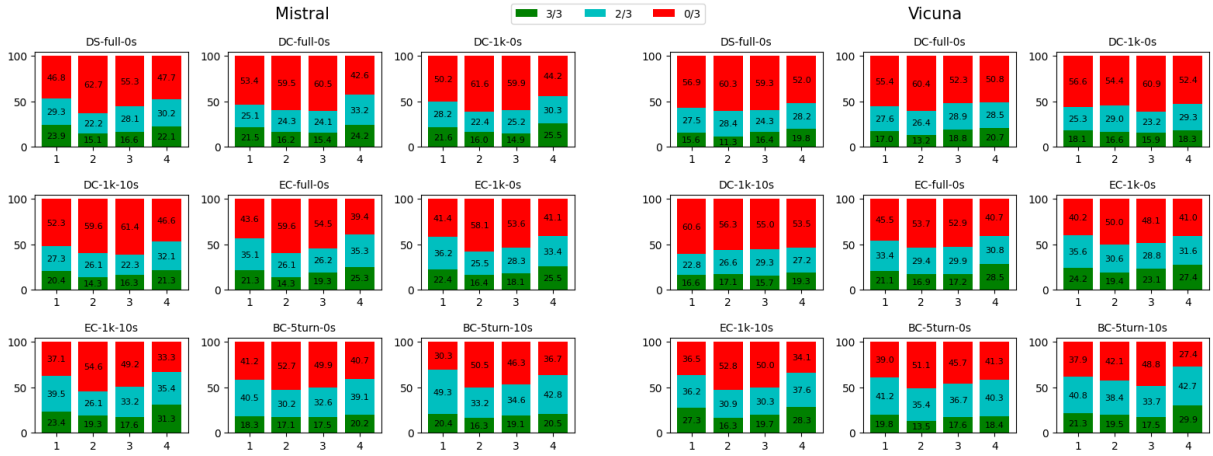

Figure 4: Per-score statistics of the agreement from 3 runs for each approach. For each combination of a score s (1-4) and an agreement rate (0/3, 2/3 or 3/3), each number is calculated as x/y where y is the number of times in any IQA dimension (T-Link, T-Press, S-Link, S-Evid), the final prediction is s and x is the number of times the agreement rate is satisfied among those y occurrences. For each score s, 2/3 means 2 out of 3 runs have s as predictions, 3/3 means all 3 runs have s as predictions.

especially in T-Press and S-Evid. Furthermore, similar to the QWK's result, T-Press and S-Evid tend to have higher consistency than T-Link and S-Link, suggesting that it is harder for LLMs to make consistent predictions on the latter dimensions. In general, these observations imply a relationship between performance and consistency of LLMs when the performance gaps are big, but when comparing approaches that are closer in performance,we see that an approach marginally better in QWK can have lower consistency (e.g., S-Link of EC-1k-0s versus EC-full-0s).

Figure 4 reports the **consistency across different scores**. This time, we group the items

Table 3: Agreement rates for the score of 4 when we treat the exact counts as the prediction instead of rounding them down to 3 (score of 4) when the number of occurrences exceeds 3.

ID	Method	Mistral		Vicuna	
		2/3	3/3	2/3	3/3
1	DC-full-0s	25.4	20.2	26.3	16.7
2	DC-1k-10s	28.4	18.3	28.1	17.7
3	DC-1k-10s	30.1	19.6	25.1	19.0
4	EC-full-0s	31.4	22.1	27.7	21.9
5	EC-1k-0s	29.3	20.2	28.5	23.6
6	EC-1k-10s	31.1	25.8	32.4	23.9
7	BC-5turn-0s	34.6	17.4	33.7	16.4
8	BC-5turn-10s	34.9	17.5	36.4	25.4

based on the final predicted scores. Specifically, we group all instances in which the final prediction are s (x-axis) regardless of the IQA dimension. For example, for a score $s = 3$, we collect all instances in all IQA dimensions where the final predictions are 3 for a total of y instances. Then, in that group, the green block (3/3) with a number p indicates that in p% of the y instances, all 3 runs have the same predicted score (i.e, 3 in this case). Similarly, a blue block (2/3) shows the percentage in which exactly 2 out of the 3 runs predict 3, and a red block (0/3) demonstrates the percentage in which 3 runs have different predictions. Overall, it is harder to reach a full agreement (3/3) for scores of 2 and 3 compared to 1 and 4 as all numbers in 3/3 (green blocks) for scores of 2 and 3 are lower than 20% (except EC-1k-0s of Vicuna for score 3). BC-5turn-10s has the highest percentages in majority voting in general (sum of 2/3 and 3/3), and its consistency for scores of 2 and 3 is lower than for scores of 1 and 4. This suggests that the LLMs are more consistent when predicting the extreme scores (1 and 4). We hypothesize that because a score of 4 is correct whenever there are at least 3 occurrences of certain IQA observations, even if the LLM misses some occurrences, it can still predict 4 as the final answer if the total number of occurrences is large; or it can overcount but the final prediction is still 4 due to the rounding down. Table 3 supports this assumption because when we use the exact counts instead of limiting it to 3, we see a decrease in consistency for both 2/3 and 3/3 compared to Figure 4. It implies that the LLMs are not very consistent for the score of 4 despite the high agreement rate from Figure 4. We leave further analyses to identify the problems of inconsistency for future work.

