Towards evaluating teacher discourse without task-specific fine-tuning data

Beata Beigman Klebanov, Michael Suhan, Jamie N. Mikeska
ETS Research Institute
bbeigmanklebanov,msuhan,jmikeska@ets.org

Abstract

Teaching quality is one of the determinants of student outcomes. Teaching simulations with feedback are one way to provide teachers with practice opportunities to help improve their skill. We investigated methods to build evaluation models of teacher performance in leading a discussion in a simulated classroom with the goal of providing feedback, particularly for tasks with little performance data.

1 Introduction

Teaching quality is one of the determinants of student outcomes (Blömeke et al., 2016; Fauth et al., 2019). The theory of practice-based teacher education (Ball and Cohen, 1999) argues that teachers need opportunities to practice core teaching skills, such as engaging students in the disciplinary discourse practices and leading classroom discussions, in situations of reduced complexity (Grossman, 2021; Forzani, 2014). For such practice opportunities to be successful and impactful, they need to be flexible, target specific difficulties, and provide learning support, in the form of timely feedback (McDonald et al., 2013; Mikeska et al., 2024).

Simulated classrooms are one environment providing such opportunities. They allow for strategic reduction in task complexity so that aspects of teaching can be isolated and practiced separately. Simulations are used in teacher education in various forms, including peers role-playing students (Davis et al., 2017; Masters, 2020), mixed-reality simulations where trained actors play the students (Bondie et al., 2021; Dieker et al., 2019), as well as emerging work where AI agents role-play students to help train teachers or tutors (Lim et al., 2025; Pan et al., 2025; Markel et al., 2023). Across all these forms, feedback to the teacher on their performance that would point out strengths and areas for growth in a constructive and actionable manner is critical (Cohen et al., 2020; Mikeska et al., 2023a).

Until recently, a bottleneck for creating automated feedback was acquiring a substantial amount of data of learners performing the simulation. Such data, with human scores, enabled the creation of machine learning based evaluation models to power automated feedback. With the advent of few-shot learning with large language models, there is an opportunity to mitigate the bottleneck, since only a handful of examples might suffice for these models to be able to evaluate a new learner's performance.

The goal of this paper is to start exploring this opportunity through two research questions: (**RQ1**) How accurately can a few-shot LLM evaluate a teacher's performance in a simulation? (**RQ2**) How do these results compare to an alternative method – a *generic* model fine-tuned on data from other tasks and used to evaluate performance in a new task? The latter approach has been successful in large-scale essay scoring, where a model trained on essays responding to a variety of essay prompts is used to evaluate essays from new, unseen prompts (Ramineni and Williamson, 2013). However, results might depend on data representation. The relatively content-agnostic essay scoring features may have been responsible for the success; recent reports suggest that transformer-based fine-tuned models do not generalize well across prompts (Shermis, 2024). To our knowledge, this is the first exploration of few-shot vs. generic fine-tuned models for evaluating teacher discourse in the absence of fine-tuning data for a task-specific model.

2 Related work

2.1 Digital teaching simulations with feedback as a learning opportunity for teachers

Using digital teaching simulations within teacher education and professional development programs can improve teachers' instructional skills, beliefs, and knowledge (Francis et al., 2018; Lee et al., 2024). Simulations are typically integrated into

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 192–200

these learning environments via cycles where teachers prepare for, engage in, and reflect on their performance, as well as receive formative feedback on how well they have enacted specific aspects of teaching (Mikeska et al., 2023b; Pecore et al., 2023). Recent research has shown that timely, personalized feedback is important to propel teachers' learning from digital teaching simulations (Cohen et al., 2020; Garrett et al., 2020; Mikeska et al., 2023a). Yet, such feedback is hard to scale, as generating it relies on extensive human resources.

2.2 Automated evaluation of teacher discourse

Recent work on automated evaluation of classroom discourse using pre-trained LLMs has explored fine-tuning (Kupor et al., 2023; Nazaretsky et al., 2023; Xu et al., 2024; Ilagan et al., 2024), zero/few-shot learning (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024; Hou et al., 2024; Asano et al., 2025), or both (Chen, 2023; Tran et al., 2024). Xu et al. (2024) noted that results are better for aspects of teaching that require less pedagogical expertise. None of these studies systematically investigated generalization across content domains, topics of discussion, or other aspects of classroom discourse.

