

Towards evaluating teacher performance in a GenAI teaching simulation of a science discussion

Beata Beigman Klebanov¹, Jamie N. Mikeska¹, Mengxuan Zhao², Catherine Flynn³, Devon Fetrow¹, Shreyashi Halder¹, Rutuja Ubale¹, Tricia Maxwell¹, Michael Suhan¹

¹ETS, Princeton, NJ, USA, ²Nesto, Canada, ³Fidelity Investment, USA

Correspondence: bbeigmanklebanov@ets.org

Abstract

GenAI can power simulated student agents that provide opportunities for educators to engage in core teaching practices, such as leading a small group argumentation-based science discussion. To realize the potential of such simulations and support teacher reflection and learning, it is necessary to provide participants with timely feedback on their performance in the simulation. This study investigates systems for automated evaluation of and feedback on teacher performance in a simulation along the dimension of making use of student ideas to move the discussion forward. We address three research questions: (a) How well do models fine-tuned on transcripts of teacher performance in a matching human-puppeteered teaching simulation (that served as the model during the development of the GenAI one) perform in evaluating transcripts from the GenAI teaching simulation? (b) How well does a system using a few-shot LLM perform on the same task? (c) How do educators perceive the quality and usefulness of the automatically generated feedback? The findings underscore the importance of a rigorous evaluation of automated evaluation and feedback systems.

1 Introduction

Learning how to engage in core teaching practices, such as facilitating a classroom discussion, is critical for educators to be effective practitioners (McLeskey et al., 2019; O’Flaherty and Beal, 2018). Research has indicated that practice-based, professional learning opportunities can develop teachers’ ability to engage in core teaching practices (Davis and Palincsar, 2023; Hurlbut and Krutka, 2020). Digital teaching simulations (DTS) can serve as practice spaces where educators can try out aspects of specific teaching practices in online environments with student avatars or agents who can interact with educators in real-time (Ledger et al., 2022; Lim et al., 2025). Formative feedback on educators’

practice can be beneficial in this learning process (Kelleci Alkan et al., 2024; Nygren et al., 2025).

Up until recently, the work on DTS included a human operating the student avatars behind the scenes. Recent advances in GenAI are now powering the exploration of simulations where GenAI is playing the students. Ideally, the last twenty years of work on DTS, including effective scenarios, student knowledge profiles, and feedback models, would be utilized by reimplementing the simulations using GenAI students to support scalability and access. In this study, we explore just such a juncture in the context of models for evaluating teacher performance in a simulation. In this paper, we answer three research questions (RQs):

RQ1 How well do models for evaluating teacher performance in a simulation that were fine-tuned on data from a matching human-led DTS perform on GenAI simulation data?

RQ2 How well does a system using a few-shot LLM perform on the same task?

RQ3 How do educators perceive the quality and usefulness of the feedback on their performance in the GenAI simulation?

2 Related Work

2.1 Digital teaching simulations (DTS)

DTS have been growing in use to support educators in learning how to engage in core teaching practices including how to: communicate with parents, elicit student thinking, address multi-lingual students’ learning needs, facilitate discussions, and manage student behavior (Bondie et al., 2021; Ledger et al., 2025; Lindberg and Jönsson, 2023). One kind of DTS is a human-puppeteered one, where a human operates the student avatars behind the scenes with the teacher interacting verbally with student avatar(s) whose images are shown on a

computer screen. The student avatars can converse with the teacher and with each other. Research using human-led DTS has shown that these tools support improvements in teachers' skills, beliefs, and knowledge (Badiee and Kaufman, 2014; McGarr, 2021).

More recently, researchers have begun using GenAI to create chatbots that respond as K-12 students in order to provide safe practice spaces for teachers. Educational goals include enhancing teachers' data literacy (Lee and Lee, 2025), critical thinking (Drushlyak et al., 2025), problem-solving skills (Lim et al., 2025), noticing (Son, 2025), eliciting student thinking (Mikeska et al., 2025), responsive teaching (Zhang et al., 2025), or professional development (Tan et al., 2025).