6.2. VALIDITY

Because we are concerned with real-world application, we explore whether the LLM-generated scores can be used to make valid inferences about the effectiveness of a coaching intervention on discussion quality. Table 4 shows coefficients representing the effect on IQA score growth of the instructional coaching intervention from the RCT in which our data were collected. We report growth coefficients for each IQA dimension and ODQ when calculated using the four best-performing LLM approaches based on reliability (rows 8 and 10 in Table 2), as well as when using human- and BERT-generated scores for comparison. The focus for our analysis is β_{11} - the estimate of the treatment effect and its magnitude. A statistically significant coefficient for β_{11} allows us to infer a treatment effect of the instructional coaching intervention.

Table 4: Growth coefficients (β_{11} in our multilevel model) and p-values for scores produced by human raters, BERT, and the four LLM prompting methods with reliability higher than BERT. * = statistically significant at $p < 0.05$, ** = significant at $p < 0.01$, *** = significant at $p < 0.001$.

IQA Dimension	Human		BERT		Mistral-BC-5turn-10s	
	Coef	Std. error	Coef	Std. error	Coef	Std. error
T-Link	0.36	0.28	0.29	0.20	0.10	0.12
T-Press	–0.46	0.32	–0.46	0.31	–0.23	0.33
S-Link	0.69	0.36	0.36	0.25	0.64	0.07
S-Evid	2.30*	0.82	1.77*	0.70	2.04*	0.79
ODQ	1.29**	0.40	1.21***	0.32	0.85*	0.32
IQA Dimension	Mistral-EC-1k-10s		Vicuna-BC-5turn-10s		Vicuna-EC-1k-10s	
	Coef	Std. error	Coef	Std. error	Coef	Std. error
T-Link	–0.20	0.30	0.03	0.35	0.13	0.31
T-Press	–0.15	0.32	–0.09	0.37	–0.05	0.34
S-Link	0.45	0.30	0.38	0.35	0.29	0.37
S-Evid	1.55	0.76	1.64*	0.79	2.03*	0.76
ODQ	0.61	0.39	0.13	0.37	0.71	0.44

When calculated using human scores, the coefficients indicating the effect of the treatment over time are positive for ODQ as well as for its three component IQA dimensions (T-Link, S-Link, and S-Evid), all aspirational in nature, and negative for T-Press, the transitional dimension. These coefficients are statistically significant in the case of S-Evid and ODQ, suggesting a treatment effect on student use of evidence/explanation and overall discussion quality. BERT scores allow us to draw the same conclusions: while the coefficients are not quite as large, they are positive for the aspirational dimensions and negative for the one transitional dimension. In addition, BERT-based scores show statistical significance in the same two cases as the human scores. Also, because the standard deviations for the BERT-based outcomes are smaller (see Table 5), the magnitude of the effect size of the treatment is similar even though the coefficients are smaller. Thus we would be led to make the same inferences about the treatment effect regardless of whether we used human- or BERT-generated scores. (For more discussion of the relationship between inferences made with human- and BERT-generated scores, see (Matsumura et al., 2024)).

Similarly, growth coefficients calculated using scores from Mistral-BC-5turn-10s allow us to make the same inference as coefficients calculated using scores from human raters, with positive coefficients in every case but T-Press and statistically significant coefficients for S-Evid and ODQ (though the latter are significant only at the $p < 0.05$ threshold, while coefficients using human-generated scores are significant at the $p < 0.01$ threshold).

The pattern of results for the four different sets of LLM prediction models are more mixed, despite their reliability metrics being better than BERT. Only Mistral-BC-5turn-10s allows for the identification of a treatment effect in the same cases as human ratings and BERT-based scores - i.e., a statistically significant effect on growth for both S-Evid and ODQ. The other three prediction models replicated human findings only partially. The two Vicuna approaches identified a treatment effect for S-Evid, but the treatment effect for overall discussion quality did not meet the traditional cutoff for statistical significance (p = .08; often considered to be

Table 5: Standard deviations of scores for human and LLM scoring methods.

