Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o

Yuya Asano*[1], Beata Beigman Klebanov[2], and Jamie N. Mikeska[2]

[1]Intelligent Systems Program, University of Pittsburgh, USA
[2]ETS Research Institute, USA
`yua17@pitt.edu`, `{bbeigmanklebanov, jmikeska}@ets.org`

Abstract

Engaging students in a coherent classroom discussion is one aspect of high-quality instruction and is an important skill that requires practice to acquire. With the goal of providing teachers with formative feedback on their classroom discussions, we investigate automated means for evaluating teachers' ability to lead coherent discussions in simulated classrooms. While prior work has shown the effectiveness of large language models (LLMs) in assessing the coherence of relatively short texts, it has also found that LLMs struggle when assessing instructional quality. We evaluate the generalizability of task formulation strategies for assessing the coherence of classroom discussions across different subject domains using GPT-4o and discuss how these formulations address the previously reported challenges—the overestimation of instructional quality and the inability to extract relevant parts of discussions. Finally, we report lack of generalizability across domains and the misalignment with humans in the use of evidence from discussions as remaining challenges.

1 Introduction

High-quality STEM instruction is well-organized and structured to provide opportunities for students to engage in productive scientific sensemaking, build their conceptual understanding, and link science ideas within and across lessons (Chen and Li, 2010; Roth et al., 2011). In fact, effective organization and structure are key features attended to in observational protocols for assessing teachers' practice, including the Framework for K-12 Science Education (National Research Council, 2012), Danielson's Framework for Teaching (Danielson, 2013), and the Classroom Assessment Scoring System protocol (Pianta, 2008). One specific high-leverage teaching practice that requires effective structuring is the facilitation of *coherent* content-focused discussions, as teachers need to ensure that students understand how the ideas that are discussed relate to and build upon one another and ensure that the work the students are doing supports progress towards addressing the discussion's learning goal (Carpenter et al., 2020; Stein et al., 2008).

Facilitating such discussions is a difficult skill to learn (Hanuscin et al., 2016; Plummer and Tanis Ozcelik, 2015; Ramsey, 1993). To help teachers develop these skills, it is important to provide them with ample practice opportunities paired with accurate assessments of their current skills and targeted personalized feedback (Ferrini-Mundy et al., 2007; Wang and Demszky, 2023; Xu et al., 2024). However, the assessment of teaching practice has limitations, including resource constraints, scalability challenges, and varying evaluator competence, as it is usually done by human evaluators (Kelly et al., 2020; Kraft et al., 2018).

Prior research has sought to overcome the limitations of manual assessment of classroom discussions by using natural language processing (Alic et al., 2022; Nazaretsky et al., 2023; Ilagan et al., 2024; Demszky et al., 2021; Suresh et al., 2019). These studies were mostly limited to analyzing turn-level teaching moves such as classifying open-ended and close-ended questions (Alic et al., 2022), labeling certain teaching strategies (Nazaretsky et al., 2023; Ilagan et al., 2024; Suresh et al., 2019), and identifying speaker contributions (Demszky et al., 2021). Assessment of discussion coherence is potentially more challenging because connections between ideas are not necessarily linear but can be hierarchical (Tao et al., 2015), and the overall coherence is not necessarily an accumulation of locally coherent moves.

Large language models (LLMs) have been successful in assessing the coherence of relatively short text, such as essays in an English proficiency

---

*This research was conducted during an internship at ETS.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 716–736
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

test (Naismith et al., 2023) and news article summaries (Liu et al., 2023; Liusie et al., 2024). However, it is still challenging for LLMs to assess classroom instruction. For example, LLMs' scores on instructional quality do not correlate with human ratings, and they fail to extract relevant utterances from classroom transcripts (Wang and Demszky, 2023). Also, they overestimate instructional quality and struggle to summarize it (Xu et al., 2024). We hypothesize that the discrepancy between LLMs' success in assessing coherence and failure to analyze instructional quality in classrooms could lie in the formulation of LLMs' tasks (Tran et al., 2024). Our goal is to evaluate the generalizability of task formulation strategies previously used to assess the coherence of short documents with LLMs to evaluate classroom discussions holistically:

RQ1 Do the task formulation strategies that work well for the coherence of short documents generalize to longer classroom discussions?

RQ2 Do the effective strategies from RQ1 generalize across subject domains (math and science)?

Our contributions are as follows:

1. We demonstrate that task formulation strategies in prior work can generalize to extended discussions, but the generalization across subject domains remains challenging.
2. We show that the strategies result in a reduction of GPT-4o's overestimation bias.
3. A closer look at the results suggests that while GPT-4o extracts utterances relevant to aspects of discussion coherence, it sometimes uses them differently from humans when justifying their answers, which raises concerns in practical real-world applications.

2 Related Work

2.1 Automated assessment of instructional quality

Prior research on automated evaluation of instructional quality in classroom discussions focused on detecting specific teacher or student discourse "moves" that characterize high-quality instruction using human-annotated corpora. Such "moves" are defined at the utterance-level and include building on student responses (Bywater et al., 2019; Demszky et al., 2021; Nazaretsky et al., 2023; Suresh et al., 2022; Tran et al., 2023), asking questions (Alic et al., 2022; Feldhus et al., 2024; Jensen et al., 2021; Tran et al., 2023), and giving supportive statements (Hunkins et al., 2022). These models are used to give feedback to teachers, showing, for example, the frequency of the target behavior in the discussion (Demszky et al., 2023; Jensen et al., 2020; Mikeska et al., 2024; Jensen et al., 2021).

More recently, LLMs have been used for holistic assessment of classroom interactions, including how effectively teachers support cognitive and language development (Whitehill and LoCasale-Crouch, 2024), to what extent classroom interactions exhibit encouragement and warmth (Hou et al., 2024), and how well tutors respond to students' math errors (Kakarla et al., 2024). However, LLMs still face challenges. For instance, ChatGPT (gpt-3.5-turbo) has low correlations with human evaluation and often fails to generate insightful and relevant suggestions for improvement (Wang and Demszky, 2023). Moreover, it overestimates instructional quality, and using its extractive summaries as inputs for the classification of instruction practices does not improve the results (Xu et al., 2024). Tran et al. (2024) have explored different task formulations to improve LLM's assessment of instructional quality, but its best-performing method is only compatible with the metrics based on the number of utterances satisfying certain criteria. We investigate prompting and task formulation strategies that are informed by recent LLM literature and can be applied to do a holistic coherence evaluation of a classroom discussion.

2.2 Automated assessment of coherence

Prior work on evaluating the coherence of a text benefited from deep neural networks, including long short-term memory (Mesgar and Strube, 2018), rational graph convolutional networks (Mesgar et al., 2021), and pretrained language models (Duari and Bhatnagar, 2022; Jeon and Strube, 2022; Zhong et al., 2022). However, these methods considered local coherence and were evaluated on tasks that could exploit it such as judging coherent and incoherent sentence pairs (Duari and Bhatnagar, 2022; Mesgar et al., 2021; Zhong et al., 2022) and short source-summary pairs (Mesgar et al., 2021; Zhong et al., 2022). BBScore (Sheng et al., 2024) captures global text coherence but treats utterances as a sequential process. This is not always the case for classroom discussions. Indeed, local coherence based on similarities of adjacent utterances had low correlations with human ratings of classroom discourse coherence (Boyle and Crossley, 2024).

LLMs enable a more holistic evaluation of coherence without modeling local coherence. Naismith et al. (2023) evaluated the coherence of pieces of writing in an English language test used for higher-education admissions, based on the Common European Framework of Reference for Languages. Liu et al. (2023) proposed using an automatic Chain of Thought (Auto CoT) to generate steps for LLMs to follow when evaluating coherence. Liusie et al. (2024) showed prompting LLMs to compare texts is more effective than prompting them to assign numerical scores. However, these studies used relatively short texts, such as essays and summaries. We incorporate their insights and evaluate the methods with long, multi-party classroom discussions.