2.2 Expert evaluation of classroom discourse

Research has suggested that teachers who engage in productive talk moves while facilitating discussions are more likely to support student engagement, sensemaking, and learning (Kazemi and Hintz, 2023; Michaels and O'Connor, 2015; Soysal and Soysal, 2022; Wei et al., 2018). Talk moves are varied prompts that teachers use to support productive talk among students, such as revoicing, pressing for reasoning or evidence, encouraging multiple perspectives, and others. Teachers can benefit from opportunities to learn about talk moves and how to employ them during instruction (Lefstein et al., 2020; O'Connor and Michaels, 2019). Such opportunities expand teachers' ideas about what productive talk moves are and let them try out these moves – ideally in peer rehearsals or DTS – before using them in their classrooms.

Facilitating productive discussions requires that teachers enact a complex array of instructional skills simultaneously. In previous work, researchers have decomposed the practice of facilitating discussions by identifying the key components that make up this practice. For example, TeachingWorks' decomposition of this practice involved noting several "discussion enabling" skills, such as anticipating student thinking and setting up the task, and "discussion leading" skills, such as launching a discussion, probing student contributions, and recording and representing content.¹ In a similar vein, Mikeska et al. (2019) identified five key dimensions of facilitating argumentation-focused discussions, such as attending to student thinking,

facilitating a coherent and connected discussion, and developing students' conceptual understanding. Other researchers have investigated how teachers engage in various aspects of facilitating discussions across content areas (Davis et al., 2017; Shaughnessy et al., 2019; Tyminski et al., 2014), further illustrating the key skills essential to this practice. Research has shown that timely, personalized feedback is important to propel teachers' learning from DTS (Cohen et al., 2020; Garrett et al., 2020). Yet, such feedback is hard to scale, as generating it relies on extensive human resources.

2.3 Automated evaluation of instructional quality

Substantial prior research on automated evaluation of instructional quality in classroom discussions focused on detecting various teacher or student talk moves, such as building on student responses (Bywater et al., 2019; Demszky et al., 2021; Nazaret-sky 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), encouraging peer interaction (Ilagan et al., 2024), and giving supportive statements (Hunkins et al., 2022).

In addition to the traditional supervised machine learning approaches, using pre-trained LLMs with fine-tuning (Chen, 2023; Kupor et al., 2023; Tran et al., 2024; Xu et al., 2024; Beigman Klebanov et al., 2026) and zero/few-shot learning (Asano et al., 2025; Chen, 2023; Hou et al., 2024; Tran et al., 2024; Wang and Demszky, 2023; Beigman Klebanov et al., 2025) have also been explored.

To the best of our knowledge, the question of generalization of scoring models between human-led and GenAI-led simulations for the same task has not yet been addressed. Such exploration is timely: Well-designed simulation tasks that were used extensively with the human-led DTS technology over the last two decades could provide a natural starting point for scaling up using GenAI to play the students. It is an open question how to best capitalize on the extensive prior research and data. In this study, we investigate how well models for automatically evaluating teacher performance in a simulation generalize from a human-led version to the GenAI version of the same simulation task.

¹<https://library.teachingworks.org/curriculum-resources/teaching-practices/leading-a-discussion/>

3 Methods

3.1 The Mystery Powder (MP) simulation

The Mystery Powder (MP) task was developed in a previous project to support teachers in learning how to facilitate argumentation-focused discussions (Mikeska et al., 2021). In this task, the teacher's goal is to facilitate a discussion with five elementary student avatars in Mursion's simulated classroom in which the students achieve the **learning goals** – they come to consensus about the identity of the MP and which properties are most important to identify it. Prior to facilitating the discussion, the teacher learns that the five students already worked together to collect data on five different properties (e.g., color, texture, reaction with vinegar) of six known powders (e.g., sugar, baking soda). Each student also worked individually or in pairs to gather data on one MP, which was one of the six known powders, and used the data to generate a written claim and evidence-based reasoning about the MP's identity. The five student avatars begin the discussion with different claims about what the MP is and which properties are useful to identify it. The students can modify their ideas during the discussion in response to persuasive arguments.

The GenAI version of the simulation (Mikeska et al., 2026) was based on the human-led one (Mikeska et al., 2021). Student personas and knowledge profiles for three of the five students were used in the GenAI simulation. A knowledge profile specifies what the student does and does not understand about the content topic, what their claims are regarding the identity of the MP and the properties necessary to identify it, what properties they checked to arrive at their claim, and what needs to be said during the discussion for them to change their minds. Personas included traits such as being reserved or outgoing, which impacted their willingness to volunteer and how elaborate their responses were.