Scoring Method	T-Link	T-Press	S-Link	S-Evid	ODQ
Human	0.99	0.96	1.01	1.34	0.88
BERT	0.52	0.70	0.52	1.13	0.62
Mistral-BC-5turn-10s	0.69	0.76	0.69	1.20	0.60
Mistral-EC-1k-10s	0.68	0.75	0.81	1.22	0.63
Vicuna-BC-5turn-10s	0.70	0.83	0.79	1.21	0.52
Vicuna-EC-1k-10s	0.89	0.71	0.82	1.22	0.71

marginally significant). The S-Evid score from Mistral-EC-1k-10s was marginally significant ($p = 0.054$), but it failed to identify a treatment effect for ODQ. Despite the falloff in statistical significance, it is important to note that these mixed results did not contradict human scores: the coefficients are comparable and are generally in the same direction as human scores. It is possible that as future work expands the size of our dataset and refines the reliability of our LLM scores, we will be able to recover more of the research inferences made with human scores.

It is notable that the performance for the LLM approaches was best for S-Evid. One difference between this dimension and the other three IQA dimensions was the distribution - the mean was more central and the distribution was not skewed (see Table 7). It is also a little surprising that only one of the four LLM approaches identified a treatment effect for overall discussion quality. Given that S-Evid was one of the three components of that measure, it is likely the inability to detect an effect on overall discussion quality is because the LLM approaches were further away from human estimates for T-Link and S-Link (both rarely occurring codes). LLM approaches only closely replicated S-Evid, whereas BERT-based estimates were similar across all IQA dimensions. The distribution of scores might also play a part here. LLM-based scores had a much narrower distribution than human scores in every case but S-Evid (see Table 5). This could help explain why the LLM approaches struggle with S-Link, T-Link, and T-Press (and, by extension, ODQ), where human scores were clustered mostly at extreme scores (see Table 7 in Appendix A). Finally, we observed variability in the pattern of results for LLM approaches which suggests the three performance-relevant factors we experimented with may affect the validity of LLM-generated scores.

7. Conclusion

We experimented with three factors affecting the performance of two LLMs in the automated assessment of classroom discussion quality. Our results show that the two LLMs have similar performance on reliability measures and that task formulation is the most important factor that impacts the performance and inference time. A shorter context length generally yields higher results but requires more computational time. Furthermore, providing few-shot examples is a very effective technique to boost the performance of an LLM if it can utilize the cues from those examples. Further optimization on how to sample few-shot examples (Liu et al., 2022; Zhao et al., 2021) is left for future work.

Additionally, a brief count representing different levels of agreement across 3 runs shows that approaches that are noticeably better in prediction results are more likely to have higher consistency, but further analyses are still needed due to the overall low consistency. We would

also like to examine in future research how our findings generalize to other classroom discussion corpora and assessment schemes.

Since discussion quality scores are typically generated to allow researchers and practitioners to make inferences about discussion quality, we tested the most reliable prompt approaches to see whether the scores they generated would allow us to make the same inferences as benchmark scores (in this case, human scores and scores from a pre-trained BERT model). In most cases we cannot make the same inferences with LLM-generated scores as with human- or BERT-generated scores, even though these LLM-generated scores had better agreement with human scores (measured by QWK) than BERT-generated scores did. This finding indicates, first, that human-LLM agreement may not be sufficient as a test of performance, depending on what use LLM scores are meant for, since we see that different sets of scores with good reliability can still lead to different conclusions about the activity being scored. One takeaway of this finding is to impress on researchers the importance of examining whether and how automated measures reproduce research inferences. Second, it indicates the importance of prompting methods. While no LLM matched human scores as closely as BERT did, one (Mistral) still showed good reliability and allowed us to make the same conclusions regarding the direction, statistical significance, and size of the treatment effect, but only when using the BC-5turn-10s method and not EC-1k-10s. Thus Mistral's usefulness for identifying a treatment effect depends in part on the three performance-affecting factors we experimented with. Third, while our evidence is limited at this point and it is premature to draw conclusions comparing and contrasting BERT-based and generative LLM approaches, perhaps as others engage in work to compare and contrast automation methods some patterns might emerge in the future.

Furthermore, these findings do not indicate that the other LLM approaches with good reliability cannot be put to good use. For example, since the best-performing LLM approaches often agree with human raters in identifying high-quality teaching moves and student use of evidence and explanation, they can potentially be useful in formative assessment contexts. For these purposes, finding a balance between the inference time and performance of LLMs is crucial, as it might not be worth sacrificing too much inference time for small performance gains. As LLMs proliferate and the affordances they provide to education research become more relevant, it is important not merely to rely on flagship models but to understand what factors are most important for prompting LLMs to produce valid assessments.

8. Limitations and Future Work

Due to our budget, we did not experiment with stronger open-source LLMs such as Llama3 or commercial LLMs such as GPT-4, which are more powerful and have a higher token limit. Additionally, although several other IQA dimensions could be tested using the same approach, we only worked on 4 of them. However, our approach can be seamlessly transitioned to new LLMs and other counting-based IQA dimensions as it is a prompt-based method.

Furthermore, human labor and example selection techniques can provide better examples instead of choosing few-shot examples by random sampling from the data as we did. As our results demonstrated the importance of having few-shot examples, we plan to experiment with in-context learning example retrievers such as LLM Retriever (Wang et al., 2024) to test whether an off-the-shelf retriever can find better few-shot examples that are dependent on the input context compared to our fixed sample-based example set.