3 Data

We used the dataset collected in previous studies Mikeska et al. (2023, 2025), where elementary pre-service teachers facilitated an argumentation-focused discussion in mathematics or science with five fictional student avatars controlled by a human actor using voice modulation software. The human actor is instructed to reflect each avatar's personality, background, and interest (e.g., "Emily is an introverted, studious, independent, serious, and literal child.") given by the researchers. Science discussions involved the Mystery Powder (**MP**) task (Mikeska et al., 2021), where students constructed arguments about the identity of a mystery powder based on its properties such as color, texture, and weight, and determined which properties were useful to identify it. The mathematics discussions focused on the Ordering Fractions (**OF**) task (Howell et al., 2021), where the learning goal was to evaluate and contrast strategies for ordering fractions with varying numerators and denominators. The teachers were given handouts on the simulated environment, the students' work before the discussion, and the goal of the discussion a week prior to the discussion. The teachers had up to 20 minutes to lead the discussion. Each discussion was video-recorded, transcribed, and timestamped for manual evaluation. One teacher facilitated at most two discussions in the dataset. Table 1 shows snippets of example transcripts from the MP task; Table 2 shows the descriptive statistics of the datasets.

The rubrics for human scoring have five dimensions, each with 2-3 supporting indicators (GO Discuss Project, 2021). Depending on the data collection phase, dimensions have three or four discrete levels and indicators are continuous between 1-3 or 1-4. This study focuses on Indicator 2A ("Overall Coherence of the Discussion") in Dimension 2 ("Facilitating a Coherent and Connected Discussion"). This indicator measures if a teacher leads a well-organized discussion focused on the content and uses the time allotted to address the given learning goal (the full rubrics are in Appendix A). Raters were current and retired K-12 teachers in STEM (Nazaretsky et al., 2023). About 27% of the discussions were double-scored; the intraclass correlations (Shrout and Fleiss, 1979) were 0.630 (MP) and 0.588 (OF). Both have moderate reliability (Koo and Li, 2016), commensurate with other dimensions (Ilagan et al., 2024; Nazaretsky et al., 2023) and other publicly available data on coherence (Gopalakrishnan et al., 2019). Raters optionally provided quotes to justify their scores.

We map a score x on the 1-4 scale to 1-3 by $\frac{2}{3}x + \frac{1}{3}$. The score distributions are in Figure 1.

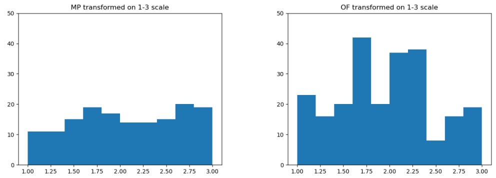

Figure 1: Score histograms in MP (left) and OF (right).

4 Experiment Setup

The MP and OF datasets were used differently. We used the MP data to develop prompts and select promising strategies; the OF data was used to test the generalization of the MP-based selections to a new domain (RQ2). For the MP data, we created four sets: five discussions used for reference (Refs), two development sets (Dev1 and Dev2), and a test set. The test set ($n = 36$) was the same as in prior work modeling other rubric dimensions (Ilagan et al., 2024; Nazaretsky et al., 2023). Using all the non-test data, we identified the five Refs discussions (see Section 5.2), then randomly chose 71 development discussions from the remainder of the data, randomizing by teachers (all discussions by the same teacher were in the same partition). We then divided the set of 71 discussions into two groups (35 and 36): Dev1 was used for experimentation with prompts, and Dev2 was used to select the most promising strategies for final testing on the MP and OF test data. For the OF dataset, we first sampled, by teacher, half the discussions for the test set ($n = 106$, from 71 teachers) and then

Coherent (human rating = 2.6)	Incoherent (human rating = 1.4)
Teacher: How about we all take about a minute to look at our own shared workspaces? ... just talk to your partner next to you about things you want to bring to the discussion about how you got to your answer, your claim, your reasoning, and just think of some evidence. So that way, when someone has a question, you can answer that question because in this discussion, it's going to be all of you having more of a discussion, and me just listening and answering questions here and there. Does that sound good? Will: Yeah. Okay. ⋮ Teacher: ... I'm going to let you take the lead like I said, so we have to make sure that we don't talk over each other and that once, and I don't have to have a conversation at all. ... So the conversation can go, Carlos to Jayla. Mina can talk to Jayla. You don't have to raise your hand, and you don't have to go in order. ... I just want to make sure that everyone understands and make sure everyone has the right answer. ... Carlos: Well, my question is for Mina and Will, and I was just wondering why you think that it's flour? Will: Well, we think that it's flour because we looked at the texture and the color and the weight, and they all matched flour. So it was pretty obvious. ⋮ Teacher: Sometimes it's easier to learn from classmates. It's sometimes easier to learn from your classmates than a teacher teaching and lecturing you, huh? Emily: Yeah. I thought everyone had really good ideas. [End of discussion]	Teacher: Today we're going to review what we've been doing for the last couple of classes. We are going to be working on identifying a substance based on its properties. Can anybody tell me what properties are? All right, Mina, what are properties? Mina: ... the properties are ... like what the powder has. Teacher: Right, like maybe characteristics? ⋮ Teacher: When we're looking at properties, you might think of a bear might have different properties than a snake. ... A bear has fur, a snake has scales. ⋮ Teacher: Yeah. What about you Jayla and Emily? You still think it's baking soda? Jayla: Yeah. Teacher: Well, you guys are right. It's baking soda. [End of discussion]

Table 1: Snippets of a coherent discussion and an incoherent one from the MP task.

	Mystery Powder (MP)	Ordering Fractions (OF)
# Transcripts	157	241
# Teachers	81	142
Av. # Utterances per Transcript	97.6	99.5
Av. # Words per Transcript	1919.6	2090.2
Av. Duration (mins)	14.5	16.7
Av. Coherence score	2.05	1.93

Table 2: Descriptive statistics of the datasets.

chose five discussions from the rest for the OF Refs set.

We test our method with GPT-4o on Azure OpenAI,[1] setting the temperature to 0 to reduce randomness. We evaluated GPT-4o predictions vs human scores using Pearson and Spearman correlations and mean squared error (MSE). For double-rated discussions, we averaged the two scores.

5 Task formulation strategies

We describe how we design our prompts. The actual prompts are in Appendix B.

5.1 Prompts to assess a single discussion

NAIVE BASELINE We prompt GPT-4o to score discussion coherence on a scale of 1-3 based on

---

[1] `https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models`

the rubric given to human raters. This rubric implements a score-level characterization strategy that describes what to expect to see in a discussion at a given score level. In addition, we give the background information on the topic, the learning goal of the classroom discussion, and the names of the student avatars. We add a new characteristic based on the justifications provided by the raters: They often pointed out that a coherent discussion had an introduction with clear and detailed learning goals and expectations. NAIVE BASELINE implements CoT (Wei et al., 2022), instructing GPT-4o to output the reasoning behind its score.

STRUCTURED COT (ST. COT) The rubric in the NAIVE BASELINE prompt characterizes highly coherent and incoherent discussions. We hypothesize that this design may prevent GPT-4o from understanding the aspects of coherence, each of which can be present or absent, or done well or badly, in a given discussion. We summarize these aspects into six bullet points and instruct GPT-4o to consider them when deriving a holistic score. This aspect-based rubric resembles the CoT prompt used to evaluate the coherence of shorter texts (Liu et al., 2023).