The GenAI simulation showed good usability, with over 70% of early pilot participants indicating that the preparation materials provided clear information about instructional goals, the participants understood what to do, and the simulation platform was easy to navigate (Mikeska et al., 2026). Most participants agreed that the students' contributions were on-topic and grade-level appropriate; however, they pointed out that student responses sometimes felt mechanical and not realistic (Mikeska et al., 2026). A key finding was that participants

varied in the extent of their use of students' ideas to advance the discussion, indicating the important role to be played by formative feedback, especially for those who rarely used students' ideas.

3.2 Feedback report

An evaluation model based on a few-shot LLM was used to classify each teacher utterance into making use of student ideas to move the discussion towards the learning goals (class 1) or not (class 0). The model will be described in section 3.4. The following five feedback components were shown in separate tabs, based on Mikeska et al. (2024). The feedback report was available immediately following the simulation.²

Overview tab presented a brief overview of what it means to productively use students' ideas to move the discussion towards the learning goals and showed a table comparing the count and frequency of teacher utterances with effective use of students' ideas in the discussion facilitated by the current teacher to a range corresponding to typical high-quality discussions. The latter were selected from human-led simulation data that was holistically scored on this dimension in prior work.

Examples tab showed up to three examples of teacher utterances where the teacher productively used students' ideas and three examples where they did not. This was followed by three positive and three negative examples based on a high-quality discussion.

Using Student Ideas tab showed a pie chart with the number and proportion of teacher turns where the teacher effectively used students' ideas in the discussion facilitated by the current teacher and in a high-quality discussion.

Flow of Discussion tab showed a timeline graph where each dot represented a teacher turn where student ideas were used effectively. Large gaps in the timeline could correspond to stretches of the discussion where the teacher drew on their own ideas, not the student's.

Next Steps tab showed strategies, prompts, and follow-up questions that can be used to increase productive use of student ideas.

²An example report can be viewed in full at: <https://figshare.com/s/fd716fcd57bcbd14f34d>

3.3 Data

3.3.1 Participants

We used data from two data collections, DC1 and DC2, both approved by our institution's IRB. Data from DC1 was used for annotator training and for selecting examples for the few-shot scoring model. Data from DC2 was used to evaluate the automated models for assessing teacher performance that powered the feedback reports.

DC1. 13 employees from an educational non-profit participated, aged 35 to 64 years. They identified as White (85%), Asian or Asian American (8%) and Black or African American (8%). 62% were female, 38% were male. 85% had worked as teachers, 23% had taught elementary science.

DC2: The 43 participants were: paraeducators (28%), in-service teachers (51%), and principals/administrators (21%). Ages ranged from 25 to 64 years. Participants identified as White (51%), American Indian or Alaskan Native (9%), Asian or Asian American (9%), Black or African American (33%), Hispanic or Latino (18%), and Native Hawaiian or Pacific Islander (5%). 63% were female, 35% were male and 2% were non-binary. 70% taught elementary science.

3.3.2 Data collection procedure

DC1. A study opportunity was shared with employees in an educational non-profit. Those interested completed the intake and consent forms. The MP simulation links, along with the background questionnaire and post-simulation survey, were sent to participants. One simulation link used GPT-4o while the other used Llama; both were voice-activated conversations. Participants were encouraged to try both models. After the simulation, participants completed the background questionnaire with demographic, educational, and professional information and the post-simulation survey, which gathered information about participants' experience with the simulation. We collected the 17 simulation transcripts used in the current study from 13 participants.

DC2. Participants were recruited through professional networks. Those interested completed an intake survey to check eligibility (current elementary paraeducator or an elementary in-service teacher or principal who had experience working with paraeducators) and signed a consent form. The link to the MP simulation with GPT-4o students was sent to participants along with a background

questionnaire and the post-simulation survey. In the same email, participants also received a disclaimer that feedback is work in progress and may not be fully accurate.³ Participants engaged with a voice-activated MP simulation and, upon completion, received a feedback report. The conversations were transcribed automatically. In this study, we used the transcripts, the feedback reports, and the post-simulation survey responses about the feedback evaluation.