Fine-tuning the LLMs, which has not been explored in this study, is a potential way to increase the performance further. Although increasing computation requirements is one of our concerns, the main bottleneck that discourages us from using fine-tuning is the lack of data. Even with a parameter-efficient fine-tuning method such as LoRA (Hu et al., 2022), our dataset is too small to utilize it. A future direction is to use data augmentation for fine-tuning. The sources can be other classroom discussion datasets such as TalkMoves (Suresh et al., 2019) or DiscussionTracker (Olshefski et al., 2020) or synthetic classroom discussion data generated by LLMs (e.g., GPT-4), in which the latter has gained a lot of attention in other fields lately due to the rise of generative LLMs (Hämäläinen et al., 2023; Li et al., 2023). As we update our prompts and measures, we will also update the findings regarding inter-rater reliability and reproducibility of research inferences. Our hope is that these analyses will provide further insights about which innovations for prompt engineering seem to be most important to aid researchers in measuring discussion quality.

Since the experiments were conducted using a specific dataset (English Language Arts classes in a Texas district) and specific student demographics, a potential algorithmic bias might be present (Baker and Hawn, 2022).

Furthermore, we have resisted the desire to over-generalize our findings, restricting our validity argument to the ability to generate valid research-based inferences with some of our automated measures. While our validity argument has focused on the use of these automated methods in education research, the potential remains for rich opportunities for their use for formative assessment purposes as well. Indeed, other work in the field (e.g., Jacobs et al. 2022; Jensen et al. 2020; Demszky and Liu 2023) has focused on the usefulness of providing automated feedback directly to teachers. Similarly, LLM-generated instructional quality scores of the kind developed in this paper could also be used to provide teachers and instructional coaches with information useful for reflecting on teacher practice.

9. Acknowledgments

We thank the Learning Research and Development Center for the grant "Using ChatGPT to Analyze Classroom Discussions" and the Learning Engineering Tools Competition.

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., and Jurafsky, D. 2022. Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, Eds. Association for Computational Linguistics, Seattle, Washington, 224–233.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. Standards for educational & psychological testing.

Baker, R. S. and Hawn, A. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education 32,* 4 (Dec), 1052–1092.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C.,

CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. T. Hadsell, M. F. Balcan, and H. Lin, Eds. NIPS '20, vol. 33. Curran Associates Inc., Red Hook, NY, USA, 1877–1901.

CORRENTI, R., MATSUMURA, L. C., WALSH, M., ZOOK-HOWELL, D., BICKEL, D. D., AND YU, B. 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.

CORRENTI, R., STEIN, M. K., SMITH, M. S., SCHERRER, J., MCKEOWN, M. G., GREENO, J. G., AND ASHLEY, K. 2015. *Improving Teaching at Scale: Design for the Scientific Measurement and Learning of Discourse Practice*. American Educational Research Association, 315–332.

DEMSZKY, D. AND LIU, J. 2023. M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*. L@S '23. Association for Computing Machinery, New York, NY, USA, 59—-69.

DEMSZKY, D., LIU, J., MANCENIDO, Z., COHEN, J., HILL, H., JURAFSKY, D., AND HASHIMOTO, T. 2021. Measuring conversational uptake: A case study on student-teacher interactions. 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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, Online, 1638–1653.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

HÄMÄLÄINEN, P., TAVAST, M., AND KUNNARI, A. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. N. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, Eds. CHI '23. Association for Computing Machinery, New York, NY, USA, 1–19.

HENNESSY, S., HOWE, C., MERCER, N., AND VRIKKI, M. 2020. Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture, and Social Interaction 25*, 100404.

HU, E. J., SHEN, Y., WALLIS, P., ALLEN-ZHU, Z., LI, Y., WANG, S., WANG, L., AND CHEN, W. 2022. LoRA: Low-rank adaptation of large language models.

HUANG, Z., XU, W., AND YU, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR abs/1508.01991*.

JACOBS, J., SCORNAVACCO, K., HARTY, C., SURESH, A., LAI, V., AND SUMNER, T. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education 112*, 103631.

JENSEN, E., DALE, M., DONNELLY, P. J., STONE, C., KELLY, S., GODLEY, A., AND D'MELLO, S. K. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery, New York, NY, USA, 1–13.

JENSEN, E., L. PUGH, S., AND K. D'MELLO, S. 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. LAK21. Association for Computing Machinery, New York, NY, USA, 302–312.

JIANG, A. Q., SABLAYROLLES, A., MENSCH, A., BAMFORD, C., CHAPLOT, D. S., CASAS, D. D. L., BRESSAND, F., LENGYEL, G., LAMPLE, G., SAULNIER, L., ET AL. 2023. Mistral 7b. *CoRR abs/2310.06825*.

