

Data-lean fine-tuning of models for evaluating teacher performance in a GenAI-led elicitation simulation

Beata Beigman Klebanov Andrew Hoang Jamie Mikeska Benny Longwill
Sanjna Kashyap Shreyashi Halder Aakanksha Bhatia*

ETS Research Institute, Princeton, NJ, USA

{bbeigmanklebanov,ahoang,jmikeska,blongwill,skashyap,shalder001}@ets.org
aakankshabhatia01@gmail.com

Abstract

Recent advances in the capabilities of conversational agents based on large language models make them a very promising tool for role playing K-12 students in order to train educators in conversational teaching practices, such as eliciting student thinking, explaining disciplinary content, and facilitating a classroom discussion. In fact, such simulations can and have been developed relatively quickly and without data to machine-learn from – neither classroom data nor human-simulated data. To enhance the usefulness and effectiveness of such teaching simulations, it is necessary to provide pedagogically sound, timely, and personalized feedback to the educator about their simulation performance. In this study, we present experiments on fine-tuning models to evaluate educator performance in an elicitation teaching simulation. The models are developed with data collected during usability testing of the simulation and evaluated on real user data. We show that even with relatively little fine-tuning data, robust performance can be obtained.

1 Introduction

Digital teaching simulations (**DTS**) have increasingly been integrated into teacher education settings to support educators in learning how to engage in high-leverage teaching practices, such as eliciting student thinking, facilitating productive discussions, and communicating with families, and have been shown to improve educators' skills, knowledge and beliefs (Dalinger et al., 2020; Dieker et al., 2019; Ersozlu et al., 2021; Lee et al., 2024; Mikeska et al., 2025a). One of the most common types of DTS involves educators in conversing with one or more student avatars via an online interface where the avatars appear visually and talk to each other and the educator in real time. In these simulations, there is a human-in-the-loop called a

simulation specialist who has been trained to use specialized technology, such as voice modulation software and game controllers, and to enact various student personas that align with the student avatars' pre-specified personalities, backgrounds, previous experiences, and content understandings (Bondie et al., 2021).

Research has shown that productive simulation-based learning experiences involve instructional or learning cycles whereby teacher educators or professional development facilitators support educators in preparing for, engaging in, and then debriefing and reflecting on their simulation performances (Lottero-Perdue et al., 2023; Mikeska et al., 2023a; Woods et al., 2024). These instructional cycles can also include formative feedback from a teacher educator, coach, educators' peers, or via written feedback reports. Research has suggested that such feedback supports educators in identifying their instructional strengths and areas for growth and in learning how they can apply what they are learning to K-12 classrooms (Cohen et al., 2020; Horn et al., 2023; Mikeska et al., 2023b; Pecore et al., 2023). However, despite the empirical evidence for the efficacy of DTS, they are challenging to scale due to the need for the simulation specialist to power the student avatars during each simulation.

To address this scaling challenge, recent research has begun exploring how generative AI (**GenAI**) can be used to simulate student responses (Marquez-Carpintero et al., 2025; Hong et al., 2025; Liu et al., 2026; Pan et al., 2025; Markel et al., 2023; Lim et al., 2025; Zhang et al., 2026; Zhuang and Zhang, 2025; Lee et al., 2025, 2026; Gíslason, 2026; Ziying et al., 2026; Son, 2025). While users perceive GenAI teaching simulations as useful to support their learning (Lim et al., 2025; Markel et al., 2023; Mikeska et al., 2025b; Zhang et al., 2026) and manual analyses are being applied to explore user activity in a simulation (Lee et al., 2026; Son, 2025), we are not aware of studies focusing

*The author was at ETS at the time of research.

on automated assessment of teacher performance in a GenAI simulation for formative purposes.

Prior work on automated evaluation of teacher performance has been conducted on annotated classroom discussions data, such as mathematics discussions (TalkMoves (Suresh et al., 2022), NCTE (Demszky and Hill, 2023)), ELA discussions (Tran et al., 2024), small-group discussions in DTS (Asano et al., 2025; Nazaretsky et al., 2023), and tutoring dialogues (Cao et al., 2025; Macina et al., 2023). In contrast with at least a hundred teacher performances in each of the above, we explore feasibility of evaluating teacher performance in a newly created GenAI-led simulation with very little development data – only 40 performances, collected during usability testing of the simulation.

2 Related work

A variety of constructs related to effective teaching have been modeled computationally based on human-annotated datasets, predominantly of real or simulated classroom discussions and tutorial dialogues. These include specific teaching actions such as asking effective questions (Alic et al., 2022; Jensen et al., 2020; Kelly et al., 2018), encouraging students' interaction (Ilgan et al., 2024), supporting student motivation (Hunkins et al., 2022), enacting conversational uptake (Demszky et al., 2021; Jensen et al., 2020); more comprehensive analyses of various teacher and student "talk moves", common types of teacher and student contributions to classroom discourse (Suresh et al., 2022; Tran et al., 2024; Cao et al., 2025; Wang et al., 2025), as well as holistic characteristics of discussions such as coherence (Asano et al., 2025), extent of instructional support (Whitehill and LoCasale-Crouch, 2024), and instructional quality (Xu et al., 2024).

Methods used for automated analyses of educational discourse ranged from feature-based machine learning (Donnelly et al., 2017; Jensen et al., 2020; Kelly et al., 2018; Hunkins et al., 2022; Samei et al., 2014), to deep neural networks, including transformers (Demszky et al., 2021; Schlotterbeck et al., 2021; Kupor et al., 2023; Nazaretsky et al., 2023; Alic et al., 2022; Xu et al., 2024; Ilgan et al., 2024), to large language models (LLMs) using techniques such as few-shot prompting, chain-of-thought, and parameter-efficient fine-tuning (Wang and Demszky, 2023; Kupor et al., 2023; Whitehill and LoCasale-Crouch, 2024; Hou et al., 2024; Wang et al., 2025; Cao et al., 2025).