3 Feedback in a teaching simulation

The context of this paper is ongoing work on developing new tasks for digital teaching simulations focused on the core teaching competency of leading a math or science argumentation-based discussion in an elementary classroom. After engaging in a simulation, teachers receive an automated feedback report. The report was designed by teacher education researchers and professionals to cover indicators of teaching quality (Mikeska et al., 2024). For each indicator, the report provides a comparison to a typical high-quality discussion and shows utterances where the target behavior did and did not occur; see Figure 1. The high-quality discussions are those that received a high score on a holistic rubric such as shown in Table 1. The comparison to high-quality discussions shows whether the teacher has engaged in the target behavior often enough.

4 Method

4.1 Data

We use data collected in multiple studies where a simulation was used as part of pre-service teacher coursework (Mikeska et al., 2023b, 2022). Before the simulation, the teacher is shown the prior work of two or three groups of simulated students; each group is designed to have a specific knowledge profile. For example, in task S1 students need to identify the mystery powder – one of six known powders – and the properties needed for the identification. The groups differ in what they think the powder is and what properties are needed for identification, as they explain in their prior work. Teachers are given up to a week to prepare to lead a 20-min discussion with a specific learning goal. In S1, the goal is to have the class reach consensus on the mystery powder and the necessary properties to identify it. In another task (M1), the goal is reaching consensus on methods for ordering fractions with different numerators and denominators.

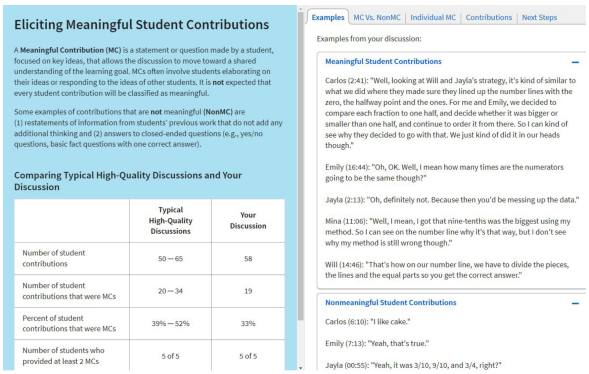

Figure 1: A screenshot of a part of a feedback report for indicator 1b – elicitng meaningful student contributions.

1	[Beginning] The teacher does not make any use of student ideas during the lesson.
2	[Developing] The teacher makes use of some student ideas in a limited way. This can mean: A missed opportunity to move the lesson forward; Only occasiona use of students' ideas when there were multiple opportunities throughout the lesson; An attempt was made to use the students'ideas, but the teacher did not do so in a way that moved the lesson forward.
3	[Well-prepared] The teacher makes use of student ideas to move the lesson forward.

Table 1: Holistic scoring rubric, indicator 1c.

Each discussion transcript was scored using a multi-dimensional rubric (Mikeska et al., 2021). Dimension 1 focused on the teacher's skill in attending to students' ideas equitably. Dimension 1 covered three indicators: how well the teacher (1a) incorporated all the key ideas that appear in the students' prior work into the discussion; (1b) elicited meaningful contributions from all students; and (1c) attended to and made use of each of the relevant student ideas shared during the discussion. We focus on 1b and 1c, for which raters scored teacher

performance on the scale of 1–3 or 1–4 (allowing scores like 1.3 or 2.7) and provided justifications by selecting one or more specific teacher (1c) or student (1b) utterances in the transcript where the target behavior clearly did (label 1) or did not (label 0) occur. These justifications form the bulk of the utterance-level annotations used in this study; some additional selections were made by research staff. Table 6 in the Appendix shows a snippet of a discussion, with justifications. We collapsed the top two levels on the 1-4 scale into score 3, as the fourth level was originally added to separate out the strongest performances. Table 1 shows the rubric for indicator 1c. The inter-rater reliability was estimated on double-scored data from task S1: r = 0.52 for 1b and r = 0.53 for 1c, indicating moderate agreement (Dancey and Reidy, 2007).