3.3.3 Human annotation

Five of the authors of this paper, all experienced with content analysis, participated in the annotation. Rater training was done using DC1 data. The rubric provided detailed guidelines about how to annotate each teacher utterance as either using students' ideas to move the discussion towards the learning goals (1) or not (0).⁴ For establishing inter-annotator agreement (IAA) and for providing data for evaluating the automated scoring models, 10 transcripts were sampled randomly from DC2 data. Each of the teacher's turns (utterances) in each of these transcripts, for a total of $n = 381$ utterances, was annotated according to the rubric. Average IAA across 10 pairs of annotators was Cohen's $\kappa = 0.60$, range 0.46–0.78. Table 1 shows the statistics. We will provide annotated examples in section 3.4.

3.3.4 Feedback evaluation questionnaire

We used a four-point Likert scale (very useful, useful, somewhat useful, and not useful) to gauge participants' perceptions of the usefulness of the feedback report in supporting paraeducators' professional learning. We used a five-point Likert scale (strongly agree, agree, neutral, disagree, strongly disagree) to gauge perceptions about the clarity of the information provided in each feedback component and about its alignment with their perception of their performance in the simulation. We included open-ended questions asking participants for suggestions about improving the feedback report. Two project members engaged in qualitative content analysis of the responses to categorize participants' suggestions for improvement.

³Please note that some of the transcriptions in the feedback report may not be 100% accurate as this is still in the early stages of development. We will be refining that as time progresses."

⁴The full rubric can be accessed at: <https://figshare.com/s/b0290905abeb3813ef1f>

Table 1: Descriptive statistics of the annotated dataset (10 transcripts).

Descriptive statistic	Mean	Std	Min	Max	Median
# teacher turns (utterances)	38.1	33.8	12	130	29.5
# teacher turns annotated as 1	4.5	5.0	0	14	3.6
# words per teacher turn	12.3	14.5	1	100	8.0

3.4 Models

3.4.1 Fine-tuned models

We use the utterance-level binary classifiers originally developed for the MP task by Nazaretsky et al. (2023). A 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. The model uses the Hugging Face DistilBERT base model (Sanh et al., 2019). A prediction for an utterance is generated by taking the average probability for class 1 assigned to the given utterance by seven classifiers fine-tuned on different subsets of the human simulation MP data. If the probability is above 0.5, class 1 is predicted; 0 is predicted otherwise.

3.4.2 Few-shot model

GPT-4o (API version 2024-12-01-preview) was instructed to judge the last utterance in a discussion snippet as making use of student ideas to move the discussion towards the learning goals (yes) or not (no). A snippet contained the four utterances preceding the target teacher utterance. Making use of student ideas requires that teachers not only invite students to share their ideas but that they use the ideas shared as a springboard for further discussion. Making use of student ideas can involve teachers prompting students to consider each other’s ideas, whether they agree or not, and why; helping students connect their ideas to the learning goals; summarizing students’ key points and prompting them to reflect on the similarities and differences across their ideas. Table 2 shows four of the eleven few-shot examples that were chosen from DC1 data. The model was run three times; the utterance was classified as yes if all runs predicted yes.

3.4.3 Model evaluation metrics

We compared the models’ predictions against the gold standard that is the majority vote of the five human annotators. We use the following metrics: Accuracy; Cohen’s κ ; Precision, Recall, and F1 score for retrieving class 1 instances (Prec.-1, Recall-1, F1-1) and for class 0 (Prec.-0, Recall-0, F1-0).

4 Results

4.1 Classification performance (RQ1 & RQ2)

Table 3 shows the results for the original fine-tuned model (**FT original**) where classification was based on the average probability across seven models, for the version where all seven runs had to agree on class 1 (**FT unanimous**), and for the **Few-shot** model. Few-shot outperformed FT original on all metrics. FT unanimous sacrificed Recall-1, but still had low Precision-1. All models have very low Precision-1: When they declare an utterance as class 1, they are correct in only one fifth (FT original) to one third (Few-shot) of the time.

4.2 Feedback evaluation by users (RQ3)

More than 80% indicated that the report overall was useful or very useful for helping both early career and experienced paraeducators identify and notice their strengths and areas for growth in using students’ ideas. Table 4 shows results for feedback components. At least 80% agreed with each statement. Results for the Next Steps tab were similar. Two participants were unable to access the reports, so $n = 41$. To improve the reports, participants suggested including more detail and clearer explanations of results and examples (27%) and removing redundant and inaccurate examples (5%).