Wang et al. (2025) compared LLM prompt-engineering, full fine-tuning with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and parameter-efficient fine-tuning (PEFT) of LLMs on the TalkMoves dataset. They found that PEFT approaches were stronger on the talk moves Restating and Revoicing of the students' utterances but performed comparably to RoBERTa on the other constructs. However, these findings were obtained with fine-tuning on hundreds of transcripts totaling about 1.5M teacher words. Beigman Klebanov et al. (2025) studied few-shot vs fine-tuned models on about 1K teacher utterances in small-group discussion DTS and found fine-tuned models to be superior; however, the annotated data included only select, clear-cut utterances per transcript. Tran et al. (2024) evaluated both fine-tuned BERT and few-shot LLMs on detecting productive teaching moves and found that fine-tuned models were generally competitive though yielded to LLMs in certain configurations. Focusing specifically on the amount of fine-tuning data in the broader context of text classification, Zhang et al. (2025) compared BERT-like fine-tuned models, zero-shot LLMs, and LLM-internal-states-based methods on six text classification tasks and concluded that BERT-like models were stronger for what the authors termed "general text classification," but the strong performance of BERTs was predicated on tuning on moderate to abundant data. When data size was reduced to a few hundred (up to 450 instances, comparable to our scenario), they sank below GPT-4o results for the three datasets on which both model types of evaluated. In the current study, we investigate a data-lean scenario focusing on fine-tuned models, with an LLM-based few-shot model as a baseline.

3 The GenAI-led teaching simulation

The simulation was designed to support elementary pre-service and in-service teachers and paraeducators in developing the teaching competency of eliciting student thinking in elementary mathematics. During the simulation, teachers engaged in voice conversation with the Cecilia chatbot. Cecilia's responses were generated in accordance with a predefined student profile (section 3.1), allowing teachers to practice instructional strategies to elicit Cecilia's method for solving the addition word problem and her understandings and misunderstandings around addition. Cecilia was designed to respond to the teacher's questions with

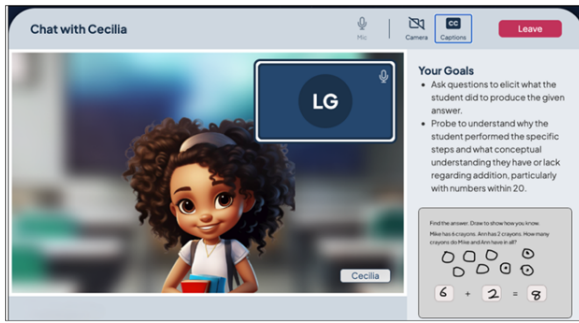


Figure 1: Interface of the GenAI teaching simulation.

the minimal amount of information that was responsive to the question, thus allowing the teacher to practice elicitation. That is, when asked a question like ‘What did you do to solve the problem?’, Cecilia would describe the first step but not the full process and would not divulge her reasoning unless explicitly prompted to do so by the teacher.

3.1 The strategies for adding task

The Strategies for Adding task is based on design criteria for elicitation tasks (Qi and Sykes, 2016). Cecilia’s knowledge profile specifies both her understanding and gaps in knowledge when solving addition problems. In this task, Cecilia solved the word problem: “Mike has 6 crayons. Ann has 2 crayons. How many crayons do Mike and Ann have in all?” Cecilia solved this problem and showed her understanding in her written work (Figure 1 bottom right). She understands that 6 represents Mike’s crayons, 2 represents Ann’s crayons, and 8 represents the total. Cecilia counts on from the first addend and uses models (e.g., drawings) to represent and solve the problem. However, she cannot add numbers fluently. She does not understand that counting on from the larger addend (regardless of the position) is more efficient. She is not familiar with any strategies besides drawing and counting for addition and she does not understand the commutative property. This knowledge profile represents a step in a typical progression of learning addition for early elementary students.

In the simulation, teachers were not provided with detailed information about Cecilia’s conceptual understanding. They were instructed to elicit what Cecilia did and did not understand regarding addition (top right – ‘Your Goals’ – in Figure 1). A 10-minute preparation time was provided, allowing teachers to consider their elicitation strategies.

3.2 Implementation of the simulation

Chatbot default dialogue responses were generated using Azure OpenAI’s GPT-4o model (v2024-08-06) via Microsoft Azure’s Chat Completions API. Each request was issued as a single-turn completion conditioned on a system prompt that encoded the knowledge profile of the simulated student, instructional constraints, and worked examples along with few-shot dialog examples and the full conversation history.

Response generation followed a hybrid pipeline combining LLM-based dialogue generation with deterministic validation and filtering. Mathematical expressions in user input were normalized using GPT-4o, then parsed and evaluated using Python’s Abstract Syntax Tree (AST) library to enforce structural constraints on practice problems (i.e., a single addition operation with two operands and a sum of up to 20). Inputs that failed the structural requirements triggered replacement of the initial response with predefined messages (e.g., “Oops! I don’t understand this!”).

In addition to structural validation on user input, regular expression-based checks were applied to detect mathematical vocabulary that Cecilia should not be familiar with (e.g., equation) and out-of-range numerical values (i.e., any number greater than 20), with violations resulting in appropriate canned responses (e.g., “I don’t know what that means” and “That’s too many to draw”, respectively). If no constraints were violated, a final LLM-based simplification step adjusted lexical complexity and limited responses to a single procedural step. A more comprehensive description of the implementation of the chatbot and the evaluation of the accuracy of Cecilia’s responses in terms of responsiveness to the teacher’s questions; appropriateness in age, emotional tone, amount of information provided, and coherence with respect to prior conversation; and alignment with the student’s knowledge profile can be found in Mikeska et al. (2025b).

4 Data

4.1 Participants

Study 1: A total of 45 educators (paraeducators, in-service teachers, and principals) participated in a usability study to test the simulation. The age range was 25 to 64 years. Among them, 67% identified as White, 18% as Hispanic or Latino, 13% as Black or African American, and 4% as Asian or Asian American. The majority of educators were

female (82%). Regarding educational background, 67% held a Master's degree, 13% a Bachelor's degree, 11% a Doctoral degree, 7% other degrees, and 2% a high school diploma. Additionally, 76% of the educators had prior experience in teaching mathematics. Their familiarity with digital teaching simulations varied: 11% of were very familiar, 31% were somewhat familiar, 38% were a little familiar, and 20% were not familiar. We collected data from 45 participants; however, due to technical issues, we had to exclude 5 transcripts from the analysis. The 40 transcripts collected in this study comprised the **development** dataset.

Study 2: This study was the first implementation pilot of the simulation within a teacher education context. The data was collected from 10 pre-service teachers (PSTs) enrolled in a mathematics method course at a private college in the Northeastern USA. PSTs were aged 18-24. At the time of data collection, 90% of them had obtained a high school diploma. Half of the participants had experience as substitute teachers or in other teaching capacities. In terms of familiarity with digital teaching simulations, 50% were not familiar, 30% were somewhat familiar, and 20% were a little familiar. The gender and ethnicity data for PSTs were not collected.