We use data from six discussion tasks, two in science (S1, S2) and four in math (M1 through M4). Table 2 shows short descriptions. For three tasks – S1, M1, M2 – we have fine-tuning data. For S2, M3, and M4 we have only data to evaluate models; these three tasks will serve as the new tasks to answer RQ2. Table 3 describes the test sets. Data used to develop few-shot models and to fine-tune the BERT model will be described with the models.

4.2 Models

4.2.1 Few-shot models

We implemented the model setup found to be the best for assessing various aspects of classroom discourse in the literature: Tran et al. (2024) investigated task formulations, zero vs few-shot, random or selected examples, and length of context for scoring models implemented using Mistral and Lacuna LLMs. Across multiple constructs and both LLMs, the authors found that an utterance-level classifier with ten few-shot examples (4 positive, 3 negative, and 3 hard negative) and with four prior utterances as context resulted in the best prediction of human holistic scores when aggregated into transcript-level scores. After sampling the test data described in Table 3, we sampled 3 transcripts per task to serve as sources of the ten examples for the tasks S2, M3, and M4 for which we had little data available. For tasks S1, M1, and M2, we sampled one transcript at a time from a larger set until the target 10 examples were identified. It took 8, 6, and 5 transcripts for the three tasks, respectively. Teacher education researchers and assessment developers selected the examples.

We use a state-of-art LLM, **GPT-4o**, and an open source smaller model, DeepSeek-R1:7b (**DS-R1**), distributed through Ollama 0.5.1,[1] both with temperature of 1.0 and default settings for all hyperparameters. Prediction is done on each utterance three times; majority vote decides the final label.

The prompt is a template into which task-specific information is put when the model is used to evaluate data from that task. The template elements are the domain (math or science), task information (e.g., identifying the mystery powder), learning goal (see Section 4.1), and few-shot examples. Below is the template of the GPT-4o prompt for 1c, with few-shot examples appended as user and assistant turns in the messages array sent to the model:

Instructions

Answer yes or no to the following question:

Given the dialogue between a teacher and students in a *{domain}* classroom about *{task_info}*, in the last turn, did the teacher attempt to make use of students' ideas to move the discussion towards the learning goal?

Learning Goal

{learning_goal}

Student names

Jayla, Will, Emily, Mina, Carlos

Output structure

Output must be one of the following words:

yes

no

To take advantage of DeepSeek-R1's "thinking", the examples are included in the system prompt, and the instructions for output structure state that the answer should be on the last line of the output.

4.2.2 Fine-tuned models

We use the utterance-level binary classifiers for indicators 1b and 1c originally developed for the S1 task by Nazaretsky et al. (2023). For indicator 1c, the teacher's utterance to be classified is represented as an embeddings vector and enriched by the embeddings vector of the preceding student utterance as context. For indicator 1b, we are classifying the students' utterances as providing or not providing a meaningful contribution, as evidence for the teacher's success in eliciting such contributions. Therefore, for indicator 1b, the target utterances are students' and the context is the preceding teacher or another student's utterance. The models use Hugging Face DistilBERT[2] base model (un-

[1] https://ollama.com

[2] https://huggingface.co/docs/transformers/model_doc/distilbert

M1	The teacher leads a discussion of three student-generated strategies for ordering the given fractions from least to greatest.
M2	The teacher leads a discussion with the students about an unconventional student-generated method for generating fractions between two given fractions. The discussion is focused on the strengths and weaknesses of the strategy, and its applicability to other situations.
M3	This discussion is grounded in students' work on a story problem in which they have used fraction multiplication. Prior to the discussion, the students individually critiqued one another's work, making the critique aspect of argumentation more clearly available to the teacher.
M4	This discussion focuses on students' work to generate meaningful understandings and representations of division by a fraction.
S1	The discussion focuses on reaching group consensus on the identity of an unknown powder based on its properties and what is known about a set of common powders. In addition to identifying the mystery powder, students discuss which properties are most useful and why.
S2	In this task, the teacher supports the students in discussion whether the amount of matter is conserved during a physical change, in this case, the mixing of ingredients to produce lemonade.