5 Discussion

5.1 Model generalization

The model fine-tuned on human-led MP data did not generalize well to GenAI-led data. One possible reason is the difference in data quality. In human-led simulation data, the transcripts were provided by a transcription agency, whereas those of the GenAI simulation were produced automatically, without human involvement. Indeed, annotators commented that it was often difficult to make out what the participant said; the annotation protocol prescribed class 0 in such cases. One of the five annotators systematically marked technical errors; she thus marked 27% of teacher utterances. Out of these, the fine-tuned model mis-classified 13% as

Table 2: Examples of input snippets with classifications of the target teacher utterance as making use of students’ ideas to move the discussion towards the learning goals. Teacher utterances are shown as transcribed by ASR.

Example 1 was classified as yes: Carlos: I think the mystery powder’s texture was smooth and its color was white. Teacher: OK, so Carlos thinks that the mystery powder is baking soda. Jayla and Mina, do you agree or disagree with him? Jayla: I agree, I think it’s baking soda. Mina: I disagree, I think it’s flour. Target teacher utterance: Mina, what do you think about Carlos’s idea? To check how the mystery powder looks in water and reacts with vinegar?
Example 2 was classified as yes: Teacher: Even if it’s 22g of it. Mina: Yes, even if it’s 22 grams, it’s still baking soda because it bubbled with vinegar! Teacher: So The weight doesn’t match. But you still think it’s baking soda. Is DO You Still think it’s useful to check the weight? Mina: Weight isn’t useful for identifying the mystery powder because the reaction with vinegar is more important for figuring it out. Target teacher utterance: Why is the reaction with vinegar more important than weight, Mina.
Example 3 was classified as no: Teacher: Mina What are the texture color and weight for flour. Mina: Sorry I didn’t understand, could you please repeat that? Teacher: What are the texture, color and weight for flour. Mina: Flour is smooth, white, and weights 24 grams. Target teacher utterance: Very good. But are there any other power powders that have the same texture, color, and weight.
Example 4 was classified as no: Teacher: That is an interesting point. The mystery powder is 24g. And we know from our classes’s chart that quite a few of the, the substances were around 24g. Carlos or Jayla. How did you consider the weight. Of the mystery powder. Carlos: The mystery powder is baking soda. Jayla: I think it’s baking soda. Mina: But I think it’s flour. Target teacher utterance: All right. We’re trying to come to an understanding where we all think it’s the same thing. So we’ve talked about the mystery powder’s reaction with water. And Mina, you had mentioned that you did not test the mystery powder with vinegar. If you had to do the experiment over again. Do you think you would mix the mystery powder with vinegar?

Table 3: Models’ classification performance on test set of n = 381 teacher utterances.

Model	Accuracy	κ	Prec.-1	Recall-1	F1-1	Prec.-0	Recall-0	F1-0
FT original	0.74	0.19	0.21	0.56	0.31	0.94	0.76	0.84
FT unanimous	0.85	0.25	0.31	0.36	0.33	0.93	0.91	0.92
Few-shot	0.83	0.42	0.36	0.87	0.51	0.98	0.82	0.89

class 1, including: “More details on why is it that is, for example, based on Carlos and Jayla” and “What do you think about the two strategies that empower.” Some of the language may be typical of class 1 utterances, misleading the fine-tuned model.

The few-shot model was misled in 5% of the cases.

Another potential reason for non-generalization is that teachers speak differently in AI-led vs human-led simulations. For instance, the annotated transcripts averaged 12.3 words per teacher utter-

Table 4: Participants’ perceptions of the quality of components (tabs) of the feedback report. Tab names (shown in the names of the columns) were inserted in _ slots in the questions. We combined strongly agree and agree (shown in the table). The remaining responses were neutral, disagree, or strongly disagree. $n = 41$.

Survey Question	%participants who agreed			
	Overview	Examples	Using Student Ideas	Flow of Discussion
I understood the description at the top of the _ tab.	97		97	93
I was able to compare my ability to use student ideas to a teacher facilitating a high-quality discussion using the information provided in the _ tab.	93	87	87	85
The information provided in the _ tab aligned with how well I used student ideas during my discussion.	81	85	81	90
I understood the examples from a high-quality discussion.	95			
I understood the examples from my discussion.	86			

ance (Table 1). In the human-led MP simulations, the number was 25.1. Some of this gap may be due to garbled speech recognition, but some could be due to a different experience conversing with AI students, who sometimes appeared “stilted and tedious” (section 3.1).