In this study, as part of the coursework, PSTs participated in two instructional cycles using the Strategies for Adding simulation. During the first cycle, PSTs prepared for the simulation by reviewing the simulation's goal and the scenario's background. The 10-minute preparation phase included reviewing Cecilia's written work and considering questions or prompts to use to elicit her problem-solving process and conceptual understanding of addition. Next, they engaged in the simulation, followed by a reflection session during their 50-minute class time. Their mathematics methods teacher educator facilitated the reflection activity, during which PSTs reviewed Hallman-Thrasher et al. (2020) categorization of questions by purpose and analyzed their own transcript. Each PST then compared their transcript with a partner's to further reflect on how to elicit Cecilia's thinking and modify their questioning strategy in the future.

After one week, the second instructional cycle took place. After the second simulation, PSTs engaged in a reflection activity and categorized their questions again. In this reflection, they focused on the information each question gathered about Cecilia's mathematical understanding. They participated in a whole-class discussion where PSTs

analyzed Cecilia's conceptual understanding and knowledge gaps based on their simulation experiences. Afterward, they completed a debrief assignment to reflect on their elicitation techniques. The 20 transcripts collected in this study comprised the **test** dataset.

4.2 Annotation of aspects of teaching quality

Facilitating productive talk in classrooms with and between students requires that teachers use specific instructional moves to gauge students' initial and developing ideas and prompt them to share their thinking and understanding (Michaels and O'Connor, 2015; Wei et al., 2018). Eliciting student thinking requires teachers to attend to specific student ideas, ask follow-up questions, and explain their reasoning (Estapa and Davis, 2023; Shaughnessy and Boerst, 2018). In this study, we focus on three key constructs related to the practice of eliciting student thinking summarized in Table 1.

| Construct | Description |
|-----------|--------------------------------|
| ASI | Attends to Specific Ideas |
| AQTA | Asks Questions Tied to Actions |
| SWE | Show Work and Explain |

Table 1: Summary of elicitation constructs.

Two raters with content analysis expertise participated in the annotation. Each teacher turn was annotated for each of the three constructs. The development set was used for developing rubrics, rater training, joint coding, individual coding, and double coding to monitor agreement. For the test set, one transcript was coded jointly and discussed, while the remaining 19 transcripts (266 teacher turns in all) were coded independently to evaluate inter-annotator agreement. The coding was reliable, achieving Cohen's κ of 0.88, 0.86, and 0.93, for AQTA, ASI, and SWE, respectively.

Table 2 exemplifies the flow of the conversation and the annotations of the teacher's utterances.

The first four teacher utterances contain at least one effective elicitation move each. The first two include asking Cecilia about something that she has done (provided the answer 8 in utterance#1, drew circles in utterance#3; AQTA = 1), specifically to explain the process ('how did you...') or reasoning ('why did you...'), respectively (SWE = 1). The third teacher's utterance (#5 in the Table) includes asking for Cecilia's reasoning about what she has just done (AQTA = 1, SWE = 1), following up

| U# | Speaker | Utterance | Classifications | | |
|----|---------|---|-----------------|------|-----|
| | | | ASI | AQTA | SWE |
| 1 | Teacher | Cecilio, how did you come up with the answer of eight? When we asked. About the crayons. | 0 | 1 | 1 |
| 2 | Cecilia | I drew circles for the crayons. Mike had 6 and Ann had 2. I counted 6,7,8. | | | |
| 3 | Teacher | That’s great. I noticed. I noticed when you drew your circles that you have three circles. With a dot inside of them. Why did you put the dot inside those three circles? | 0 | 1 | 1 |
| 4 | Cecilia | I put dots to show where I started counting Ann’s crayons. | | | |
| 5 | Teacher | So why do you have. Three dots instead of just two dots. If Anne only has two crayons. | 1 | 1 | 1 |
| 6 | Cecilia | The first dot is where I started counting from Mke’s crayons. | | | |
| 7 | Teacher | Do you worry that you might get confused by having three dots to represent Anne’s two? Crayons. | 1 | 0 | 0 |
| 8 | Cecilia | Nope! I know which dots go to Ann’s crayons because I still count them starting at Mike’s crayons. | | | |
| 9 | Teacher | You were able to come up. With the correct answer, Cecilia. But I worry in the future. Your extra dot might cause some confusion. | 0 | 0 | 0 |
| 10 | Cecilia | Okay. I’ll try to remember that. | | | |
| 11 | Teacher | Maybe. A helpful strategy going forward would be to put the number one. In the circle where you started instead of using dots. | 0 | 0 | 0 |
| 12 | Cecilia | Okay, I’ll try that! | | | |

Table 2: An excerpt from a conversation between the Teacher (human participant) and Cecilia (the GenAI powered student). Data is shown as it appears in the automated transcription. Human classifications of each teacher utterance for the three constructs are shown in the last three columns.

on Cecilia’s idea of putting the dots where she starts counting Ann’s crayons (ASI = 1). The next teacher utterance continues the follow-up on Cecilia’s three dots idea (ASI = 1), though not about a specific action just performed (AQTA = 0) and not asking a process or reasoning question, but an attitude question (‘do you worry...’, SWE = 0). The last two teacher utterances do not contain any of the effective moves. The final utterance, specifically, is the teacher’s attempt to teach Cecilia to improve her strategy for adding – which is out of scope for the simulation that focuses on eliciting the student’s ideas rather than on changing them. A more effective elicitation strategy would have been to move on after the first three exchanges and ask Cecilia to solve a different problem that would help the teacher further explore Cecilia’s mathematical understanding, for example, asking her to add more than two numbers or bigger numbers, or checking if she is familiar with the commutative property by solving 2+6 instead of 6+2.

4.3 Datasets

Table 3 shows the descriptive statistics of the data.

| | Development set | | | Test set | | |
|-----|-----------------|------|-----|----------|------|-----|
| #Pa | 40 | | | 10 | | |
| #Tr | 40 | | | 20 | | |
| #Ut | 630 | | | 275 | | |
| #W | 11K | | | 3.6K | | |
| | ASI | AQTA | SWE | ASI | AQTA | SWE |
| %1 | 28 | 26 | 23 | 27 | 40 | 29 |

Table 3: Descriptive statistics of the datasets. #Pa: Participants. #Tr: Transcripts. #Ut: Teacher utterances. #W: Total word count of all teacher utterances. %1: Percentage of teacher utterances labeled as class 1.

5 Models for evaluating teacher discourse

The goal of the evaluation models is to power formative feedback to teachers immediately following the simulation. Two types of feedback are created. First, for every construct, we show up to 3

utterances where the teacher engaged in the target behavior and up to 3 utterances where they did not. As an additional resource, examples from other teacher’s performances are provided. Second, for every construct, we show the overall frequency of occurrence of utterances where the target behavior occurred, and compare it to the range calculated using data from well-performing teachers. To support both types of feedback, the evaluation models are implemented as classifiers. For each construct, we classify each teacher utterance as showing the target behavior (class 1) or not (class 0).