Table 2: Task descriptions.

Task	M1	S1	M2	S2	M3	M4
#transcripts	40	34	40	40	37	35
			Indicator 1c:			
#utts. (K)	1.8	1.5	1.5	1.8	1.8	1.5
#labeled utts. (K)	.39	.29	.46	.63	.49	.89
Average score	2.3	2.4	2.4	2.5	2.4	2.4
Std of scores	.60	.58	.60	.51	.50	.58
%1 in labeled utts.	52	67	71	27	60	69
			Indicator 1b:			
#utts. (K)	2.1	1.9	1.7	2.1	2.1	1.9
#labeled utts. (K)	.60	.43	.62	.77	.74	.65
Average score	2.3	2.4	2.4	2.5	2.4	2.4
Std of scores	.65	.57	.65	.45	.57	.68
%1 in labeled utts.	36	59	72	53	60	45

Table 3: Description of the test data. For each indicator, we show the number of teacher (1c) or student (1b) utterances, overall and labeled. Next are average and std of the holistic transcript scores. The last rows show the percentage of labeled utterances that have the label 1 (where the target behavior occurs).

cased, 66M parameters) (Sanh et al., 2019) with PyTorch 2.2.2 (Paszke et al., 2019). Details of the fine-tuning process and licensing information can be found in the Appendix (see Technical Details). We fine-tuned the classifiers on data from 120 transcripts – 40 from each of M1, S1, and M2 – that were sampled after the test data was partitioned out.

4.3 Model Evaluation

We evaluate the models in two ways. First, we use the utterance level data classified as 1 or 0 using rater justifications to evaluate the accuracy of utterance-level predictions. We think about these as "easy" data, in the sense used in the NLP and machine learning literature to refer to clear-cut, uncontroversial cases – as opposed to "hard" cases where human annotators disagree or where the correct label might be genuinely unclear (Leonardelli et al., 2021; Loukina et al., 2018; Jamison and Gurevych, 2015; Beigman Klebanov and Beigman, 2014). Having just "easy" evaluation data does not allow for a comprehensive evaluation of utterance-level predictions, but being able to classify the easy cases correctly can serve as a first-cut test, as a model that can't handle the easy cases would have low face validity. This evaluation most directly supports the feedback component where example class 1 and class 0 utterances are shown (see Figure 1).

The second evaluation is at the level of the transcript, where we derive a holistic score from utterance-level predictions (number of utterances classified as 1) and compare it to the holistic score provided by human raters. In the easy-vs-hard framework, this evaluation includes both easy and hard instances, since predictions are made on all utterances, some of which are expected to be harder than others. In the NLP literature, there is evidence that for some tasks, training a model on only the easy data results in better performance on not just the easy test cases but on the hard ones, too (Jamison and Gurevych, 2015), presumably because the system isn't getting confused by training examples that could be ambiguous or controversial. The transcript-level evaluation supports the feedback where the overall statistics of the target behavior in the current teacher's discussion are compared to those in high-quality discussions (see Figure 1).

5 Results

5.1 Utterance-level

Table 4 shows the results for the utterance level classification. To answer RQ1: GPT-4o has average accuracy of 0.73 across two indicators × six tasks, range = 0.62–0.81, std = 0.064. DeepSeek-R1 is much less successful, with average accuracy of only 0.56, range = 0.38–0.69, std = 0.10.

To answer RQ2, we compare the performance

of GPT-4o on tasks S2, M3, and M4 to that of the BERT model fine-tuned on data from tasks M1, S1, and M2. The average accuracy of the BERT model on two indicators × three new tasks is 0.76, range = 0.54–0.91, std = 0.13. The average accuracy of the GPT-4o model on the same data is 0.72, range = 0.63–0.81, std = 0.07. While BERT has a higher average, it is more volatile, with a poor performance of 0.54 on indicator 1c for task S2.