QUOTES Wang and Demszky (2023) have shown that LLMs cannot extract utterances relevant to instructional quality from classroom transcripts and that instructing LLMs to generate reasoning does not improve the correlation with human graders. On the other hand, Naismith et al. (2023) have found that LLMs cite examples from essays that contribute to coherence and that asking for rationale increases the correlation with humans. This line of work implies that the evaluation by LLMs can be improved if they can quote the right examples. Thus, we ask GPT-4o to provide quotes supporting the rating in CoT.

NEGATIVE FOCUS Prior work has shown that LLMs overestimate instructional quality (Xu et al., 2024). This tendency, known as leniency bias (Thakur et al., 2024), is observed when LLMs act as a judge even outside of education. Li et al. (2024) achieved better alignment between LLM and human judgment by training LLMs to generate a critical review before the final judgment. Since we use GPT-4o without fine-tuning, we ask GPT-4o to "conclude to what extent (mostly, somewhat, or seldom) the teacher **failed** to shape a coherent discussion and build ideas toward a learning goal" to make GPT-4o's reasoning critical.

5.2 Comparison between discussions

The comparison strategy is motivated by the finding that LLMs are better at comparison than assigning numerical scores, including for evaluating the coherence of summaries (Liusie et al., 2024). However, we need $O(N^2)$ comparisons to compare all discussions and fully rank them. To reduce the cost, we compare a discussion with a small set of reference discussions. Reference discussions (referred to as Refs in Section 4) are chosen so that they (1) are not in the test set, (2) are rated by two raters, and (3) have an average score between 1.7 and 2.3 on the scale of 1-3 (i.e., middle-level performance). Of all the discussions that fit the criteria, we picked five with the smallest difference in the ratings between the two raters. The number five is based on the literature on the evaluation of automated summaries that found the comparison to 4-5 reference summaries was optimal (Nenkova and Passonneau, 2004). For each reference discussion, we ask the LLM whether the discussion-to-score is better/worse than or similar to the reference. If the discussion-to-score is better than the reference, we assign a score of 3; if it is similar – 2; worse – 1. For the final continuous score, we average the scores across the reference discussions.

We incorporate the comparison paradigm into the NAIVE BASELINE prompt and the best-performing formulation strategy for a single discussion on Dev 2 by changing the LLM's task from rating to comparison. The definition of coherence in NAIVE BASELINE stays the same, except that it is now characterized by highly coherent, moderately coherent, and incoherent, instead of the score levels. We call this NAIVE BASELINE COMPARISON. Also, we apply the comparison formulation to the CoT outputs of the best strategy for a single discussion on Dev 2 because the reasoning provided by CoT might be a good summary of the degree of coherence of a discussion. We call this <STRATEGY NAME> (2 STEP), where <strategy name> is determined in the next section.

6 Results on MP dev data (Dev 2)

The top pane of Table 3 shows the results for single-discussion strategies on the MP Dev 2 set. ST. COT has the lowest MSE. QUOTES has the best Pearson correlation but has the worst MSE. NEGATIVE FOCUS trails behind the other methods. Thus, we combine the two most promising strategies, ST. COT and QUOTES. The combination shows the

Single discussion strategies	Pearson	Spearman	MSE
Naive Baseline	.503 (.458-.542)	.447 (.405-.495)	.533 (.516-.558)
St. CoT	.480 (.405-.580)	.499 (.426-.594)	**.335** (.276-.373)
Quotes	**.542** (.491-.608)	.497 (.443-.557)	.598 (.564-.620)
Negative Focus	.469 (.436-.504)	.478 (.456-.510)	.483 (.459-.504)
St. CoT+Quotes	.512 (.468-.567)	**.565** (.526-.629)	.359 (.329-.387)
Comparison strategies	Pearson	Spearman	MSE
Naive Baseline Comparison	**.584** (.562-.604)	**.607** (.587-.631)	**.326** (.316-.334)
St. CoT+Quotes Comparison	.555 (.538-.572)	.596 (.557-.625)	.496 (.456-.549)
St. CoT+Quotes (2-Step)	.538 (.505-.590)	.550 (.506-.628)	.352 (.316-.399)

Table 3: Pearson and Spearman correlations (higher numbers are better) and MSE (lower numbers are better) for the single-discussion formulations (top) and the comparison formulations (bottom) on Dev 2. We report an average and a range of five runs. The best result is in bold, and the second-best result is underlined.

	MP (n = 36)			OF (n = 106)		
Strategies	Pearson	Spearman	MSE	Pearson	Spearman	MSE
Naive Baseline	.574 (.548-.599)	.578 (.547-.611)	.493 (.481-.504)	.167 (.140-.210)	.139 (.110-.183)	.754 (.715-.788)
St. CoT + Quotes	.663 (.592-.730)	.607 (.541-.692)	.272 (.233-.317)	**.416** (.389-.447)	**.420** (.394-.451)	**.405** (.379-.439)
Naive Baseline Comparison	**.708** (.686-.732)	**.702** (.664-.736)	**.236** (.219-.252)	.308 (.280-.341)	.328 (.295-.365)	.523 (.507-.553)

Table 4: Results on test sets; reported are the average and range of five runs. The best performance is in bold.

best or second-best performance and outperforms the Naive Baseline on all metrics. Therefore, we create St. CoT+Quotes (2 step) as a 2-step comparison strategy.

The bottom pane of Table 3 shows the results for comparison strategies. The results support the effectiveness of comparing the discussion-to-be-scored with references. The comparison versions of Naive Baseline and St. CoT+Quotes perform better than their single-discussion versions on all metrics, both in terms of average performance and stability (narrower range), apart from MSE for St. CoT+Quotes. The results do not support the two-step formulation. This implies that the description of a discussion does not capture the information necessary for the comparison, consistently with prior literature (Xu et al., 2024).

For the final evaluation of test data, we select Naive Baseline Comparison, as it showed the best performance on Dev 2. We also evaluate the St. CoT+Quotes single-discussion formulation, since it performs best in the more resource-lean scenario without reference discussions. The Naive Baseline scoring scenario will also be evaluated on test data to check whether gains over baseline are replicated in the test results.

7 Final test results

Table 4 shows the results on the test sets of MP and OF. To answer RQ1 (generalizability of strategies to classroom discussion), we compare the rows. Our results support the generalization of the strategies evaluated on short text to long classroom discussions: Both St. CoT+Quotes and Naive Baseline Comparison outperformed Naive Baseline on all metrics. RQ2 (generalizability across subject domains) is answered by comparing the columns. We observe that the performance on the OF data is generally much worse, across formulations and metrics, than on MP data.

We further analyze how the task formulation strategies address the limitations of LLMs in assessing instructional quality found in the literature.

Overestimation of quality One of the limitations is that LLMs tend to overestimate the instructional quality (Xu et al., 2024). To check this tendency, we plot GPT-4o predictions vs human scores in Figure 2, using the runs with the median MSE out of five. The Naive Baseline exhibits overestimation, as most of the points are above the diagonal; the median scores are 2.75 (MP) and 2.5 (OF). St. CoT+Quotes reduces the median scores to 2.5