JIANG, Z., ARAKI, J., DING, H., AND NEUBIG, G. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics 9*, 962–977.

JUNKER, B. W., WEISBERG, Y., MATSUMURA, L. C., CROSSON, A., WOLF, M., LEVISON, A., AND RESNICK, L. 2005. *Overview of the instructional quality assessment*. Regents of the University of California Oakland, CA.

KANE, M., CROOKS, T., AND COHEN, A. 1999. Validating measures of performance. *Educational measurement: Issues and practice 18,* 2 (Sum.), 5–17.

KANE, M. T. 2013. Validating the interpretation and uses of test scores. *Journal of Educational Measurement 50,* 1, 1–73.

KARMAKER SANTU, S. K. AND FENG, D. 2023. TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, Singapore, 14197–14203.

KOJIMA, T., GU, S. S., REID, M., MATSUO, Y., AND IWASAWA, Y. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrav, K. Cho, and A. Oh, Eds. NIPS '22. Curran Associates Inc., Red Hook, NY, USA, 22199 – 22213.

KUPOR, A., MORGAN, C., AND DEMSZKY, D. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint*.

LI, Z., ZHU, H., LU, Z., AND YIN, M. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, Singapore, 10443–10461.

LIU, J., SHEN, D., ZHANG, Y., DOLAN, B., CARIN, L., AND CHEN, W. 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*, E. Agirre, M. Apidianaki, and I. Vulić, Eds. Association for Computational Linguistics, Dublin, Ireland and Online, 100–114.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*.

LUGINI, L. AND LITMAN, D. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, N. Slonim and R. Aharonov, Eds. Association for Computational Linguistics, Brussels, Belgium, 57–67.

LUGINI, L. AND LITMAN, D. 2020. Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1475–1480.

LUGINI, L., LITMAN, D., GODLEY, A., AND OLSHEFSKI, C. 2018. Annotating student talk in text-based classroom discussions. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP*

for Building Educational Applications, J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, and H. Yannakoudakis, Eds. Association for Computational Linguistics, New Orleans, Louisiana, 110–116.

MATSUMURA, L. C., CORRENTI, R., LITMAN, D., PIERCE, B., AND TRAN, N. 2024. Automated measures of classroom discussion quality for research inferences. Under review.

MATSUMURA, L. C., GARNIER, H. E., SLATER, S. C., AND BOSTON, M. D. 2008. Toward measuring instructional interactions "at-scale". *Educational Assessment 13,* 4, 267–300.

MERCER, N. 2010. The analysis of classroom talk: Methods and methodologies. *British Journal of Educational Psychology 80,* 1, 1–14.

MESSICK, S. 1989. Meaning and values in test validation: The science and ethics of assessment. *Educational Research 18,* 2 (Mar.), 5–11.

MOSS, P. A. 2016. Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy, and Practice 23,* 2, 236–251.

NAZARETSKY, T., MIKESKA, J. N., AND BEIGMAN KLEBANOV, B. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. LAK2023. Association for Computing Machinery, New York, NY, USA, 122–132.

OLSHEFSKI, C., LUGINI, L., SINGH, R., LITMAN, D., AND GODLEY, A. 2020. The discussion tracker corpus of collaborative argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, Marseille, France, 1033–1043.

PAN, L., WU, X., LU, X., LUU, A. T., WANG, W. Y., KAN, M.-Y., AND NAKOV, P. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Toronto, Canada, 6981–7004.

PANCHENDRARAJAN, R. AND AMARESAN, A. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang, and Y. Yao, Eds. Association for Computational Linguistics, Hong Kong.

POWERS, D. E., BURSTEIN, J. C., CHODOROW, M. S., FOWLES, M. E., AND KUSICH, K. 2002. Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research 26,* 4, 407–426.

POWERS, D. E., BURSTEIN, J. C., CHODOROW, M. S., FOWLES, M. E., AND KUSICH, K. 2015. Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education 28,* 2, 130–142.

RAUDENBUSH, S. W. AND BRYK, A. S. 2002. *Hierarchical linear models: Applications and data analysis methods*. Sage.

ROBINSON, J. D., CHUANG, C., SRA, S., AND JEGELKA, S. 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, Austria.

SUN, S., KRISHNA, K., MATTARELLA-MICKE, A., AND IYYER, M. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 807–822.

SURESH, A., JACOBS, J., LAI, V., TAN, C., WARD, W., MARTIN, J. H., AND SUMNER, T. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. In *In the Proceedings of the Spring AAAI 2021 Symposium on Artificial Intelligence for K-12 Education*.

SURESH, A., SUMNER, T., JACOBS, J., FOLAND, B., AND WARD, W. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. *Proceedings of the AAAI Conference on Artificial Intelligence 33,* 01 (Jul.), 9721–9728.

TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S., BIKEL, D., BLECHER, L., FERRER, C. C., CHEN, M., CUCURULL, G., ESIOBU, D., FERNANDES, J., FU, J., FU, W., FULLER, B., GAO, C., GOSWAMI, V., GOYAL, N., HARTSHORN, A., HOSSEINI, S., HOU, R., INAN, H., KARDAS, M., KERKEZ, V., KHABSA, M., KLOUMANN, I., KORENEV, A., KOURA, P. S., LACHAUX, M.-A., LAVRIL, T., LEE, J., LISKOVICH, D., LU, Y., MAO, Y., MARTINET, X., MIHAYLOV, T., MISHRA, P., MOLYBOG, I., NIE, Y., POULTON, A., REIZENSTEIN, J., RUNGTA, R., SALADI, K., SCHELTEN, A., SILVA, R., SMITH, E. M., SUBRAMANIAN, R., TAN, X. E., TANG, B., TAYLOR, R., WILLIAMS, A., KUAN, J. X., XU, P., YAN, Z., ZAROV, I., ZHANG, Y., FAN, A., KAMBADUR, M., NARANG, S., RODRIGUEZ, A., STOJNIC, R., EDUNOV, S., AND SCIALOM, T. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv abs/2307.09288*.

TRAN, N., PIERCE, B., LITMAN, D., CORRENTI, R., AND MATSUMURA, L. C. 2023. Utilizing natural language processing for automated assessment of classroom discussion. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O. C. Santos, Eds. Springer Nature Switzerland, Cham, 490–496.

TRAN, N., PIERCE, B., LITMAN, D., CORRENTI, R., AND MATSUMURA, L. C. 2024. Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, B. Paassen and C. D. Epp, Eds. International Educational Data Mining Society, Atlanta, Georgia, USA, 500–510.

WANG, D., SHAN, D., ZHENG, Y., GUO, K., CHEN, G., AND LU, Y. 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*, M. Feng, T. Käser, and P. Talukdar, Eds. International Educational Data Mining Society, Bengaluru, India, 515–519.

WANG, L., YANG, N., AND WEI, F. 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)*, Y. Graham and M. Purver, Eds. Association for Computational Linguistics, St. Julian's, Malta, 1752–1767.

WANG, R. AND DEMSZKY, D. 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)*, E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, Eds. Association for Computational Linguistics, Toronto, Canada, 626–667.

WANG, X., WEI, J., SCHUURMANS, D., LE, Q. V., CHI, E. H., NARANG, S., CHOWDHERY, A., AND ZHOU, D. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, Kigali, Rwanda.

WHITEHILL, J. AND LOCASALE-CROUCH, J. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining 16,* 1 (Jun.), 34–60.

WILKINSON, I. A. G., MURPHY, P. K., AND BINICI, S. 2015. *Dialogue-Intensive Pedagogies for Promoting Reading Comprehension: What We Know, What We Need to Know*. American Educational Research Association, 37–50.

XU, P., LIU, J., JONES, N., COHEN, J., AND AI, W. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Association for Computational Linguistics, Mexico City, Mexico, 4375–4389.

ZECHNER, K. AND LOUKINA, A. 2020. Automated scoring of extended spontaneous speech. In *Handbook of Automated Scoring: Theory into Practice*, D. Yan, A. Rapp, and P. Foltz, Eds. Chapman and Hall/CRC, 365–382.

ZHAO, Z., WALLACE, E., FENG, S., KLEIN, D., AND SINGH, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds. Proceedings of Machine Learning Research, vol. 139. PMLR, 12697–12706.

ZHENG, L., CHIANG, W.-L., SHENG, Y., ZHUANG, S., WU, Z., ZHUANG, Y., LIN, Z., LI, Z., LI, D., XING, E., ZHANG, H., GONZALEZ, J. E., AND STOICA, I. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Vol. 36. Curran Associates, Inc., 46595–46623.

ZHOU, Y., MURESANU, A. I., HAN, Z., PASTER, K., PITIS, S., CHAN, H., AND BA, J. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, Kigali, Rwanda.

APPENDICES

A. RELATED WORK

Table 6 shows the coding constructs automated in existing work, automated method, subject matter and grade level of students in the data, and evidence used to evaluate performance.

B. DATASET STATISTICS

The average length of a transcript is 3421 tokens and the median is 3537 tokens. The shortest transcript consists of 1986 tokens and the longest one consists of 6393 tokens. Table 7 shows the statistics of the 4 focused IQA dimensions in our dataset.

C. OTHER IQA DIMENSIONS

We briefly list other IQA dimenions that were not studied in this work in Table 8.