5.1 Baselines

5.1.1 Few-shot models

Each construct was evaluated through LLM prompting using Azure OpenAI GPT-4o structured outputs. Each prompt instructed the model to identify and extract exact conversational turns from the transcript containing evidence for a single construct. Structured output returned verbatim excerpts with original role (i.e., student or teacher), message string, and turn identifier fields, enabling turn-level binary classification downstream. An empty list was returned when no evidence was found. All prompts were constructed as single system messages containing task instructions, the standardized problem context, construct-specific requirements with between two and five positive and negative examples picked by the subject matter expert on the team, and the full chat history transcript. Appendix B provides the prompts.

5.1.2 Fine-tuned baseline

As a baseline encoder model, we used DeBERTa (He et al., 2021) model deberta-v3-base¹ (He et al., 2023) from HuggingFace’s Transformers library (Wolf et al., 2020). We picked DeBERTa as it was the latest widely used iteration of BERT-style family of models (at the time of analysis) that achieved state-of-the-art performance on natural language understanding benchmarks like MNLI (Williams et al., 2018) and SQuAD v2.0 (Rajpurkar et al., 2018), improving on both BERT and RoBERTa models (He et al., 2021). The model has 12 layers and a hidden size of 768, 86M backbone parameters with a vocabulary of 128K tokens. This model was pre-trained using 160GB of data. For fine-tuning for our classification task, the learning rate was set to 5e-5, trained for 20 epochs. Only the

¹<https://huggingface.co/microsoft/deberta-v3-base>

current teacher utterance were used as input to classify the utterance. We fine-tuned the model on data from four development set folds and evaluated on the fifth fold, in a 5-fold cross-validation design.

5.2 Experimental model

Using the development set, we explored a number of strategies for improving the baseline fine-tuned classifier. We used 5-fold cross-validation on the development set to evaluate the effectiveness of the strategies, separately and in combination. We discuss each strategy in turn.

Weighed loss to account for category imbalance (WB): As can be seen in Table 3, the class distribution is somewhat skewed, with class 1 accounting for fewer than 30% of the teacher utterances for each construct in the development data. In weighted loss, the weights were inversely proportional to the class frequencies.

Context (CO): Prior research has shown that incorporating a limited context of a few utterances can be effective (Tran et al., 2024; Cao et al., 2025). We experimented with a sliding window of three utterances, from three utterances preceding the current one, two preceding and one subsequent, all the way to three subsequent utterances as context. Based on the results on the development set, we selected the scheme with one utterance preceding the current one (the student’s utterance right before the current teacher utterance) and two utterances following the current one (the student’s response and the teacher’s next utterance) as context.

Pre-training using real classroom discussion transcripts (CD): In order to attune the model to teacher-student educational discourse, we pre-trained on the publicly available data of classroom discussion transcripts from the NCTE corpus (Demszky and Hill, 2023). The corpus contains 1,660 transcripts of 45-60 minute 4th and 5th grade mathematics lessons collected by the National Center for Teacher Effectiveness. Cao et al. (2025) found that pre-training on classroom discussion data was useful for developing better classifiers of the same talk moves in tutorial dialogues. Our data isn’t tutoring, in the sense that the teacher’s goal isn’t to help the student learn, but to elicit the students’ current state of knowledge; nevertheless, classroom discourse in the domain of mathematics could be useful pre-training material. We experimented with 1, 5, 10, and 15-epoch pre-training and found that 10-epoch pre-training was the most effective.

Multi-label classification (ML): We observed that while the three teaching constructs capture different aspects of a high quality teacher performance in the simulation, they sometimes happen in the same teacher utterance; see, for example, the conversation excerpt shown in Table 2. Quantitatively, we found that Pearson’s correlation between the list of binary labels for AQTA and SWE across the 630 teacher utterances in the 40 development set simulations was $r = 0.59$, suggesting a tendency to occur in the same utterances. The other correlations were weaker but still significant: $r = 0.37$ for AQTA and ASI and $r = 0.25$ for ASI and SWE. We implemented a multi-label classification with zero to three labels for each teacher utterance.

Selecting the final model for benchmarking: Table 4 shows the results for each strategy on its own (rows WB through CD), using 5-fold cross-validation on the development set. In addition to the four evaluation metrics (see section 5.3) for each of the three constructs, yielding $4 \times 3 = 12$ numbers for each model, we also provide a summary evaluation in the form of the count of cases, out of 12, where the result was substantially better (\uparrow) or worse (\downarrow) than the fine-tuned baseline, using a change of 0.04 or more as a cutoff. We observe that all strategies were net-positive, namely, increases outnumbered decreases, but (a) all incurred some decreases, and (b) the best net result was only 2 out of 12, suggesting overall mild positive impact. We experimented with combining strategies and picked the best combination, shown in row F(inal): The model that included pre-training on the classroom (NCTE) data, multi-label learning, and context. That model achieved a net effect of 8 out of 12, without incurring any decreases. We call this model **FT-Exp** henceforth and benchmark it against the baselines on the test set to evaluate whether the improvement potential is realized on a new set of transcripts unseen during development.

5.3 Model evaluation

To evaluate the models for their ability to identify teacher utterances that realize the given teaching construct, we use both a retrieval paradigm (precision, recall, and F1 scores for the rarer positive class) and a classification paradigm (Cohen’s κ).

For each of the fine-tuned models (FT-BL and FT-Exp), we used five models that we fine-tuned on four out of the five cross-validation folds each. We compute the evaluation metrics for each model on the test set and show the average performance

| Mod. | Met. | ASI | AQTA | SWE | \uparrow | \downarrow |
|------|----------|-----|------|-----|------------|--------------|
| BL | Prec. | .54 | .74 | .61 | | |
| | Rec. | .58 | .72 | .72 | | |
| | F1 | .55 | .73 | .64 | | |
| | κ | .37 | .62 | .52 | | |
| WB | Prec. | .58 | .68 | .58 | 1 | 1 |
| | Rec. | .61 | .79 | .75 | 1 | |
| | F1 | .56 | .73 | .65 | | |
| | κ | .41 | .61 | .52 | 1 | |
| CO | Prec. | .58 | .66 | .63 | 1 | 1 |
| | Rec. | .60 | .64 | .72 | | 1 |
| | F1 | .59 | .72 | .65 | 1 | |
| | κ | .42 | .60 | .54 | 1 | |
| ML | Prec. | .52 | .70 | .66 | 1 | 1 |
| | Rec. | .70 | .76 | .70 | 2 | |
| | F1 | .58 | .71 | .65 | | |
| | κ | .39 | .60 | .54 | | |
| CD | Prec. | .55 | .69 | .64 | | 1 |
| | Rec. | .73 | .83 | .67 | 2 | 1 |
| | F1 | .62 | .74 | .62 | 1 | |
| | κ | .44 | .63 | .50 | 1 | |
| F | Prec. | .58 | .76 | .67 | 2 | |
| | Rec. | .70 | .76 | .69 | 2 | |
| | F1 | .61 | .76 | .67 | 1 | |
| | κ | .44 | .66 | .56 | 3 | |

Table 4: Dev-set performance of the baseline (BL), each strategy on its own (WB: Weight Balance; CO: Context; ML: Multi-Label; CD: Classroom Discourse) and the final selected model (F). The last two columns count the number of metrics with increase (\uparrow) and decrease (\downarrow) of at least 0.04 wrt BL.

as the final test set evaluation. In addition, we show the evaluation metrics on the test set for the best aggregate model, as explained in the next section.