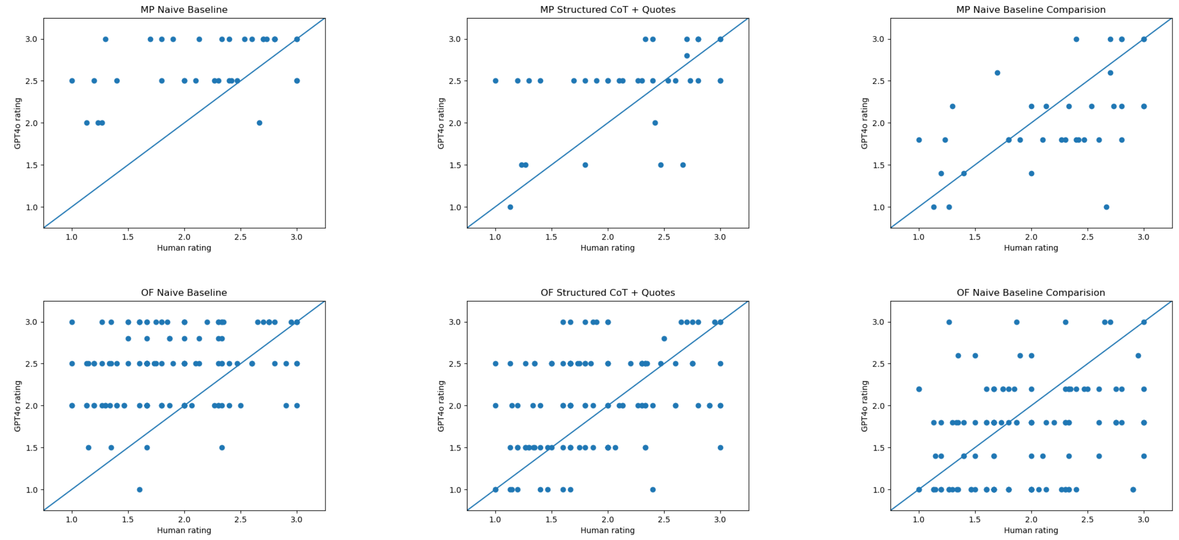

Figure 2: Scatter plots of GPT-4o predictions vs human scores. The top row is MP (Dev 2), and the bottom row is OF (discussions not in Refs or test set). The points above the diagonal are over-estimations by GPT-4o.

(MP) and 2 (OF). Naive Baseline Comparison further pushes down the scores; >60% of the discussions receive 2±0.2 points for MP, and >90% of the discussions receive 1.6 ± 0.6 points for OF. Thus, our results confirm the over-scoring by Naive Baseline observed in the literature and suggest that the St. CoT+Quotes and Naive Baseline Comparison formulations help reduce it.

Inability to provide relevant quotes Another limitation is that many quotes provided by LLMs are unfaithful or irrelevant (Wang and Demszky, 2023). Therefore, we investigated the quotes produced by St. CoT+Quotes, using the same runs as above. We sampled six discussions (three MP, three OF) with at least 50 words in their human justifications. St. CoT+Quotes provided more quotes than humans: 4.33 vs 1.83 per discussion, on average. All quotes given by St. CoT+Quotes exist in the discussions. In five out of six discussions, the system picked at least one quote also picked by a human. Generally, the system picked quotes relevant to the specific aspect; see Table 5 for example output. However, the *interpretations* of the quote were not always aligned. For example, “Do you all feel like you learned something from this?” was cited as justification for a critical evaluation of the teacher’s conclusion by a human rater (“Conclusion is minimal”), whereas the system picked the same quote to justify the evaluation “The teacher summarized important conclusions consistently throughout the discussion.” Thus, while the system’s quotes show the grounding of the aspects in the discussion, they do not consistently support the *evaluation* of the aspect. Other five discussions are in Appendix C.

8 Discussion

Our results show the promise for generalizability of task formulation strategies for the evaluation of the coherence of short text to the evaluation of teachers’ ability to lead coherent discussions. Further analyses suggest that these strategies address previously reported challenges: overestimation bias (Xu et al., 2024) and the inability to extract relevant parts of discussions (Wang and Demszky, 2023). However, we also highlight two challenges in deploying GPT-4o in classroom discussion assessment.

The first challenge is the generalizability across subject domains (RQ2). This would be a bottleneck in real-world applications because it implies that we should tune prompts for every subject, increasing cost and time and making it difficult to generalize without a decent-sized corpus. Moreover, inconsistent scoring across subjects threatens fairness in professional development and evaluation of teachers. For instance, if we used GPT-4o’s assessment to determine the need for professional development, teachers in the more mis-scored subjects (math in our example) would suffer from more mis-allocation of resources such as coaching sessions than teachers in other subjects even when teaching practice is comparable.

The second challenge is the misalignment in the interpretation of discussions and quotes between

Human	The discussion is mostly organized, purposeful, and focused on the content at hand. The teacher presents the task and sets expectations for the discussion. She systematically moves from group to group, sharing key ideas. Transitions are timely and appropriate. Conclusion is minimal ("Yeah, you did. And everyone had good methods. Do you all feel like you learned something from this? Sometimes it's easier to learn from classmates. It's sometimes easier to learn from your classmates than a teacher teaching and lecturing you, huh?")
GPT	1. The teacher did not state a clear, detailed learning goal and expectation at the beginning. The teacher mentioned that the students would discuss their findings and answer questions, but did not explicitly state the learning goal of constructing an argument based on properties. For example, the teacher said, 'in this discussion, it's going to be all of you having more of a discussion, and me just listening and answering questions here and there.' 2. The teacher moved between groups of ideas in a timely manner with apparent rationale. The teacher facilitated the discussion by asking each group to share their findings and then prompted students to ask questions, which kept the discussion focused and organized. For example, 'All right. So, let's just pick one person from our group... Mina and Will, go ahead and pick one person to share what your claim was, what your evidence was and what your answer was.' 3. The teacher instructed with deliberation, ensuring that students understood the process and the importance of respectful conversation. For example, 'I'm going to let you take the lead like I said, so we have to make sure that we don't talk over each other.' 4. The teacher introduced well-chosen new ideas that benefited student understanding at relevant moments. For example, the teacher clarified the importance of checking all properties but also understanding which ones are most useful, 'Sometimes in science, and like Carlos said, Carlos did check the weight. He just realized that the weight was not as useful.' 5. The teacher summarized important conclusions consistently throughout the discussion. For example, 'Does everyone understand how Carlos did that?' and 'Do you all feel like you learned something from this?' 6. The teacher connected a series of ideas to the learning goal by guiding students to understand which properties were most useful in identifying the mystery powder. For example, 'Carlos, I'm reading your answer. Could you maybe tell us maybe one of the properties that aren't that important? Maybe that would help Mina and Will.' Overall, the discussion was coherent and well-facilitated, but the initial learning goal could have been stated more clearly.

Table 5: Human evaluation and STRUCTURED COT + QUOTES (GPT) output for the coherent example in Table 1. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations.

humans and GPT-4o. This is concerning when we base feedback for teachers on GPT-4o's outputs because it would give teachers incorrect action items or miss opportunities for improvement. In the example in Table 5, the human evaluation suggests the conclusion is the area for improvement; the teacher could have elaborated more than just saying "Do you all feel like you learned something from this? ..." However, ST. COT+QUOTES identifies it as a good conclusion. Instead, it recommends "explicitly stating the learning goal of constructing an argument based on properties," which was already achieved according to the human evaluation. This misalignment could undermine the validity and usability of GPT-4o in generating something more than scores, including feedback. A potential remedy could be retrieving relevant pre-defined human-written feedback based on the score, but it cannot fully utilize LLMs' advantages in flexibility and personalization. This motivates future work on in-depth analysis of human and LLM quotes and on improving the evaluation of quotes selected by GPT-4o.

9 Conclusion

We evaluated task formulation strategies to assess the coherence of classroom discussions. Our results show that strategies previously evaluated for assessing the coherence of short text, such as essays or summaries, successfully generalize to assessing much longer texts—transcripts of 20-minute-long simulated classroom discussions. We reveal that these strategies help GPT-4o tackle the limitations pointed by the literature: overestimation of instructional quality and failure to quote relevant utterances from discussions. However, they do not show cross-domain generalization even within the same simulated setting. Our study serves as a step toward supporting teachers' development with automated personalized feedback by providing accurate automated evaluation of the target skill, though challenges still remain.