D. EXAMPLE PROMPTS

Figures 5, 6, 7 and 8 show example prompts for Direct Score, Direct Counting, Extractive Counting and Binary Counting, respectively. The last lines of the prompts are incomplete to let

Table 6: Related work automating discussion quality constructs

Article	Constructs automated	Automation method	Subject matter	Grade level	Reliability/validity evidence
(Alic et al., 2022)	Funneling/focusing teacher questions	fine-tuned RoBERTa, unsupervised approaches	math	4th-5th grade	comparison with human ratings, qualitative analysis of results
(Demszky and Liu, 2023)	Counting instances of teacher uptake of student contributions	fine-tuned BERT	Data 4	high school	
(Demszky et al., 2021)	Counting instances of teacher **uptake** of student contributions	Fine-tuned BERT	Math discussions	4th-5th grade	Comparison of human and automated coding, linear correlation between automated uptake measure and student satisfaction/instructional quality
(Jacobs et al., 2022)	Teacher talk move labels based on Accountable Talk: Keeping everyone together, Getting students to relate, Restating, Revoicing, Press for accuracy, Press for reasoning	BI-LSTM	Not specified	Elementary, middle & high school	Qualitative analyses of teacher perception of automated feedback
(Jensen et al., 2020)	Applied labels to teacher talk: Instructional Talk, Questions, Authentic Questions, Elaborated Evaluation, High Cognitive Level, Uptake, Goal Specificity, ELA Terms	Data 3	ELA	1st & 3rd grade	correlations between automated estimates and human coding
(Kupor et al., 2023)	Labled transcripts with five Accountable Talk codes: Adding on, Connecting, Eliciting, Probing and Revoicing students' ideas	fine-tuned RoBERTa and GPT-3	online computer science classes	not specified	comparison to human coding, relevance/usefulness of feedback, correlation with student outcomes (attendance, students' section ratings and assignment completions)
(Lugini and Litman, 2020)	Labeling of human-segmented argument units with Claim, Evidence, Warrant	Two neural network models and pre-trained BERT	ELA	high school	Comparison to human coding
(Nazaretsky et al., 2023)	Teacher attention to student ideas	BERT	not specified	Simulated elementary students	Score in attention to student ideas rubric
(Suresh et al., 2021)	Sentence contains/does not contain a teacher talk move, plus one of six teacher talk move labels based on Accountable Talk: Keeping everyone together, Getting students to relate, Restating, Revoicing, Press for accuracy, Press for reasoning	BI-LSTM	Not specified	Not specified	Comparison to human coding
(Xu et al., 2024)	Labeling of teacher talk with five codes for teacher metacognitive modeling: Objective, Unpacking, Self-Instruction, Self-Regulation, Ending, as well as codes from Mathematical Quality Instruction rubric	multiple fine-tuned BERT variants and Llama-2-7b	not specified (SimSE dataset), math (NCTE dataset)	not specified (SimSE dataset), 4th-5th (NCTE dataset)	comparison to human coding

Table 7: Data distribution and mean (**Avg**) of 4 focused *IQA* rubrics for Teacher (T) and Student (S) with their relevant *ATM* codes. An *IQA* rubric's distribution is represented as the counts of each score (1 to 4 from left to right) (n=112 discussions).

IQA Rubric			**Relevant *ATM* code**
Short Description	**Distribution**	**Avg Score**	**Code Label**
T-Link: *T* connects *S*s	[69, 23, 9, 11]	1.66	Recap or Synthesize S Ideas
T-Press: *T* presses S	[8, 13, 11, 80]	3.46	Press
S-Link: S builds on other's ideas	[84, 7, 10, 11]	1.54	Strong Link
S-Evid: S supports their claims	[38, 17, 9, 48]	2.60	Strong Text-based Evidence
			Strong Explanation

Table 8: Other IQA dimensions that have not been studied in this work and their definitions.

IQA Dimension	IQA Dimension's Description
Participation in Learning Community	What percentage of Ss participated in the discussion about a text?
Wait Time	Did Teacher provide individual Students with adequate time in the class discussion to fully express their thoughts?
Rigor of Text	How rigorous were the text(s) used as the basis for the discussion? Did they contain sufficient 'grist' to support an academically challenging discussion?
Rigor of Class Discussion	Thinking about the text discussion as a whole and the questions Teacher asked Students, were Students supported to analyze and interpret a text (e.g., consider the underlying meaning or literary characteristics of a text, etc.)?
Segmenting Text	Does Teacher stop during the reading of the text to ask questions or clarify ideas?
Guidance Toward Constructing the Gist	Does Teacher ask open-ended questions and facilitate discussion to guide Students to construct the gist of the text (i.e., a coherent representation of the text)?
Developing Community	Does T help create a learning community within the classroom?

the LLMs complete the text (i.e., provide the answer). These prompts were iteratively developed through a trial-and-error process. Specifically, we modify the prompts while ensuring that the task formulation is intact (DS, DC, EC, EC) until we have a satisfying performance on the training data.

Based on the given dialogue between a teacher and students in a classroom, rate how well did Teacher support Students in connecting ideas and positions to build coherence in the discussion about a text on a scale of 1-4 (low-high) as follows:
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 during the lesson, Teacher connects Students' contributions to each other and shows how ideas/ positions shared during the discussion relate to each other.
2: Once during the lesson, Teacher connects Students' contributions to each other and shows how ideas/ positions shared during the discussion relate to each other 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.