5.2 Human subjectivity

We observed low class 1 precision – a lot of false positives. Are these completely off-target, or difficult or borderline cases? We used the judgements of five annotators to explore this question. We created five versions of the gold standard. In the most permissive one, an utterance is classified as 1 if *any* of the annotators classified it as 1. In the second, two out of five votes suffice. The third is the majority vote we have used so far. The fourth requires a four out of five majority, the fifth – unanimous agreement. Figure 1 shows the results.

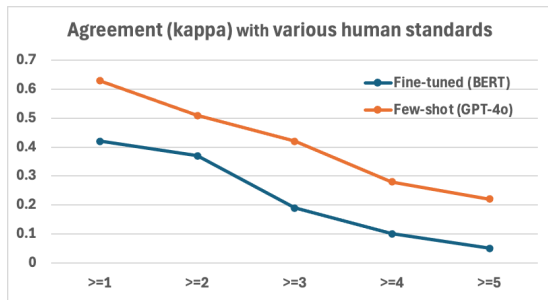


Figure 1: Model performance against increasingly conservative human gold standard.

Performance degradation for the two most conservative versions is to be expected, since the gold class 1 instances are subsets of those for the majority gold standard. The pattern to the left of the majority vote is the more interesting one. For the few-shot model especially, many of the class 1 predictions agree with the minority of annotators,

reaching $\kappa = 0.63$ with the most permissive version (the leftmost orange data point in Figure 1).

To further clarify whether the few-shot model captures the behavior of the annotator who gave the most class 1 classifications, we computed the model’s κ against that annotator. The κ was 0.21, much below the 0.42 for the majority and the lowest among the five annotators. The few-shot model is not modeling the most permissive annotator but minority judgments across different annotators.

5.3 User acceptance of inaccurate results

The 0.36 class 1 precision of the few-shot model is quite low: Only about one in three positive examples from the current teacher’s discussion was a true positive. Yet participants overwhelmingly said that they understood the examples, were able to compare their own discussion with a high-quality one using the information provided, and that the information was aligned with their own perception of their performance. In fact, only 2 out of 41 respondents mentioned removing redundant or inaccurate examples to improve the feedback, suggesting that they noticed problems with the examples. It could be that the rest did not examine the examples in detail or focused on the curated examples from high-quality discussions. It could also be a result of a social desirability bias – participants telling us what they think we want to hear, which is that the feedback was clear, useful, and accurate.

The discussion in section 5.2 suggests another possibility. Some of the examples, while incorrect, may not be obviously so. Those who think them correct may be in agreement with a minority of trained raters. This raises the hazard of leading the user to a misinterpretation of the practiced construct. Remediations include feedback that shows only curated examples from other users, alerting the users to be critical towards the examples, or

providing this information to the teacher educator who can mediate the interpretative process.

5.4 Limitations

Most participants were White and female; we will strive to recruit more inclusively. Most participants had teaching experience; the behavior of pre-service teachers in the simulation is yet to be explored. Sample sizes were small; the findings need to be confirmed with a larger sample.

6 Conclusion

With the advent of GenAI, scaling promising interventions is becoming feasible. An attempt to do just that is the context of this study: A GenAI simulation was developed based on the human-led digital teaching simulation. We found that the models for evaluating the teacher's performance in the GenAI simulation fine-tuned on data from the human-led simulations did not perform well. A few-shot LLM-based evaluation model performed better, although it still had low precision. Notwithstanding, study participants judged feedback based on its predictions as useful and accurate. We found that the model often agreed with a minority opinion among the five human raters, which may help explain why users found the feedback plausible.

Our results underscore the need for a thorough evaluation of models used to create formative feedback. First, even if the AI simulation is built to mimic the human-led one, the resulting data might be different enough for evaluation models to not generalize well. Second, even if trainees accept the feedback as useful and accurate, it may not actually be accurate, and it risks misleading them into an erroneous understanding of the teaching construct. Lastly, while few-shot LLM-based models promise to speed up model development by not requiring data for fine-tuning, it is important to have data to assess their performance. In our case, the models suffered from low precision; more data might be needed to improve performance.

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 *Proc. of Building Educational Applications (BEA)*, pages 224–233.