Limitations

We acknowledge the limitations of our evaluation. First, the generalizability of our findings should be explored with other LLMs and datasets. Our results demonstrate some generalizability of coherence evaluation methods from other genres (essays and summaries) to our context but also show that generalization across STEM subjects within the same simulated classroom context is not straightforward since performance is lower on OF than MP. Improving generalization across content domains is our most immediate goal. In addition, we implicitly show the generalizability across models because the prior work our prompts are based on uses models different from ours: GPT-4 for ST. COT (Liu et al., 2023) and QUOTES (NAISMITH ET AL., 2023) and open-source LLMs, including FlanT5 and Llama2, for the comparison strategy (Liusie et al., 2024). Although our results imply that the strategies in this paper are potentially generalizable to other models, further experiments would be necessary to verify it.

Second, our implementation of the comparison formulation compares discussions only with moderately coherent reference discussions and results in excessive lowering of scores. We leave it to future work to explore strategies for selecting reference discussions that could help mitigate this excessive correction of over-scoring.

Third, human quotes are not the "gold standard" since the raters were asked to provide some examples from the discussion (see Section 3); there are potentially other good quotes that weren't selected. The analysis in Section 7 motivates future work to improve the evaluation of quotes selected by GPT-4o.

Finally, our experiments are done only in simulated classrooms. These are important for scaling up practice opportunities by allowing teachers to repeat the cycle of practice and reflection on their teaching without harming real students by their mistakes (Dalinger et al., 2020; Dieker et al., 2014). Generalizability to real classrooms with real students is also important. However, since our goal of scaling up feedback aligns better with the advantages of simulated classrooms, we prioritized this exploration on data from simulated discussions, leaving exploration of real-life discussions to future work.

Ethical considerations

We would like to address potential ethical concerns. First, giving student names and the whole discussions to GPT-4o is not a breach of privacy. In this work, we are not using data from real elementary students. Instead, all the data comes from responses from elementary student avatars in a simulated classroom. The student avatars are operated by an adult, called a simulation specialist, who is trained to use specialized equipment (e.g., game controllers, voice modulation software, etc.) to sound, move, and respond like upper elementary students (cf. Section 3). Each teacher participant signed a consent form that provides their written approval for the research use of the video-recorded discussion and who it can and cannot be shared with. Video recordings are only shared outside of our research team if the participant has consented to that use. For this study, no video recordings were used; we used de-identified transcripts for analyses.

Second, LLMs could be susceptible to their algorithmic biases. Our work addresses bias concerns by showing how to reduce overestimation (bias against low-performing teachers) in Section 7. The model's explanations could be biased, too, and might not be pedagogically sound (cf. Section 7). As discussed above, instead of giving teachers the model's explanations as they are as feedback, we plan to use the scores and outputs to provide the teachers with feedback by, for example, retrieving relevant pre-defined human-written feedback.

The costs of using GPT-4o and collecting and scoring discussion data for model development could also be a barrier to applying our results to the real world. However, the best performing method, STRUCTURED COT + QUOTES, is zero-shot and does not require any reference discussions. Thus, it works well in resource-constrained settings. The price of GPT-4o is \$0.00250 / 1K input tokens and \$0.01000 / 1K output tokens as of writing. Since the average number of words in discussions is around 2K and the output is usually no more than 500 words (cf. Tables 2 and 5), the cost per discussion is less than \$0.1. Therefore, our method scales well at low cost.

Acknowledgments

This study used data collected on previous projects funded by the National Science Foundation (grants #2037983, #1621344, and #2032179). The opinions expressed herein are those of the authors and not the funding agency. We are grateful for the elementary pre-service teachers who participated in these earlier studies and for the contributions of practicing and retired teachers in scoring these discussion transcripts.

References

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of Innovative Use of NLP for Building Educational Applications (BEA)*, pages 224–233.

Jessica Boyle and Scott Crossley. 2024. Semantic similarity of teacher and student discourse linked to quality ratings from classroom observations. In *Proceedings of Educational Data Mining (EDM)*, pages 797–801.

James P Bywater, Jennifer L Chiu, James Hong, and Vidhya Sankaranarayanan. 2019. The teacher responding tool: Scaffolding the teacher practice of responding to student ideas in mathematics classrooms. *Computers & Education*, 139:16–30.

Stacey L Carpenter, Jiwon Kim, Katherine Nilsen, Tobias Irish, Julie A Bianchini, and Alan R Berkowitz. 2020. Secondary science teachers' use of discourse moves to work with student ideas in classroom discussions. *International Journal of Science Education*, 42(15):2513–2533.

XI Chen and Yeping Li. 2010. Instructional coherence in chinese mathematics classroom—a case study of lessons on fraction division. *International Journal of Science and Mathematics Education*, 8:711–735.

Tara Dalinger, Katherine B Thomas, Susan Stansberry, and Ying Xiu. 2020. A mixed reality simulation offers strategic practice for pre-service teachers. *Computers & Education*, 144:103696.

C. Danielson. 2013. *The Framework for Teaching: Evaluation Instrument*. Danielson Group.

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1638–1653.

Lisa A Dieker, Jacqueline A Rodriguez, Benjamin Lignugaris/Kraft, Michael C Hynes, and Charles E Hughes. 2014. The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, 37(1):21–33.

Swagata Duari and Vasudha Bhatnagar. 2022. Ffcd: A fast-and-frugal coherence detection method. *IEEE Access*, 10:85305–85314.

Nils Feldhus, Aliki Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of Information Technology for Social Good*, page 225–230.

Joan Ferrini-Mundy, Gail Burrill, and William H Schmidt. 2007. Building teacher capacity for implementing curricular coherence: Mathematics teacher professional development tasks. *Journal of Mathematics Teacher Education*, 10:311–324.

GO Discuss Project. 2021. Scoring. Qualitative Data Repository.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Deborah Hanuscin, Kelsey Lipsitz, Dante Cisterna-Alburquerque, Kathryn A Arnone, Delinda van Garderen, Zandra de Araujo, and Eun Ju Lee. 2016. Developing coherent conceptual storylines: Two elementary challenges. *Journal of Science Teacher Education*, 27:393–414.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In *Proceedings of Artificial Intelligence in Education (AIED)*, pages 60–74.

Heather Howell, Jamie Mikeska, Jessica Tierney, Benjamin Baehr, and Penny Lehman. 2021. Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics: The ordering fractions task. *ETS Research Memorandum No. 21-10*.

Nicholas Hunkins, Sean Kelly, and Sidney D'Mello. 2022. "beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In *Proceedings of Learning Analytics and Knowledge (LAK)*, pages 230–238.

Michael Ilagan, Beata Beigman Klebanov, and Jamie Mikeska. 2024. Automated evaluation of teacher encouragement of student-to-student interactions in a simulated classroom discussion. In *Proceedings of Innovative Use of NLP for Building Educational Applications (BEA)*, pages 182–198.

Emily Jensen, Meghan Dale, Patrick J. Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K. D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of Human Factors in Computing Systems*, pages 1–13.

Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. In *Proceedings of Learning Analytics and Knowledge (LAK)*, page 302–312.

Sungho Jeon and Michael Strube. 2022. Entity-based neural local coherence modeling. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 7787–7805.

Sanjit Kakarla, Danielle Thomas, Jionghao Lin, Shivang Gupta, and Ken Koedinger. 2024. Using large language models to assess tutors' performance in reacting to students making math errors. In *AI for Education: Bridging Innovation and Responsibility at the AAAI Conference on AI (AAAI)*.

Sean Kelly, Robert Bringe, Esteban Aucejo, and Jane Cooley Fruehwirth. 2020. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28:62–62.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Matthew A Kraft, David Blazar, and Dylan Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4):547–588.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, and 1 others. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 139–151.

Mohsen Mesgar, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2021. A neural graph-based local coherence model. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2316–2321.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 4328–4339.