Comparing GPT-4o and BERT on the three tasks on which BERT was fine-tuned – M1, S1, and M2 – we observe that BERT shows stronger performance, as expected. BERT's average accuracy on two indicators × three tasks is 0.81, range = 0.70–0.89, std = 0.08. GPT-4o's average accuracy on the same data is 0.74, range = 0.62–0.79, std = 0.06. These results indicate that it is worthwhile to fine-tune a model on available data for scoring new performances belonging to the tasks on which the model was fine-tuned. For utterance-level scoring of data from new tasks, one might want to go with GPT-4o, as its performance is comparable to BERT's on average but more stable across tasks.

	M1	S1	M2	S2	M3	M4
			Indicator 1c:			
BERT	**.73**	.70	**.87**	.54	**.91**	**.81**
DS-R1	.53	.52	.69	.38	.47	.46
GPT-4o	.62	**.75**	.79	**.63**	.73	.68
			Indicator 1b:			
BERT	**.84**	**.81**	**.89**	.67	**.81**	**.84**
DS-R1	.50	.65	.69	.57	.69	.58
GPT-4o	.77	.75	.78	**.69**	**.81**	.79

Table 4: Accuracy of classifying teacher utterances as making use of student ideas or not (Indicator c) and student utterances as providing a meaningful contribution or not (Indicator b), on labeled test data. Best performance per indicator per task is boldfaced. Gray background marks BERT performance on tasks on which the BERT model was fine-tuned.

5.2 Transcript-level

Table 5 shows the correlations between the human holistic indicator score and the number of teacher (1c) or student (1b) utterances that were classified as 1 (exhibiting the target behavior). To answer RQ1: GPT-4o averages r = 0.46 across the two indicators × the six tasks, range = 0.14–0.73, std = 0.18. DeepSeek-R1 averages r = 0.44 for the same data, range = .22–.64, std = 0.15. Thus, the two few-shot models show comparable moderate performance and substantial volatility across tasks.

To answer RQ2, we compare the few-shot models to BERT on the three new tasks – S2, M3, and M4. BERT performs at r = 0.39 on average across two indicators × three tasks, range = 0.19–0.57, std = 0.14. GPT-4o average performance on the same data is r = 0.32, range = 0.14-0.52, std = 0.13. DeepSeek-R1 averages r = 0.34, range = 0.22-0.51, std = 0.10. At the aggregate level of the full transcripts, the generic fine tuned model tends to show stronger performance than few-shot models, although all the models achieve only low-medium correlations with the human holistic scores and are quite volatile.

Across the three tasks on which the BERT model was fine-tuned (tasks M1, S2, M2), it outperforms the few-shot models: BERT averages r = 0.67, range = 0.55–0.79, std = 0.09. GPT-4o averages r = 0.59, range = 0.45–0.73, std = 0.09. DeepSeek-R1 averages r = 0.54, range = 0.33–0.64, std = 0.11. For the transcript level, the results suggest that the fine-tuned generic model is the model of choice.

	M1	S1	M2	S2	M3	M4
			Indicator 1c:			
BERT	**.74**	.59	**.55**	.32	.19	**.41**
DS-R1	.33	**.63**	.53	.22	**.27**	.32
GPT-4o	.45	.62	**.55**	**.35**	.14	.22
			Indicator 1b:			
BERT	**.79**	**.70**	**.64**	**.50**	.34	**.57**
DS-R1	.64	.59	.54	.32	**.37**	.51
GPT-4o	.73	.62	.59	.36	.31	.52

Table 5: Pearson's correlation between the human holistic indicator score and the number of teacher utterances automatically labeled exhibiting the target behavior. Best performance per indicator per task is boldfaced. Gray background marks BERT performance on tasks on which the BERT model was fine-tuned.

6 Discussion

Our results suggest that a fine-tuned generic model is worth creating if only to score the data from the tasks it was fine-tuned on, as it shows stronger performance than few-shot models in these cases, both at the uttereance and at the transcript level. However, for evaluating data from new tasks for which it is not feasible to fine-tune a model due to lack of data, the situation is less clear-cut. In particular, at the utterance level, the generic fine-tuned model shows more volatile performance across tasks than the few-shot one, failing to capture the "easy" dis-

tinctions between utterances in one of the six evaluated cases (accuracy = 0.54 on task S2).