Yuya Asano, Beata Beigman Klebanov, and Jamie Mikeska. 2025. Exploring task formulation strategies

to evaluate the coherence of classroom discussions with GPT-4o. In *Proc. of Building Educational Applications (BEA)*, pages 716–736.

Farnaz Badiie and David Kaufman. 2014. Effectiveness of an online simulation for teacher education. *Journal of Technology and Teacher Education*, 22(2):167–186.

Beata Beigman Klebanov, Andrew Hoang, Jamie Mikeska, Benny Longwill, Sanjna Kashyap, Shreyashi Halder, and Aakanksha Bhatia. 2026. [Data-lean fine-tuning of models for evaluating teacher performance in a GenAI-led elicitation simulation](#). In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 546–561, San Diego, California, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Michael Suhan, and Jamie Mikeska. 2025. [Towards evaluating teacher discourse without task-specific fine-tuning data](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 192–200, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

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.

James Bywater, Jennifer 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.

Gaowei Chen. 2023. Can ChatGPT detect student talk moves in classroom discourse? A preliminary comparison with BERT. In *Proceedings of EDM*.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educ. Evaluation & Policy Analysis*, 42(2):208–231.

Elizabeth Davis, Matthew Kloser, Andrea Wells, Mark Windschitl, Janet Carlson, and J. 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.

Elizabeth Davis and Annemarie Palincsar. 2023. Engagement in high-leverage science teaching practices among novice elementary teachers. *Science Education*, 107(2):291–332.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

- Marina Drushlyak, Tetiana Lukashova, Volodymyr Samonia, and Olena Semenikhina. 2025. Artificial intelligence in education: Chatgpt-based simulations in teachers' preparation. *Informatyka, Automatyka, Pomiarowy Gospodarce i Ochronie Środowiska*, 15(1):144–152.
- Nils Feldhus, Alik Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proc. Inf. Techn. for Social Good*, page 225–230.
- 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*.
- 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 *Proc. of AIED*, pages 60–74.
- 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 *Proc. of LAK*, pages 230–238.
- Amanda Hurlbut and Daniel Krutka. 2020. Where do we start?: Initiating a practice-based teacher education program around high-leverage practices. *Journal of Teacher Education and Educators*, 9(2):169–199.
- 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 *Proc. of Building Educational Applications (BEA)*, pages 182–198.
- Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. In *Proc. of LAK*, page 302–312.
- Elham Kazemi and Allison Hintz. 2023. *Intentional talk: How to structure and lead productive mathematical discussions*. Routledge.
- Özge Kelleci Alkan, Nuri Aksoy, Taibe Kulaksız, Hatice Kaplan, Büşra Durmaz, Mihriban Özcan, and Bilge Kalkavan. 2024. A multi-feedback system integrated simulation-based teacher training to scaffold pre-service teachers' teaching skills: A phenomenological approach. *Education and Information Technologies*, 29(15):20691–20713.
- Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.
- Susan Ledger, Madeline Burgess, Natasha Rappa, Brad Power, Kok Wai Wong, Timothy Teo, and Bruce Hilliard. 2022. Simulation platforms in initial teacher education: Past practice informing future potentiality. *Computers & Education*, 178:104385.
- Susan Ledger, Mailizar Mailizar, Sue Gregory, Miriam Tanti, David Gibson, and Stacy Kruse. 2025. Learning to teach with simulation: historical insights. *Journal of Computers in Education*, 12(1):339–366.
- Jiwon Lee and Jeongmin Lee. 2025. Generative ai chatbot for teachers' data-informed decision-making. *Educational Technology & Society*, 28(3):298–317.
- Adam Lefstein, Dana Veder-Weiss, and Aliza Segal. 2020. Relocating research on teacher learning: toward pedagogically productive talk. *Educational researcher*, 49(5):360–368.
- 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.
- Sofia Lindberg and Anders Jönsson. 2023. Preservice teachers training with avatars: A systematic literature review of "human-in-the-loop" simulations in teacher education and special education. *Education sciences*, 13(8):817.
- Oliver McGarr. 2021. The use of virtual simulations in teacher education to develop pre-service teachers' behaviour and classroom management skills: Implications for reflective practice. *Journal of Education for Teaching*, 47(2):274–286.
- James McLeskey, Bonnie Billingsley, Mary Brownell, Lawrence Maheady, and Timothy Lewis. 2019. What are high-leverage practices for special education teachers and why are they important? *Remedial and Special Education*, 40(6):331–337.
- Sarah Michaels and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue*, pages 347–362.
- Jamie Mikeska, Beata Beigman Klebanov, Aakanksha Bhatia, Shreyashi Halder, and Michael Suhan. 2025. Evaluating the use of generative AI to support learning opportunities for teachers to practice engaging in key instructional skills. In *Proc. of AIED*, pages 378–391.
- Jamie 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 *Proc. of AIED*, pages 251–258.