5.4 Robustness through aggregation

Due to non-determinism, it is common practice to train multiple times and aggregate the results. It is also possible to intentionally diversify the models by training multiple models on various sub-samples of the data. Aggregation of predictions from multiple such models could improve robustness. Requiring more models to agree is likely to improve precision but hurt recall, while using a permissive aggregation, e.g., positive if at least one model predicts positive, could help recall but hurt precision.

To gauge the optimal aggregation strategy, one ideally would train multiple models based on different subsets of the fine-tuning data for each of the cross-validation folds, resulting in $5 \times 5 = 25$ dif-

ferent models. This would require sub-partitioning of each fold into 80% and 20% again; given that a fold already contains only 8 transcripts, this design would further reduce the training data and result in a small evaluation set for the aggregation models. We therefore opted for a different design to gauge the optimal aggregation strategy: We randomly picked one of the FT-Exp models fine-tuned on four development folds and re-tuned it with four different random seeds. This created some variation in the models, and we considered how many positive votes we should require for the best balance of precision and recall. The results are shown in Table 6 in Appendix A. The results showed that while the most permissive model has the best recall and the most conservative had the best precision, as expected, the best precision and recall combination is achieved by a majority vote – requiring three out of the five models to assign the teacher utterance to class 1 in order for the aggregate to be class 1.

Based on these findings, for the test set, we will provide the results of the majority aggregation of five models (Agg FT-Exp), each fine-tuned on four out of five development set folds.

6 Results

Table 5 shows the results. We observe that fine-tuning was quite effective, the small number of transcripts notwithstanding. Both FT models improved over the Few Shot baseline on all metrics, on average across the three target constructs.

Next, we observe that FT-Exp gained in recall (+0.12) more than it lost in precision (-0.06) with respect to FT-BL, on average across constructs, resulting in improvements in F1 and κ scores, both on average and for each of the constructs.

Let us next compare the results for FT-Exp, shown as average performance on the test set across five models trained on four out of five development folds, to Agg FT-Exp, namely, a single system that aggregates the classifications generated by the five systems above using a majority vote. The results show clearly that the aggregated model is similar or better than the average single model.

7 Discussion

Overall, there is no reason to prefer few-shot (FS) models over fine-tuned (FT) ones, as the only cases of lower scores for FT models are recall for ASI for FT-BL and precision for ASI for FT-Exp, but, in both cases, these were compensated by a larger in-

| Model | Metric | ASI | AQTA | SWE | Av. |
|----------|----------|-----|------|-----|-----|
| Few Shot | Prec. | .53 | .65 | .73 | .64 |
| | Recall | .45 | .59 | .61 | .55 |
| | F1 | .49 | .62 | .67 | .59 |
| | κ | .29 | .41 | .53 | .41 |
| FT BL | Prec. | .68 | .82 | .71 | .74 |
| | Recall | .41 | .71 | .70 | .61 |
| | F1 | .48 | .75 | .71 | .65 |
| | κ | .36 | .60 | .59 | .52 |
| FT Exp. | Prec. | .50 | .85 | .70 | .68 |
| | Recall | .71 | .72 | .77 | .73 |
| | F1 | .58 | .78 | .73 | .70 |
| | κ | .38 | .64 | .61 | .54 |
| Agg. | Prec. | .54 | .84 | .70 | .69 |
| | Recall | .81 | .77 | .78 | .79 |
| | F1 | .65 | .81 | .74 | .73 |
| | κ | .48 | .69 | .60 | .59 |

Table 5: Test-set performance of the two baselines and the best experimental models.

crease in precision and recall, respectively, showing overall gains in κ and similar or better F1 scores. For AQTA and SWE, the FT-Exp model crossed the $\kappa \geq 0.60$ threshold that is often taken to be sufficient for formative purposes, showing that fine-tuning, even with relatively little data, can result in usable models. That said, one of the constructs – ASI – remained challenging for all the models.

The FT-Exp model is the most promising, not only due to quality of performance, but also in terms of maintenance and cost of exploitation. Since FT-Exp employs multi-label learning, there is just one model dealing with all the three constructs. Being a relatively small BERT-family model, it can be installed locally and avoid incurring cost per query, as do the larger and more powerful LLMs, like GPT-4o used here for the few-shot model.

Our results also suggest that it may be beneficial to introduce some variation in the models by fine-tuning multiple models on somewhat different subsets of the fine-tuning data. In particular, majority aggregation of five such models results in improved recall across the board, without paying a price in precision. A different aggregation strategy to prioritize recall or precision could be devised, as well as optimizing the number of models and the extent of difference between them; more development data would be necessary for this optimization task. Our results show that a strategy evaluated on development set using a model variation method

that does not require data sub-partitioning resulted in a robust aggregation strategy that generalized to a blind test set with a different model variation method. While using multiple models to attain more robust performance is not a new finding in the LLM world, our results show that model aggregation can be successfully developed and evaluated using relatively little data.

Limitations

Although our results cover multiple effective elicitation moves, they are all studied in the context of a single task. Additional elicitation tasks in upper elementary mathematics and science are currently under development. We intend to replicate the study design when early data become available.

In addition, while the test set contained data from the target population of users and was collected in the target context of use of the simulation, the data come from a small number of pre-service teachers. As the simulation is used in subsequent data collections, we will examine the performance of the automated evaluation models on additional data to gauge the model's generalization.

We note that the focus of the current paper is the evaluation of teacher performance in the simulation. The usefulness of an evaluation like this is related but not fully determined by the simulation itself constituting an effective practice based *learning* opportunity for educators. In particular, even if the simulation has some deficiency in terms of the realism of the simulated student's discourse (Scarlatos et al., 2026), automated evaluation of the teacher's performance could still be an efficient way to gauge the teacher's current skill with elicitation strategies, if validated for this purpose. We thank a BEA reviewer for bringing this issue up.