It could be that the S2 data is particularly difficult to classify; however, since GPT-4o shows much stronger performance, it is likelier that the class distribution difference between the fine-tuning data and the S2 data is to blame for BERT's failure of generalization for S2. Indeed, Table 3 shows that S2 data had an exceptionally low proportion of 1s – only 27%. This occasional generalization failure of a generic fine-tuned model illustrates its weakness compared to a few-shot model, namely, its dependence on class distribution in the fine-tuning data, to which the few-shot models are more robust.

Another interesting finding is that the wide gap between GPT-4o and DeepSeek-R1 at the utterance level (average accuracy of 0.73 vs 0.56, respectively) is closed almost entirely at the transcript level (average correlations of 0.46 and 0.44, respectively). Thus, while DeepSeek-R1 has worse face validity, as it isn't able to consistently tell apart clear-cut cases of 0s and 1s, its aggregate performance is similar to GPT-4o's. To gain further insight into this result, we checked the proportion of utterances classified as 1. GPT-4o classified 69% of all teacher utterances as 1 (indicator 1c). The number is 57% for DeepSeek-R1. DeepSeek-R1 predicts many fewer 1s than GPT-4o, including mis-classifying more "easy" 1s as 0s: The ratio of the number of false negative predictions for the labeled utterance-level data for DeepSeek-4o to that of GPT-4o is 3.5:1, while the ratio of false positives is 1:1. Still, the two systems make similar relative judgments of which transcripts have more utterances with the target behavior. Indeed, GPT-4o's and DeepSeek-R1's transcript-level predictions correlate at $r = 0.72$ on average across tasks for indicator 1c – a much stronger correlation than either of them has with the human holistic scores.

Finally, we observe that transcript-level performance on M1, S1, and M2 is stronger than on S2, M3, and M4, for both GPT-4o and BERT. It is not the case at the utterance level, apart from S2. The models were able to classify the "easy" utterance-level labeled data, but that was not always sufficient to be able to classify all cases – easy and hard – in a reasonable way, that is, in alignment with the tendency expected based on the holistic score. For GPT-4o, limiting the number of transcripts to draw the ten few-shot examples from to 3 may have resulted in examples of lower quality – when we were not limited by dataset size, we went through 5-8 transcripts to find good examples for S1, M1, and M2. Going with fewer than 8-10 transcripts per new task may not be advisable. For BERT, it is apparently not enough to fine-tune on "easy" cases to handle not only the "easy" cases for the new tasks but the hard ones, too. In future work, we will explore automated detection of harder examples. This should help focus the utterance-level models on the easy ones (where the accuracy is high) for picking examples for feedback. Identifying harder cases in the unlabeled utterances from the 120 fine-tuning transcripts, labeling them, and adding to the fine-tuning data might help improve BERT's transcript-level performance on new tasks.

The current study has a number of limitations. First, we experimented with a limited range of models; it is possible that results would change with more effective prompts or different LLMs or more sophisticated data representation for fine-tuning. Second, we considered only two teaching quality indicators; while it is encouraging that the results are aligned between these two, further work is necessary to evaluate robustness of the findings.

7 Conclusion

We investigated two approaches for evaluating teacher performance in leading a discussion in a simulated classroom in the context where no data for fine-tuning on the specific discussion task is available. One approach uses a few-shot LLM. The other approach is a "generic" model fine-tuned on data from other discussion tasks. We found that the few shot model (GPT-4o) may be preferable for analyzing utterance-level data, due to its more stable performance across tasks, while the fine-tuned BERT model performed better in the aggregate, transcript-level evaluation. Our results thus point towards a way to capitalize both on few-shot learning and on previously collected data in order to supply the most effective learning opportunity – the one with timely automated feedback – even when little prior data is available for the current performance task.

References

Yuya Asano, Beata Beigman Klebanov, and Jamie Mikeska. 2025. Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o. In *Proceedings of the 20th Workshop*

on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 716–736, Vienna, Austria. Association for Computational Linguistics.

Deborah Ball and David Cohen. 1999. Developing practice, developing practitioners: Toward a practice-based theory of professional education. In *Teaching as the learning profession: Handbook of policy and practice*, pages 3–32. San Francisco: Jossey Bass.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396.