#Dialogue
Teacher: Okay. So today we are going to read chapter one from The Many Troubles of Andy Russell. Just by hearing the title and looking at the front cover, what genre do you think it's going to be? Jordan?
Student: Someone that's getting in trouble or always making messes.
Teacher: Yes, but the genre. What genre do you think it's going to be?
...
Rating (only specify a number between 1-4):

Figure 5: An example prompt of Direct Score (DS) for (T-Link) Teacher links Student's contribution.

How many times did Teacher press Students to support their contributions with evidence and/or reasoning in the given dialogue between a teacher and students in a classroom?
#Dialogue
Teacher: What is he?
Student: It's fantasy. Nope!
Teacher: It is, it is.
Student: He's magic. Isn't it science fiction?
...
Answer: 1
#Dialogue
...
#Dialogue
Teacher: I hadn't thought of that! That's a new thought for me. Return to the carpet, please. Your time is up. I had never thought of that. Did anyone else think like Lily did? That they were thinking that she's trying to make them kinder people.
Student: I was thinking they're harassing the town.
Teacher: They are.
Student: I was thinking basically the same as Lily. I was thinking that Beatrix was trying to trick the two giants so they won't eat Greta. That was my guess.
...
Answer: 0
#Dialogue
Teacher: Okay. So today we are going to read chapter one from The Many Troubles of Andy Russell. Just by hearing the title and looking at the front cover, what genre do you think it's going to be? Jordan?
Student: Someone that's getting in trouble or always making messes.
Teacher: Yes, but the genre. What genre do you think it's going to be?
...
Answer:

Figure 6: An example prompt of Direct Counting (DC) for (T-Press) Teacher presses Student.

Based on the given dialogue between a teacher and students in a classroom, provide up to 3 examples in which Students' contributions link to and build on each other during the discussion about a text. Answer "Not found" if no example is found.
#Dialogue
Student: He says because he doesn't want it, ... I think that's what he meant to say.
Teacher: Okay. So who can tell me what Anthony just said? Damien.
Student: That not to, if they're telling you to throw the ball not hard, which means like Anthony said, not to throw it hard but also throw it light that since he doesn't have his mitt he can at least catch it. So that's why he wants them to throw the ball soft.
...
Answer:
Student: That not to, if they're telling you to throw the ball not hard, which means like Anthony said, not to throw it hard but also throw it light that since he doesn't have his mitt he can at least catch it. So that's why he wants them to throw the ball soft.
...
#Dialogue
...

#Dialogue
Teacher: Ah, I like what she said about getting more attention. Connor
Student: I think that Sco is trying to what Joshua and Essence just said, they're trying to tick them off like what Chloe does all the time, every day.
...
Answer:
Student: I think that Sco is trying to what Joshua and Essence just said, they're trying to tick them off like what Chloe does all the time, every day.

#Dialogue
Teacher 40: Okay. So today we are going to read chapter one from The Many Troubles of Andy Russell. Just by hearing the title and looking at the front cover, what genre do you think it's going to be? Jordan?
Student: Someone that's getting in trouble or always making messes.
Teacher 40: Yes, but the genre. What genre do you think it's going to be?
...
Answer:

Figure 7: An example prompt of Extractive Counting (EC) for (S-Link) Student links other's contribution

Given the dialogue between a teacher and students in a classroom, in the last turn, did Students support their contributions with reasoning?
#Dialogue
Student: Because he said if I don't fall out and so I think he's going to purposely fall out for some reason.
Teacher: Nicole.
Student: I think he's going to try to get them up. Get one of the boys up there.
Teacher: Okay, let's see. Okay, when it says up here, I found a wonderful seat up here in line 48. Can you find where that is? Sco said loudly. What do you think that tells us? What do you think Connor?
Student: I think that he is, that he's trying to make them jealous because boys likes to do that a lot and it's either that or like Elena said, he wants to distract them so they won't catch the ball and they might get interested in climbing the tree.
Answer (yes or no): yes
#Dialogue
...

#Dialogue
Student: They're being annoyed and because Sco wanted them to stop the game ...
Teacher: Ah, I like what she said about getting more attention. Connor.
Student: I think that Sco is trying to what Joshua and Essence just said, ...
Teacher: Okay. They went over and sat cross-legged in the shade of the tree. Sco looked down ... What's happening now, Nico?
Student: That's that Monk is getting really mad that he's like saying what they're doing already.
...
Answer (yes or no): no
#Dialogue
Student: Someone that's getting in trouble or always making messes.
Teacher: Yes, but the genre. What genre do you think it's going to be?
Student: Fiction.
Teacher: Fiction. Do you all agree?
Student: Yes.
Answer (yes or no):

Figure 8: An example prompt of Binary Counting (BC) for (S-Evid(b)) Student provides explanation