Due to team capacity limitation, we explored only one of the most powerful LLMs available (at the time when the few-shot baselines were developed, GPT 4.0 was the top model). It is possible that few shot learning results would be better with the newer GPT model and/or with other LLMs such as Claude. It is also possible that open-source LLMs such as Llama can also be effectively fine-tuned in a parameter-efficient way (Wang et al., 2025), though, as Tran et al. (2024), we note that the data sizes are most likely too small to move the needle for these models. We leave it to future work to explore other LLMs, especially for the ASI construct for which we do not yet have a satisfac-

tory model. We believe our findings regarding the effectiveness of pre-training using classroom data, of flexible context that may include looking ahead in the conversation, of utilizing the teacher's ability to include more than one effective strategy in a single utterance, and model variation and aggregation methods are a good starting point for further research in automated evaluation of teacher performance.

Ethics statement

The data collections that yielded the data used in this study were conducted with the approval of the IRB of the authors' institution.

Feedback to the user in newly developed tasks is an inherently complicated proposition. This is because for the feedback to be most pedagogically appropriate, it needs to take into account how users typically behave in the activity, what strong and weak performances look like, and what properties of the performances best differentiate between the two. This is very hard to do with only usability data that may not even come from the target demographic and/or the target use settings. Therefore, provision of feedback is often delayed until well-performing feedback models can be built, which may deprive the early users of some of the potential benefit of participating in the simulation. Our results are quite optimistic in this respect: With only 40 development transcripts obtained through a usability study, we attained acceptable performance for two out of the three target constructs, suggesting that feedback along these two dimensions can now be responsibly provided to subsequent users.

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.
- 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.
- 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.
- Jie Cao, Abhijit Suresh, Jennifer Jacobs, Charis Clevenger, Amanda Howard, Chelsea Brown, Brent Milne, Tom Fischhaber, Tamara Sumner, and James H. Martin. 2025. [Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7671–7684, Abu Dhabi, UAE. Association for Computational Linguistics.
- 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.
- Tara Dalinger, Katherine Thomas, Susan Stansberry, and Ying Xiu. 2020. A mixed reality simulation offers strategic practice for pre-service teachers. *Computers & Education*, 144:103696.
- Dorottya Demszky and Heather Hill. 2023. [The NCTE transcripts: A dataset of elementary math classroom transcripts](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.
- 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*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lisa Dieker, Carrie Straub, Michael Hynes, Charles 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*, 11(4):1–20.
- Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D’Mello. 2017. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, pages 218–227.
- Zara Ersozlu, Susan Ledger, and Linda Hobbs. 2021. Virtual simulation in ITE: technology driven authentic assessment and moderation of practice.
- Anne Estapa and Jeni Davis. 2023. Prospective teachers’ instructional decisions and pedagogical moves when responding to student thinking in elementary mathematics and science lessons. *International Journal of Science and Mathematics Education*, 21(5):1703–1724.
- Ingólfur Gíslason. 2026. Learning about human mathematical dialogue from dialogue with chatbots: Babbaling, gargling and funnelling. *The Mathematics Enthusiast*, 23(1):167–184.
- Allyson Hallman-Thrasher, Ginger Rhodes, and Kyle Schultz. 2020. Supporting mathematics teacher educators’ practices for facilitating prospective teachers’ mathematical explanations in content courses. *The Mathematics Enthusiast*, 17(2):883–906.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *International Conference on Learning Representations*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.
- Sumin Hong, Jewoong Moon, Taeyeon Eom, Idowu David Awoyemi, and Juno Hwang. 2025. Generative AI-enhanced virtual reality simulation for pre-service teacher education: A mixed-methods analysis of usability and instructional utility for course integration. *Education Sciences*, 15(8):997.
- Annemarie Horn, Marcia Rock, Laura Chezan, Jonna Bobzien, Olga Karadimou, and Aeshah Alturki. 2023. Effects of e coaching on the occurrence, equity, and variety of behavior specific praise during Mursion™ simulations. *Journal of Special Education Technology*, 38(4):501–514.
- 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.
- 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 the 12th International Learning Analytics & Knowledge Conference*, 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 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 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 the 2020 CHI Conference on Human Factors in Computing Systems, CHI ’20*, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Sean Kelly, Andrew M. Olney, Patrick Donnelly, Martin Nystrand, and Sidney K. D’Mello. 2018. [Automatically measuring question authenticity in real-world classrooms](#). *Educational Researcher*, 47(7):451–464.
- Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. [Measuring five accountable talk moves to improve instruction at scale](#). *arXiv preprint arXiv:2311.10749*.
- Dabae Lee, Taekwon Son, and Sheunghyun Yeo. 2025. [Impacts of interacting with an AI chatbot on preservice teachers’ responsive teaching skills in math education](#). *Journal of Computer Assisted Learning*, 41(1):e13091.
- Dabae Lee, Corey Webel, Sheunghyun Yeo, Burcu Alapala, and Rebekah Hanak. 2026. [Cultivating responsive teaching with AI: exploring preservice teachers’ questioning patterns with student-emulating agents](#). *Journal of Mathematics Teacher Education*.
- Tammy 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.
- 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.
- Bojie Liu, Yu Yang, and Guandong Xu. 2026. [GenAI-empowered virtual micro-teaching training system for preservice teachers](#). In *Behavioural and Social Computing*, pages 232–241, Singapore. Springer Nature Singapore.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *Preprint*, arXiv:1907.11692.
- Pamela Lottero-Perdue, Karen Cimino, and Julia Brandeberry. 2023. [Preservice teachers facilitating a discussion with elementary student avatars before facilitating it with real students](#). *Innovations in Science Teacher Education*, 8(4).
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.
- Julia Markel, Steven Opferman, James Landay, and Chris Piech. 2023. [GPTeach: Interactive TA training with GPT-based students](#). In *Proceedings of the 10th ACM conference on Learning@ Scale*, pages 226–236.
- Luis Marquez-Carpintero, Alberto Lopez-Sellers, and Miguel Cazorla. 2025. [Simulating students with large language models: A review of architecture, mechanisms, and role modelling in education with generative AI](#). *arXiv preprint arXiv:2511.06078*.
- 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, Dionne Cross Francis, Pamela Lottero-Perdue, Meredith Park Rogers, Calli Shekell, Pavneet Kaur Bharaj, Heather Howell, Adam Maltese, Meredith Thompson, and Justin Reich. 2025a. [Promoting preservice teachers’ facilitation of argumentation in mathematics and science through digital simulations](#). *Teaching and Teacher Education*, 154:104858.
- Jamie Mikeska, Heather Howell, and Devon Kinsey. 2023a. [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 Mikeska, Beata Beigman Klebanov, Aakanksha Bhatia, Shreyashi Halder, and Michael Suhan. 2025b. [Evaluating the use of generative artificial intelligence to support learning opportunities for teachers to practice engaging in key instructional skills](#). In *Artificial Intelligence in Education*, pages 378–391, Cham. Springer Nature Switzerland.
- Jamie Mikeska, Jonathan Steinberg, Pamela Lottero-Perdue, Dante Cisterna, Devon Kinsey, and Heather Howell. 2023b. [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.