Sigrid Blömeke, Rolf Vegar Olsen, and Ute Suhl. 2016. Relation of student achievement to the quality of their teachers and instructional quality. *Teacher quality, instructional quality and student outcomes*, 2:21–50.

Rhonda Bondie, Zid Mancenido, and Chris Dede. 2021. Interaction principles for digital puppeteering to promote teacher learning. *Journal of research on technology in education*, 53(1):107–123.

Gaowei Chen. 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*.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational evaluation and policy analysis*, 42(2):208–231.

Christine Dancey and John Reidy. 2007. *Statistics without maths for psychology*. Pearson education.

Elizabeth A Davis, Matthew Kloser, Andrea Wells, Mark Windschitl, Janet Carlson, and John-Carlos Marino. 2017. Teaching the practice of leading sensemaking discussions in science: Science teacher educators using rehearsals. *Journal of Science Teacher Education*, 28(3):275–293.

Lisa A Dieker, Carrie Straub, Michael Hynes, Charles E Hughes, Caitlyn Bukathy, Taylor Bousfield, and Samantha Mrstik. 2019. Using virtual rehearsal in a simulator to impact the performance of science teachers. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 11(4):1–20.

Benjamin Fauth, Jasmin Decristan, Anna-Theresia Decker, Gerhard Büttner, Ilonca Hardy, Eckhard Klieme, and Mareike Kunter. 2019. The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and teacher education*, 86:102882.

Francesca M Forzani. 2014. Understanding "core practices" and "practice-based" teacher education: Learning from the past. *Journal of teacher education*, 65(4):357–368.

Anthony Tuf Francis, Mark Olson, Paul J Weinberg, and Amanda Stearns-Pfeiffer. 2018. Not just for novices: The programmatic impact of practice-based teacher education. *Action in Teacher Education*, 40(2):119–132.

Rachel Garrett, Toni Smith, Melinda Griffin, and Melissa Yisak. 2020. A randomized field study of a teacher professional development program using mixed-reality simulation to develop instructional practice. *Society for Research on Educational Effectiveness*.

Pam Grossman. 2021. *Teaching core practices in teacher education*. Harvard Education Press.

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 *International Conference on Artificial Intelligence in Education*, pages 60–74. Springer.

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 the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 182–198, Mexico City, Mexico. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 291–297.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.

Tammy D Lee, Carrie Lee, Mark Newton, Paul Vos, Jennifer Gallagher, Daniel Dickerson, and Camryn Regenthal. 2024. Peer to peer vs. virtual rehearsal simulation rehearsal contexts: Elementary teacher candidates' scientific discourse skills explored. *Journal of Science Teacher Education*, 35(1):63–84.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieun Lim, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, Haewon Jung, Yoonsun Jang,

Sanghyeok Lee, and Jewoong Moon. 2025. Development and implementation of a generative artificial intelligence-enhanced simulation to enhance problem-solving skills for pre-service teachers. *Computers & Education*, 232:105306.

Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 1–12.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Heidi Masters. 2020. Using teaching rehearsals to prepare preservice teachers for explanation-driven science instruction. *Journal of Science Teacher Education*, 31(4):414–434.

Morva McDonald, Elham Kazemi, and Sarah Schneider Kavanagh. 2013. Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of teacher education*, 64(5):378–386.

Jamie 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. *ETS Research Memorandum no. RM-21-06)*.

Jamie Mikeska, Jonathan Steinberg, Pamela Lottero-Perdue, Dante Cisterna, Devon Kinsey, and Heather Howell. 2023a. Using simulated classrooms to examine elementary teachers' perceptions about, attention to, and use of formative feedback to improve their ability to facilitate science discussions. *Contemporary Issues in Technology and Teacher Education*, 23(1):48–83.

Jamie N Mikeska, Heather Howell, and Devon Kinsey. 2023b. 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.

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 *Artificial Intelligence in Education*, pages 251–258, Cham. Springer Nature Switzerland.

Jamie N. Mikeska, Calli Shekell, Adam V. Maltese, Justin Reich, Meredith Thompson, Heather Howell, Pamela S. Lottero-Perdue, and Meredith Park Rogers. 2022. Exploring the potential of an online suite of practice-based activities for supporting preservice elementary teachers in learning how to facilitate argumentation-focused discussions in mathematics and science. In *Proceedings of Society for Information Technology Teacher Education International Conference 2022*.