Jamie N. Mikeska, Beata Beigman Klebanov, Alessia Marigo, Jessica Tierney, Tricia Maxwell, and Tanya Nazaretsky. 2024. Exploring the potential of automated and personalized feedback to support science teacher learning. In *Proceedings of Artificial Intelligence in Education (AIED)*, pages 251–258.

Jamie N Mikeska, Dionne Cross Francis, Pamela S Lottero-Perdue, Meredith Park Rogers, Calli Shekell, Pavneet Kaur Bharaj, Heather Howell, Adam Maltese, Meredith Thompson, and Justin Reich. 2025. Promoting preservice teachers' facilitation of argumentation in mathematics and science through digital simulations. *Teaching and Teacher Education*, 154:104858.

Jamie N Mikeska, Heather Howell, Joseph Ciofalo, Adam Devitt, Elizabeth Orlandi, Kenneth King, and G Simonelli. 2021. Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics: The mystery powder task. *Research Memorandum No. RM-21-06, Educational Testing Service*.

Jamie N Mikeska, Heather Howell, and Devon Kinsey. 2023. Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education*, 74(5):422–436.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of Innovative Use of NLP for Building Educational Applications (BEA)*, pages 394–403.

National Research Council. 2012. A framework for k-12 science education: Practices, crosscutting concepts, and core ideas. *National Academy of Sciences*.

Tanya Nazaretsky, Jamie N Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *Proceedings of Learning Analytics and Knowledge (LAK)*, pages 122–132.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 145–152.

RC Pianta. 2008. Classroom assessment scoring system™: Manual k-3. *Paul H Brookes Publishing*.

Julia D Plummer and Arzu Tanis Ozcelik. 2015. Preservice teachers developing coherent inquiry investigations in elementary astronomy. *Science Education*, 99(5):932–957.

John Ramsey. 1993. Developing conceptual storylines with the learning cycle. *Journal of Elementary Science Education*, 5(2):1–20.

Kathleen J Roth, Helen E Garnier, Catherine Chen, Meike Lemmens, Kathleen Schwille, and Nicole Wickler. 2011. Videobased lesson analysis: Effective science pd for teacher and student learning. *Journal of Research in Science Teaching*, 48(2):117–148.

Zhecheng Sheng, Tianhao Zhang, Chen Jiang, and Dongyeop Kang. 2024. Bbscore: A brownian bridge based metric for assessing text coherence. In *Proceedings of the AAAI Conference on AI (AAAI)*, pages 14937–14945.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420.

Mary Kay Stein, Randi A Engle, Margaret S Smith, and Elizabeth K Hughes. 2008. Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, 10(4):313–340.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Finetuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of Innovative Use of NLP for Building Educational Applications (BEA)*, pages 71–81.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI conference on AI (AAAI)*, pages 9721–9728.

Wang Tao, Cai Jinfa, and Hwang Stephen. 2015. Achieving coherence in the mathematics classroom: Toward a framework for examining instructional coherence. In *How Chinese Teach Mathematics: Perspectives From Insiders*, pages 111–148. World Scientific.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2023. Utilizing natural language processing for automated assessment of classroom discussion. In *Proceedings of Artificial Intelligence in Education (AIED). Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 490–496.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment. In *Proceedings of Educational Data Mining (EDM)*, pages 500–510.

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

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

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4375–4389.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multidimensional evaluator for text generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2023–2038.

A Rubrics

Human raters evaluated a teacher's ability to lead a coherent discussion based on the rubrics in Table 6

Level Label	Description
1 Beginning	Discussion has a weak sense of organization, purpose, and focus.
2 Developing	Discussion is somewhat organized, purposeful, and focused on the content at hand. AND/OR Portions of the discussion are strongly variable with respect to organization, purpose, and focus.
3 Well-prepared	Discussion is mostly organized, purposeful, and focused on the content at hand.
4 Commendable	Discussion is organized, purposeful, and focused on the content at hand. AND The teacher uses the time allotted so that the learning goal is achieved.

Table 6: Rubrics for Indicator 2A ("Overall Coherence of the Discussion") (GO Discuss Project, 2021)

and the observation notes in Table 7 (GO Discuss Project, 2021).

B LLM Prompts

B.1 Naive baseline

The prompt used as a baseline is the following:

> Following is a discussion about <task information>. <task learning goal> <student information> <rating instruction> <coherence definition>
>
> A score of 3 is characterized by <score 3 characteristics>
>
> A score of 2 is characterized by <score 2 characteristics>
>
> A score of 1 is characterized by <score 1 characteristics>
>
> Please output your score and reasoning in the following JSON format: {"reason": "...", "score": "a float number between 1-3"}.
>
> <the discussion to score>

For Indicator 2a, only consider evidence of organization or planning that is connected to the intended student learning goal.
A score of 4 is characterized by a strong degree of coherence around the content and ideas that are discussed **and** the teacher's successful use of the available time during the discussion to address the learning goal.
A score of 3 is characterized by a strong degree of coherence around the content and ideas that are discussed. For example: • Transitions between ideas and/or groups are timely and make sense. • Instruction takes place in ways that suggest deliberation on the part of the teacher. • New ideas that are introduced are well chosen and occur at relevant moments. Note that you can score a 3 even if the teacher does not achieve the learning goal by the end of the discussion.
A score of 2 is characterized by a variable degree of coherence around the content and ideas that are discussed. For example, different portions of the discussion might be scored as a 1 or 3 if viewed separately. At least some portion of the discussion is highly coherent.
A score of 1 is characterized by a lack of coherence around the content and ideas that are discussed. For example: • Discussion has a weak sense of purpose and trajectory. • Teacher moves between ideas abruptly and without apparent rationale. • Teacher introduces new ideas that have limited potential for benefiting student understanding. • Important conclusions may be left unstated or inconsistently summarized. • Discussion may be characterized as a series of unconnected ideas taken up one at a time.

Table 7: The observation notes provided to human raters (GO Discuss Project, 2021).

<task information> and <task learning goal> are dependent on the dataset. For MP, <task information> is *"identifying a mystery powder in a science classroom"*, and <task learning goal> is *"The learning goal is that students will construct an argument about the identity of a mystery powder based on its properties and come to a consensus about which properties are most useful in identifying the unknown powder."* For OP, <task informa-

tion> is *“ordering fractions in a math classroom”*, and <task learning goal> is *“The learning goal is that students will evaluate, justify, compare, and contrast strategies for ordering fractions with different numerators and denominators.”* These descriptions are taken from the handouts given to the teachers in the dataset before they facilitate discussions (Mikeska et al., 2023, 2025).

<student information> is *“Mina, Will, Emily, Jayla, and Carlos are students.”*. <rating instruction> is *“Your task is to rate the discussion based on its coherence on a scale of 1-3.”* <coherence definition> is *“To be coherent, a discussion must be organized, purposeful, and focused on the content at hand, and the teacher must use the time allotted so that the learning goal is achieved.”*

<score 3 characteristics>: a strong degree of coherence around the content and ideas that are discussed and the teacher’s successful use of the available time during the discussion to address the learning goal. For example,

> - The teacher states a clear, detailed learning goal and expectation at the beginning.
>
> - Transitions between ideas and/or groups are timely and make sense.
>
> - Instruction takes place in ways that suggest deliberation on the part of the teacher.
>
> - New ideas that are introduced are well chosen and occur at relevant moments.

<score 2 characteristics>: a variable degree of coherence around the content and ideas that are discussed. For example, different portions of the discussion might be scored as a 1 or 3 if viewed separately. At least some portions of the discussion are highly coherent.