- 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 *Proceedings of the 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.
- John Pecore, Corey Nagle, Tadlee Welty, Minkyong 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.
- Yi Qi and Gary Sykes. 2016. [Eliciting student thinking: Definition, research support, and measurement of the ETS® National Observational Teaching Examination \(NOTE\) Assessment Series](#). *ETS Research Memorandum No. RM-16-06*.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for squad](#). *Preprint*, arXiv:1806.03822.
- Borhan Samei, Haiying Li, Fazel Keshtkar, Vasile Rus, and Arthur C Graesser. 2014. Context-based speech act classification in intelligent tutoring systems. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 236–241. Springer.
- Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. Simulated students in tutoring dialogues: Substance or illusion? *arXiv preprint arXiv:2601.04025*.
- Danner Schlotterbeck, Pablo Uribe, Abelino Jiménez, Roberto Araya, Johan van der Molen Moris, and Daniela Caballero. 2021. Tarta: Teacher activity recognizer from transcriptions and audio. In *Artificial Intelligence in Education*, pages 369–380, Cham. Springer International Publishing.
- Meghan Shaughnessy and Timothy Boerst. 2018. Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student’s thinking. *Journal of Teacher Education*, 69(1):40–55.
- Taekwon Son. 2025. Noticing classes of preservice teachers: Relations to teaching moves through AI chatbot simulation. *Education and Information Technologies*, 30(7):9161–9184.
- Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4654–4662.
- 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.
- Deliang Wang, Yaqian Zheng, Jinjiang Li, and Gaowei Chen. 2025. Parameter-efficiently fine-tuning large language models for classroom dialogue analysis. *IEEE Transactions on Learning Technologies*, 18:542–555.
- 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.
- Liwei Wei, Karen 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.
- 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*.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). *Preprint*, arXiv:1704.05426.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Hugging-face’s transformers: State-of-the-art natural language processing](#). *Preprint*, arXiv:1910.03771.
- Dawn Woods, Linda Doornbos, and Cynthia Carver. 2024. Using mixed reality simulations to prepare justice-oriented teachers. *Contemporary Issues in Technology and Teacher Education*, 24(1):95–124.
- 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.

Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao, and Xuming Hu. 2025. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs? In *Findings of the Association for Computational Linguistics: EMNLP*, page 18980–18989. Association for Computational Linguistics.

Nuodi Zhang, Fengfeng Ke, Chih-Pu Dai, Sherry Southerland, and Alex Barrett. 2026. Science and mathematics preservice teachers’ perceptions and experiences of practicing dialogic teaching in generative AI-powered virtual reality simulation. *Teaching and Teacher Education*, 171:105349.

Yuling Zhuang and Si Zhang. 2025. Pre-service mathematics teachers’ perceptions of using GenAI for practicing teacher questioning: A semester-long study. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(9):em2689.

Liu Ziyang, Hu Yongchun, and Zhang Qiaoping. 2026. Harnessing artificial intelligence for preservice teachers’ development: A scoping review of applications, benefits, and challenges. *Computers and Education Open*, 10:100330.

A Aggregation Schemes Evaluation

| #1s | Metric | ASI | AQTA | SWE | Av. |
|-----|--------|-----|------|-----|-----|
| 1 | Prec. | .53 | .68 | .65 | .62 |
| | Recall | .78 | .81 | .82 | .80 |
| | F1 | .63 | .74 | .72 | .70 |
| 2 | Prec. | .55 | .70 | .66 | .64 |
| | Recall | .77 | .80 | .80 | .79 |
| | F1 | .64 | .75 | .73 | .71 |
| 3 | Prec. | .57 | .71 | .67 | .65 |
| | Recall | .75 | .80 | .80 | .78 |
| | F1 | .65 | .75 | .73 | .71 |
| 4 | Prec. | .58 | .71 | .67 | .65 |
| | Recall | .72 | .80 | .78 | .77 |
| | F1 | .64 | .75 | .72 | .70 |
| 5 | Prec. | .57 | .73 | .67 | .66 |
| | Recall | .65 | .78 | .75 | .73 |
| | F1 | .60 | .76 | .71 | .69 |

Table 6: Development set performance of aggregation schemes for five FT-Exp models fine-tuned with different seeds. The first column shows the number of class 1 classifications across the five models for the aggregate to be class 1. Thus, models requiring only one class 1 vote have the highest recall (0.80, on average across the three constructs), whereas models requiring full agreement - five class 1 votes - have the highest precision (0.66, on average across the three constructs).

B Feedback Evaluation Module

The feedback module evaluates educator–student conversation transcripts against pedagogical criteria using structured output parsing. A single LLM call is issued per elicitation construct (see Table 1), with the model parameters shown in Table 7.

Table 7: Model parameters for the feedback module

| Parameter | Value |
|-----------------|-----------------------|
| Temperature | 0 |
| Max Tokens | 16,000 |
| Seed | 42 |
| Response Format | Structured (Evidence) |
| Model | GPT-4o (v2024-08-06) |

Note. The Evidence response format is a structured output schema containing a list of evidence sequences, each comprising consecutive conversational turns with role, content, and turn identifier fields. The same model deployment and parameters are used for all constructs.

For each construct, the prompt is sent as a single system message. The prompt follows a shared template: a task description specifying the construct name, a shared math problem context describing the student’s work, construct-specific evidence requirements with examples and non-examples, and a transcript placeholder where the full conversation history is inserted at runtime. The response is parsed into a structured Evidence object via the OpenAI structured output API.

The task description is as follows:

Your task is to collect evidence on whether the educator’s questions, prompts, and tasks are conducive to eliciting student thinking in the following criterion: *[construct name]* Return your answer as an Evidence object with lists of examples from the user (educator) from the transcript. Provide excerpts from chat history to give sufficient context for evidence. Follow the examples closely, they must be nearly identical. Return an empty list if nothing is similar to the given examples. All evidence must be written exactly as is from the transcript. Never invent or copy examples.