- Jamie Mikeska, Catherine Flynn, Devon Fetrow, Shreyashi Halder, Tricia Maxwell, Rutuja Ubale, Beata Beigman Klebanov, Abhilash Sarnad, Sanat Singh, and Mengxuan Zhao. 2026. Using multi-student generative AI teaching simulations as practice spaces for facilitating science discussions. In *National Association for Research in Science Teaching (NARST) Annual Meeting*, Seattle, WA.
- Jamie Mikeska, Heather Howell, Joseph Ciofalo, Adam Devitt, Elizabeth Orlandi, Kenneth King, and Glenn 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 Mikeska, Heather Howell, and Carrie Straub. 2019. Using performance tasks within simulated environments to assess teacher ability to engage in coordinated, accumulated, and dynamic (cad) competencies. *International Journal of Testing*, 19(2):128–147.
- Tanya Nazaretsky, Jamie Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *Proc. of LAK*, pages 122–132.
- Thomas Nygren, Marcus Samuelsson, Per-Olof Hansson, Evgenia Efimova, and Steven Bachelder. 2025. Ai versus human feedback in mixed reality simulations: Comparing llm and expert mentoring in preservice teacher education on controversial issues. *International Journal of Artificial Intelligence in Education*, pages 1–33.
- Catherine O'Connor and Sarah Michaels. 2019. Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, 97:166–175.
- Joanne O'Flaherty and Erin M Beal. 2018. Core competencies and high leverage practices of the beginning teacher: A synthesis of the literature. *Journal of Education for teaching*, 44(4):461–478.
- 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*.
- Meghan Shaughnessy, Hala Ghousseini, Elham Kazemi, Megan Franke, Megan Kelley-Petersen, and Elizabeth Sugino Hartmann. 2019. An investigation of supporting teacher learning in the context of a common decomposition for leading mathematics discussions. *Teaching and Teacher Education*, 80(1):167–179.
- Taekwon Son. 2025. Noticing classes of preservice teachers: Relations to teaching moves through AI chatbot simulation. *Education and Information Technologies*, 30(7):9161–9184.
- Yilmaz Soysal and Somayyeh Soysal. 2022. Exploring science teacher questions' influence on the students' talk productivity: A classroom discourse analysis approach. *Sage Open*, 12(2):21582440221102433.
- Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proc. of Building Educational Applications (BEA)*, pages 71–81.
- Seng Tan, Yee Tan, Chew Teo, and Guangji Yuan. 2025. Teachers' professional agency in learning with ai: A case study of a generative ai-based knowledge building learning companion for teachers. *British Journal of Educational Technology*.
- Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2023. Utilizing natural language processing for automated assessment of classroom discussion. In *Proc. of AIED, Posters and Late Breaking Results*, pages 490–496.
- 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.
- Andrew Tyminski, V Zambak, Corey Drake, and Tonia Land. 2014. Using representations, decomposition, and approximations of practices to support prospective elementary mathematics teachers' practice of organizing discussions. *Journal of Mathematics Teacher Education*, 17(5):463–487.
- Rose Wang and Dorottya Demszy. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proc. Building Educational Applications (BEA)*, pages 626–667.
- Liwei Wei, P Murphy, and Carla Firetto. 2018. How can teachers facilitate productive small-group talk? an integrated taxonomy of teacher discourse moves. *The Elementary School Journal*, 118(4):578–609.
- Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using LLMs to measure instruction quality. In *Proc. of NAACL*, pages 4375–4389.
- Nuodi Zhang, Fengfeng Ke, Chih-Pu Dai, Sherry A Southerland, and Xin Yuan. 2025. Seeking to support preservice teachers' responsive teaching: Leveraging AI-supported virtual simulation. *British Journal of Educational Technology*, 56(3):1148–1169.