<score 1 characteristics>: a lack of coherence around the content and ideas that are discussed. For example:

> - Discussion has a weak sense of purpose and trajectory.
>
> - Teacher moves between ideas abruptly and without apparent rationale.
>
> - Teacher introduces new ideas that have limited potential for benefiting student understanding.
>
> - Important conclusions may be left unstated or inconsistently summarized.
>
> - Discussion may be characterized as a series of unconnected ideas taken up one at a time.

These score characteristics are adopted from the observation notes in Table 7. Only the discussion is sent as a user input to GPT-4o, and the rest is sent as a system input.

B.2 Prompts for the single discussion strategies

Only the discussion is sent as a user input to GPT-4o, the rest is sent as a system input.

STRUCTURED COT

> Following is a discussion about <task information>. <task learning goal> <student information> <rating instruction> To do so, first, read the discussion carefully. Then, describe whether the teacher succeeded in doing or failed to do each of the following:
>
> <aspects of coherence>
>
> In the end, rate the discussion on a scale of 1-3.
>
> Please output your description and score in the following JSON format: {“description”: “1. The teacher ...”, “score”: “a float number between 1-3”}.
>
> <the discussion to score>

<aspects of coherence>:

> 1. state a clear, detailed learning goal and expectation at the beginning,
>
> 2. move between (groups of) ideas timely with apparent rationale,
>
> 3. instruct with deliberation,
>
> 4. introduce well-chosen new ideas that benefit student understanding at relevant moments,
>
> 5. summarize important conclusions consistently throughout the discussion, and
>
> 6. connect a series of ideas to the learning goal.

QUOTES The prompts is the same as NAIVE BASELINE, with the addition of the following right before "Please output your score...":

> When you rate the discussion, provide quotes from it in your reasoning to support your score.

NEGATIVE FOCUS The prompts is the same as NAIVE BASELINE, with the addition of the following right before "Please output your score...":

> When you rate the discussion, provide your reasoning and conclude to what extent (mostly, somewhat or seldom) the teacher failed to shape a coherent discussion and build ideas toward the learning goal.

STRUCTURED COT + QUOTES prompt The prompts is the same as STRUCTURED COT, with the addition of the following right before "In the end, rate...":

> When you describe each of the above aspects, provide quotes from the discussion in your reasoning to support your score.

B.3 Prompts for the comparison strategies

Only the part starting from "Here's the first discussion;" is sent as a user input, the rest is sent as a system input. We optimized the ordering of the discussions for each prompts using Dev 1 because it impacts the decisions (Liusie et al., 2024).

NAIVE COMPARISON BASELINE

> Following is a discussion about <task information> <task learning goal> <student information> <comparison instruction> <coherence definition>
>
> A highly coherent discussion is characterized by <score 3 characteristics>
>
> A moderately coherent discussion is characterized by <score 2 characteristics>
>
> An incoherent discussion is characterized by <score 1 characteristics>
>
> You may say that the first discussion has a similar coherence to the second one.
>
> Please output your decision and reasoning in the following JSON format: {"reason": "...", "The first discussion is": "similar/better/worse"}.
>
> Here's the first discussion;
>
> <a reference discussion>
>
> Here's the second discussion
>
> <the discussion to score>

<comparison instruction> is *"Your task is to determine whether the first discussion is better or worse than the second one based on their coherence."*

ST. COT+QUOTES COMPARISON

> Following are two discussions about <task information> <task learning goal> <student information> <comparison instruction> To do so, first, read both discussions carefully. Then, for each discussion, describe whether the teachers succeeded in doing or failed to do each of the following:
>
> <aspects of coherence>
>
> When you describe each of the above aspects, provide quotes from the discussions in your reasoning to support your comparison.
>
> In the end, decide whether the first discussion is better or worse than the second one. You may say that the first discussion has a similar coherence to the second one.
>
> Please output your decision and reasoning in the following JSON format: {"reason": "1. In the first discussion, ...", "The first discussion is": "similar/better/worse"}.
>
> Here's the first discussion;
>
> <the discussion to score>
>
> Here's the second discussion
>
> <a reference discussion>

ST. COT+QUOTES COMPARISON - TWO STEP Step 1 prompt is the following:

> Following is a discussion about <task information>. <task learning goal> <student information> Your task is to describe the coherence of the discussion. To do so, first, read the discussion carefully. Then, describe whether the teacher succeeded in doing or failed to do each of the following:

<aspects of coherence>

> When you describe each of the above aspects, provide quotes from the discussion to support your description.

Step 2 prompt is the following:

> Following are the evaluations of two different classroom discussions about <task information> <comparison instruction> To do so, first, read each aspect in both evaluations carefully, including the quotes in the evaluations. In the end, decide whether the first discussion is better or worse than the second one. You may say that the first discussion has a similar coherence to the second one.
>
> Please output your decision and reasoning in the following JSON format: {“reason”: “...”, “The first discussion is”: “similar/better/worse”}.
>
> Here’s the evaluation of the first discussion;
>
> <the output of step 1 for a reference discussion>
>
> Here’s the evaluation of the second discussion;
>
> <the output of step 1 for the discussion to score>

C Additional comparisons between human and ST. COT+QUOTES

Tables 8-12 have additional comparisons between justifications by human raters and ST. COT+QUOTES.

Human	The discussion was mostly organized with portions that were harder to follow. At the beginning of the lesson, the teacher said (e.g., today we're going to look at everyone's claim from our experiment. And, I want us all to work on constructing, defending and critiquing ours and others' claims) and then during the discussion, references the properties chart to help guide the students toward the learning goal.
GPT	1. The teacher did not state a clear, detailed learning goal and expectation at the beginning. The teacher mentioned, 'I want us all to work on constructing, defending and critiquing ours and others' claims,' but did not provide a detailed learning goal or clear expectations. 2. The teacher moved between groups of ideas timely with apparent rationale. For example, the teacher transitioned from discussing the importance of weight to other properties like texture and color, and then to the reactions with vinegar and water. 3. The teacher instructed with deliberation, asking students to turn and talk, and then come back together to share their thoughts. 4. The teacher introduced well-chosen new ideas that benefited student understanding at relevant moments, such as explaining why weight might not be as important and guiding students to see that color was not a distinguishing property. 5. The teacher summarized important conclusions consistently throughout the discussion, such as reiterating that weight and color were not as useful in identifying the mystery powder. 6. The teacher connected a series of ideas to the learning goal by guiding students to use the most relevant properties (texture, reaction with vinegar, and reaction with water) to identify the mystery powder as baking soda.

Table 8: Human evaluation and STRUCTURED COT + QUOTES (GPT) output. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations.

Human	The T made a small attempt to frame the purpose of the discussion at the start by asking the students if they knew what was meant by “properties”. The students did try to provide information, but the T wound up taking over and giving a weak example of what properties were. The T had each group share their claim and evidence. After each group shared, the T critiqued their evidence while reading the groups’ prework section containing their rationale about the most important properties (Key Idea 3). There was a weak review of the properties at the end of the discussion—again with the T providing the information. There was a short “consensus” discussion where the T polled the students to see if they thought the mystery powder was baking soda. The T had a very weak ending.
GPT	1. The teacher did not state a clear, detailed learning goal and expectation at the beginning. The teacher mentioned, ‘Today we’re going to review what we’ve been doing for the last couple of classes. We are going to be working on identifying a substance based on its properties,’ but did not elaborate on the specific learning goal or expectations. 2. The teacher moved between ideas timely with apparent rationale, such as transitioning from discussing properties to having students share their answers and reasoning. For example, ‘Talk among yourselves for a couple of minutes about what the mystery powder is and how you can support your answer. You guys ready to share your answers with the class?’ 3. The teacher instructed with deliberation, providing clear explanations and feedback. For instance, ‘That’s a good explanation. You guys said that baking soda is white, smooth, bubbles when mixed with vinegar, looks clear when is mixed with water and weighs 24 grams.’ 4. The teacher introduced well-chosen new ideas that benefited student understanding at relevant moments, such as explaining why weight might not be a useful property in this context: ’Weight just tells you how much you have, so I didn’t bother with it.’ 5. The teacher summarized important conclusions consistently throughout the discussion, such as reiterating the properties to consider: ‘When we do look at properties, you want to consider everything, because in this chart, something can’t be smooth and rough.’ 6. The teacher connected a series of ideas to the learning goal by discussing the properties and how they help identify the mystery powder, but the connection could have been stronger with a clearer initial statement of the learning goal.