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 *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 122–132.

Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

John L Pecore, Corey Nagle, Tadlee Welty, Minkyoung Kim, and Melissa Demetrikopoulos. 2023. Science teacher candidates' questioning and discussion skill performance in a virtual simulation using experiential deliberate practice. *Journal of Science Teacher Education*, 34(4):415–435.

Chaitanya Ramineni and David M Williamson. 2013. Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1):25–39.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mark D Shermis. 2024. Automated scoring for NAEP short-form constructed responses in reading. In *The Routledge International Handbook of Automated Essay Evaluation*, pages 117–140. Routledge.

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

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667.

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 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4375–4389.

A Appendix

A.1 Example of the data

T:	Mina and Will, why did you choose weight as an important property?
M:	Because it falls under some of the things that we can see and measure.
T:	Carlos, do you want to explain to them about why you thought that weight wasn't important?
C:	**Sure. Well, actually I don't think weight is really that important, because the weight of the object doesn't really change what the object is. If you were to add more powder, it would change the weight, but that doesn't change what the powder is.**
M:	I guess I see what you mean by that, but I still think that we found the correct thing.
T:	Jayla and Emily, do have any other points to make on the conversation of whether or not weight was important?
E:	Well, we did test the weight in ours because we thought that testing all the properties would be important, but now I'm starting to think about it. I guess weight doesn't really matter, since if we were to add more or take away some of the powder the weight would change, but it wouldn't change what the powder was, like Carlos was saying. So I get that now.
T:	Right. Will, Jayla, do you have any other points that you want to make?
W:	I guess I'm starting to see what Carlos means by that.
T:	Jayla?
J:	Oh yeah, I can see where he was coming from.

Table 6: A snippet of a Mystery Powder (S1) discussion. The blue-boldfaced teacher's utterance was marked by a human rater as an example of the teacher using a student's idea to move the discussion forward (indicator 1c), whereas the black-boldfaced utterance from Carlos was marked as an example of a meaningful student contribution (indicator 1b). T: Teacher. M: Mina. C: Carlos. J: Jayla. W: Will. E: Emily.

A.2 Technical Details of the Models

Fine-tuning. The BERT models were fine-tuned (including all transformer layers, the pooling layer, and the final dense output layer) with Adam optimizer (learning rate = 1e-5, learning warmup = 600) to minimize the binary cross-entropy loss. We used a grid search across 15 epochs with batch sizes 1, 4, and 8 for indicator 1c and across 10 epochs with batch sizes 1, 4 and 8 for indicator 1b. The parameters were tuned using 7-fold cross-validation on the fine-tuning data. For indicator 1c, we used 7 epochs with batch size 4. For indicator 1b, we used 3 epochs with batch size 4.

GPU hours. For indicator 1c, DistilBERT fine-tuning was run locally on a desktop PC with an NVIDIA GeForce RTX 3050 GPU and 16gb physical memory. All fine-tuning, including grid search for all models, took 5 hours and 14 minutes. For indicator 1b, DistilBERT finetuning was run locally on a MacBook Pro with an Apple M2 Pro chip (integrated GPU) and 16gb physical memory, and took 14 hours and 57 minutes.

For both indicators, DeepSeek-R1 test set predictions were run on the same PC, taking on average 97 minutes per transcript for indicator 1c and 66 minutes for indicator 1b. Predictions for fine-tuned models were run on the same MacBook Pro, taking on average 20 seconds per transcript for indicator 1c and 66 seconds per transcript for indicator 1b. GPT-4o predictions were generated using the Batch API via the Microsoft Azure OpenAI Service under our organization's subscription, which provides a 24-hour target turnaround for batch jobs.

Licensing of artifacts. The instance of GPT-4o used is a proprietary AI model accessible via Microsoft's Azure OpenAI Service, subject to Microsoft's licensing terms. Ollama and DeepSeek-R1 are licensed under the MIT License. DistilBERT is licensed under the Apache License, Version 2.0. PyTorch is licensed under the BSD-3-Clause.