The shared math problem context included in each prompt is:

The discussion is about how the student solved a problem by completing the following steps on a math worksheet:

1. The student drew 6 circles representing Mike's crayons while saying out loud what they were drawing. The circles are drawn in the same pattern as the six-face of a six-sided dice.
2. The student drew 2 crayons representing Ann's crayons
3. The student continued counting, starting where they left off with Mike's crayons, and continuing by counting Ann's crayons. They did not start counting from 1. They drew a dot on Mike's 6th crayon, and then dots on Ann's two crayons as they counted.
4. After solving the problem, the student wrote the numbers 6, 2, and 8 in boxes on the worksheet to complete $6 + 2 = 8$.

The construct-specific sections of each prompt are presented in sections B.1- B.3. For evaluation purposes, all and only teacher (User) utterances that are returned as part of the evidence for a particular construct are labeled 1 for that construct.

B.1 Attends to Specific Ideas

Evidence Requirements

- To count as evidence for this criterion, the teacher needs to ask a follow up probing question based on a specific idea that the student shared about their problem solving process or math understanding (could be about the given math problem or new problems that the teacher includes in the conversation). The follow up probing question should require the student to explain why they engage in specific steps or to explain their understanding related to the math problem or concepts.
- If it is the first time that the teacher is asking about something the student did (e.g., Why did you draw three circles with dots in them?), that would not count as evidence. But a follow up question based on the student's initial response (e.g., about the three circles with dots)

would count as evidence. Essentially the student needs to say something about a specific idea before the teacher's follow up prompt can be tagged as evidence for this criterion.

- The follow up prompt does not need to be linked to the immediate preceding utterance. The teacher could be following up on something that the student said earlier in the conversation and that would still count as evidence for this criterion.

Examples of Evidence

Example 1:

User: I see that there are eight crayons in all. Can you tell me why you put a dot in three of the grains.

Assistant: I didn't put a dot in three, I put them on the last three I counted.

User: Okay. So how many crayons did Mike have?

Assistant: 6 crayons.

User: So you drew six grands for Mike? And how many crayons did Anne have?

Assistant: Ann had two crayons.

User: Ann had two crayons. So why did you put a dot in the last three, the two circles?

Assistant: I put a dot on Mike's last crayon and Ann's two crayons when I counted them to make 7 and 8.

Example 2:

User: What are the dots in the middle of the circles? What do those represent?

Assistant: The dots are for counting.

User: So why did you only count three of them?

Example 3:

User: Okay, I see that. And why are there dots in three of the circles?

Assistant: I put dots when I count the crayons.

User: Okay, so why are there not dots in the first five?

Example 4:

User: Well, how do you know that's how many crowns they have altogether?

Assistant: I drew 6 for Mike and 2 for Ann. Then I counted, 6, 7, 8.

User: And why did you count? Six, seven, eight.

The following would NOT count as evidence:

- If the teacher restates the student's idea without probing further (only restates to clarify that is what the student said).

- If the teacher asks the students to recall the steps they used to solve the given addition word problem.

Non-Examples

Non-Example 1:

User: So if you remember, here is where normally. I would have manipulatives. For example, maybe some markers. Let's pretend that these markers are crayons, like the circles you drew. So. You first drew six circles. Like this for? Mike's crayons, right?

Assistant: Yes.

Non-Example 2:

User: You drew two circles next because Ann has 2 crayons, right?

Assistant: Yes!

B.2 Asks Questions Tied to Actions

Evidence Requirements

- To count as evidence for this criterion, the teacher needs to ask questions about what the student did or about why the student did specific things to solve the given addition word problem or to solve new addition or subtraction number sentences or word problems.
- These questions do not need to link to the student's preceding comment or utterance.
- If the teacher asks the student a question to confirm or clarify about what they did (something that the student already mentioned earlier in the conversation), that would count as evidence for this criterion.

Examples of Evidence

Example 1:

User: So why did you draw eight circles?

Assistant: I drew 6 circles for Mike's crayons and 2 for Ann's crayons.

Example 2:

User: I'd really like to learn too. Can you show me how you're working this? Problem. What's the first step?

Assistant: First, I draw circles.

Example 3:

User: Can you explain to me? Why? You did, the steps that you did.

Assistant: I drew circles to show the crayons for Mike and Ann. Then I counted 6, 7, 8.

Example 4:

User: Okay, so I see three circles. With a dot on the inside. I'm wondering what made you draw a third circle.

Assistant: I drew two circles for Ann and put dots as I counted.

Example 5:

User: Why did you choose that strategy?

Assistant: That's how I show the crayons.

The following would NOT count as evidence:

- If the teacher's question focuses on what the student understands and is not related to something specific they did to solve the given addition word problem or new addition or subtraction number sentences or word problems. For example, "What does addition mean?", "What does it mean to count on?"
- If the teacher's question asks the student to answer a factual (recall) question related to the given addition word problem.
- If the teacher's questions asks the student what they would do in solving additional problems or what they would do in the future. For example, gives them a new problem and asks them to solve it.

Non-Examples

Non-Example 1:

User: What does the plus sign mean?

Assistant: The plus sign means you add things together.

Non-Example 2:

User: What do you know about how commutative property can be used to solve addition problems?

Assistant: I'm not sure.

Non-Example 3:

User: How is addition similar to multiplication?

Assistant: I just know addition.

Non-Example 4:

User: How many crayons did Mike have?

Assistant: Mike had 6 crayons.

B.3 Shows Work and Explains

Evidence Requirements

- To count as evidence for this criterion, the teacher needs to prompt the student to reference or discuss one or more aspects of their

written work AND explain what they did in that written work or why the written work is the way it is.

Examples of Evidence

Example 1:

User: Which crayons were Mike? How did you show Mike's crayons?

Assistant: I drew 6 circles for Mike's crayons.

Example 2:

User: Excellent. So tell me, how did you count on from six?

Assistant: I started at 6 and then counted 7, 8.

Example 3:

User: So what was the first thing you did to solve the problem?

Assistant: I drew 6 circles for Mike's crayons.

The following would NOT count as evidence:

- If the teacher restates an idea that the student already shared.
- If the teacher asks the students to recall the steps they used to solve the given addition word problem
- If the teacher's question asks the student what they would do in solving additional problems or what they would do in the future (e.g., gives them a new problem and asks them to solve it)

Non-Examples

Non-Example 1:

User: Now I can understand. You started at six, and then you added two more, so that's how you got to. Eight. Is that right?

Assistant: Yes.

Non-Example 2:

User: So if you remember, here is where normally. I would have manipulatives. For example, maybe some markers. Let's pretend that these markers are crayons, like the circles you drew. So. You first drew six circles. Like this for? Mike's crayons, right?

Assistant: Yes.