Table 9: Human evaluation and STRUCTURED COT + QUOTES (GPT) output. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations.

Human 1	The discussion overall was coherent and focused on the learning goal. The T began the discussion by explaining that they will be talking about the fractions they ordered. The T then shares that they will begin by explaining the strategy they used and why it works for any set of fractions. After the students finished sharing their strategies and why they used them, the T then shifted to having the students think about other groups strategies. Once Mina came to the realization that her strategy did not work, the T then transitions the discussion to think about whether the other strategies used could be used with any set of fractions. To end the discussion, the T asks the student what strategy they would use if she gave them the same problem again based on the discussion. By the end the learning goal was achieved. Mina realized that her strategy would not work and the other groups realized their strategy could not be generalized to all fraction sets.
Human 2	The T opens the discussion by saying, “Today we’re going to be talking about ordering fractions and the worksheet that you guys did with three-ninths, nine-tenths, oh three-tenths, nine-tenths and three-fourths. So first I want everyone to go around the room and I will call on you. And you’re going to explain your strategy and why you think it works for any set of fractions. First, we’re going to start off with Will and Jayla. I want you to explain to the group what you did for your strategy.” She then asks each group to share their strategy. Then, she focuses on one strategy at a time and asks others what they think about it. After that (and convincing M she is incorrect), she discusses the answers from one group at a time to the last question. Finally she asks them what they would do “If I were to give you this problem again, after listening to everyone’s strategy and what the limitations were and what they did and why they did it, would you do a, would you use a different strategy for the next time? I want Mina to go first.” So overall, the discussion was pretty organized. However, there were times when the T transitioned to a different idea rather than continuing with a S’s thought; for example, when C brought up that M’s strategy would work if the fractions had the same denominator, the T said, “So now we’re going to jump back to Will and Jayla. So now we’re going to talk about the second question.” Also, the T introduced a new representation (rectangles) instead of using W and J’s number line to convince M her answer was incorrect.
GPT	1. The teacher did not state a clear, detailed learning goal and expectation at the beginning. The teacher mentioned the topic but did not articulate the specific learning objectives. For example, the teacher said, ‘Today we’re going to be talking about ordering fractions and the worksheet that you guys did,’ but did not elaborate on the learning goals. 2. The teacher moved between groups of ideas timely with apparent rationale, ensuring each student had a chance to explain their strategy and then facilitating a discussion on the strengths and limitations of each method. For instance, the teacher asked each group to explain their strategy and then prompted other students to comment on it. 3. The teacher instructed with deliberation, asking follow-up questions to ensure students explained their reasoning. For example, the teacher asked Mina to explain more about her strategy and why she chose it. 4. The teacher introduced well-chosen new ideas that benefited student understanding at relevant moments, such as when the teacher provided a visual explanation to help Mina understand why her method did not work. 5. The teacher did not consistently summarize important conclusions throughout the discussion. While the teacher did clarify points at times, there was no consistent summarization of key takeaways after each strategy discussion. 6. The teacher connected a series of ideas to the learning goal by discussing the limitations and strengths of each strategy, but this connection could have been more explicit. The teacher asked students to reflect on whether they would use a different strategy next time, which helped tie the discussion back to the learning goal.

Table 10: Human evaluation and STRUCTURED COT + QUOTES (GPT) output. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations.

Human	No discussion guidelines or introduction provided. Introduced unproductive or confusing tangents (asked Ss to define numerator and denominator but accepted simplistic description of the top number vs. the bottom number; started drawing an incorrect area model to help M demonstrate her strategy; suggested that E/C use a picture or time line to demonstrate their strategy.) T provided some clear and timely transitions between discussions of different strategies.
GPT	1. The teacher clearly stated the learning goal at the beginning: 'So today students, we are going to evaluate, justify, compare and contrast strategies for ordering fractions with different numerators and different denominators.'* This sets a clear expectation for the lesson. 2. The teacher moved between ideas timely and with apparent rationale, such as transitioning from discussing Jayla and Will's number line strategy to practicing with new fractions: 'I think this might be easier if we practice on some new numbers, okay?' 3. The teacher instructed with deliberation, ensuring students understood each step before moving on, as seen when the teacher asked students to place dots on the number line and confirmed their understanding: 'Yes. Good job. So now from this number line, we know that five-eighths is greater than three-eighth, but we still don't know where one fourth lands.' 4. The teacher introduced well-chosen new ideas at relevant moments, such as comparing fractions to one-half to determine their relative sizes: 'The next strategy is Mina's strategy. Do you want to describe to the class what you did?' 5. The teacher summarized important conclusions consistently, such as confirming the order of fractions after each exercise: 'So, now we can see that the correct order three-tenths is the least greatest then it is three-fourths and the greatest is nine-tenths.' 6. The teacher connected a series of ideas to the learning goal by revisiting different strategies and ensuring students understood the rationale behind each: 'Now to finish up this discussion, I'm going to give you guys another set of fractions and whatever strategy that you want you can decide what one is the greatest and what one's the smallest, okay?' Overall, the discussion was coherent and well-structured, effectively meeting the learning goal.

Table 11: Human evaluation and STRUCTURED COT + QUOTES (GPT) output. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations. *The human rater annotated "So today students, ..." as the evidence for introduction in a separate document.

Human	The teacher provides hardly any introduction at the beginning, but then provides a better introduction at 0:24 after “testing the waters.”$*^1$ The teacher focuses on addressing Mina’s misconception, but does so with what I think are area models that she drew before the interaction rather than using the number lines, which could do the same thing. The teacher does try to make it clear that the number line strategy is not always a good option, but most of the ideas are her own.
GPT	1. The teacher did not state a clear, detailed learning goal and expectation at the beginning. The teacher mentioned discussing strategies but did not clearly outline the learning goal: ‘we’re going to talk about what strategies are helpful when we are comparing fractions, and what strategies might be better used on a different kind of problem.’ 2. The teacher moved between ideas timely with apparent rationale, focusing on different students’ work and strategies: ‘So we’re going to start our discussion by looking at Will and Jayla’s work with their ordering fractions.’ 3. The teacher instructed with deliberation, asking students to explain their thought processes and addressing misconceptions: ‘If you could just explain to the class, your thought process for putting together the strategy, I think that might be really helpful to the other students in the class.’ 4. The teacher introduced well-chosen new ideas that benefited student understanding at relevant moments, such as using visual aids: ‘I actually drew a picture that maybe we could look at together.’$*^2$ 5. The teacher summarized important conclusions consistently throughout the discussion, reinforcing key points: ‘So when we look at this picture, can we see which one’s biggest?’ 6. The teacher connected a series of ideas to the learning goal by discussing different strategies and their effectiveness: ‘All of the strategies we’re going to talk about today do work at some point, but we want to talk about the benefits and maybe some of the things that are a little bit harder about those strategies as well.’

Table 12: Human evaluation and STRUCTURED COT + QUOTES (GPT) output. The green and red texts represent agreement and disagreement between the human and STRUCTURED COT + QUOTES, respectively. The blue texts are quotes from the discussion appearing in both evaluations. $*^1$at 0:24, the teacher said “we’re going to talk ... ” $*^2$the human rater annotated “I actually drew ...” as the evidence for the introduction of new ideas in a separate document.