

# Automated Evaluation of Teacher Encouragement of Student-to-Student Interactions in a Simulated Classroom Discussion

Michael John Ilagan

McGill University

michael.ilagan@mail.mcgill.ca

Beata Beigman Klebanov

ETS Research Institute

bbeigmanklebanov@ets.org

Jamie N. Mikeska

ETS Research Institute

jmikeska@ets.org


## Abstract

Leading students to engage in argumentation-focused discussions is a challenge for elementary school teachers, as doing so requires facilitating group discussions with student-to-student interaction. The Mystery Powder (MP) Task was designed to be used in online simulated classrooms to develop teachers' skill in facilitating small group science discussions. In order to provide timely and scaleable feedback to teachers facilitating a discussion in the simulated classroom, we employ a hybrid modeling approach that successfully combines fine-tuned large language models with features capturing important elements of the discourse dynamic to evaluate MP discussion transcripts. To our knowledge, this is the first application of a hybrid model to automate evaluation of teacher discourse.

## 1 Introduction

Scientific argumentation is an essential skill, and in elementary school classrooms, group science discussions are a natural modality for providing students with opportunities to engage in scientific argumentation (Sampson and Blanchard, 2012; Shemwell and Furtak, 2010). Accordingly, it is essential that teachers are well equipped to facilitate such discussions. But facilitating them is not straightforward. Many teachers are used to a lecture style of interaction where they deliver the facts and the students respond only to the teacher (Cazden, 1988; Lemke, 1990; Lloyd et al., 2016). In contrast, in an ideal group science discussion, students directly interact with their peers (rather than just the teacher) and engage with each other's ideas, rather than only their own and the teacher's (Fishman et al., 2017; Tenenbaum et al., 2020).

Digitally simulated classroom experiences have become increasingly used to prepare teachers for the work of teaching (Dalinger et al., 2020; Dieker et al., 2014). In a simulated classroom, the teacher-



Teacher	How did you find that?
Carlos	Well, I looked at the properties one at a time, except for the weight, and I narrowed it down that way.
Teacher	Okay. So Jayla, Emily, or Carlos, do you have any questions for Mina or Will on the ways that they found their answer?
Emily	Well, yeah. Well, what properties did you look?
Will	Well, we looked at the texture and the color, and then we measured the weight. And those all matched flour.

Figure 1: The Mursion Upper Elementary Classroom Environment, with an excerpt from a Mystery Powder discussion transcript. Two blocks of utterances (explained in section 4.2) are shown in blue and orange, respectively. Image provided by Mursion, Inc.

in-training, also called a *pre-service teacher* (henceforth, **PST**), enacts a classroom scenario, interacting in real time with student avatars puppeteered by a trained human actor equipped with voice modulating software. In contrast to practicum experiences, simulated classrooms afford development of targeted skills in an environment that is both standardized and low-stakes (Dalinger et al., 2020; Bondie et al., 2021; Cohen et al., 2020; Ersozlu et al., 2021). Automating as much as possible of the simulation would help make the learning experience more affordable and thus accessible to a wider range of teachers; it would also allow teachers to engage in multiple rounds of practice to hone their teaching skills. Of the two bottlenecks—the puppeteer enacting the student avatars and the human expert evaluating the performance—we here address the second, leaving the first to future work.

The present paper is a case study of developing automated evaluation, with supervised learning, of a PST's performance in a simulated classroom. We focus on the Mystery Powder (henceforth, **MP**)

task (Mikeska et al., 2021), a particular lesson that the PST is to teach in a simulated classroom (Figure 1) designed to develop PSTs’ competency in facilitating small group argumentation-focused science discussions at the elementary level. Successful facilitation of a discussion is complex; in this work, we address one of its dimensions, namely, the extent to which the teacher encourages student-to-student interactions where students engage directly with each other’s ideas (Mikeska et al., 2021; GO Discuss Project, 2021).

In line with the manual evaluation process that produced the training data (Mikeska et al., 2019), our approach to automated evaluation had models on two levels: classifiers identifying PST utterances as positive examples of the desired teaching practices; and regressors scoring the transcript as a whole on the same practices (Nazaretsky et al., 2023). Furthermore, we kept in mind two considerations: classifier training must deal with the fact that rater labels were non-exhaustive (only some utterances are labeled); and regressors must aggregate utterance-level information in an intuitive way.

In terms of what features were used, we built three types of models: (a) models based on the analysis of the content of what the PST said, implemented using fine-tuned large language models (henceforth, LLMs); (b) models based on the structure of the interaction, who speaks when and in relation to whose utterance; and (c) combined models using both content and structure. To our knowledge, this is the first demonstration, in the context of automated analysis of teacher discourse, of a successful combination of fine-tuned LLMs and shallow features into a hybrid model that outperforms both components in isolation across the board, for multiple levels of analysis (utterance-level and transcript-level) and multiple indicators of performance.

## 2 Related Work

### 2.1 Elements of high-quality teaching practices

Recent research has addressed automated detection of high-quality teaching practices in human-annotated corpora of real classroom transcripts. Demszky and colleagues (Demszky et al., 2021; Demszky and Hill, 2023; Alic et al., 2022) detected features associated with dialogic instruction, such as teachers’ conversational uptake (Demszky et al.,

2021) and open-ended questions (Alic et al., 2022), which they found to benefit classroom outcomes such as student satisfaction and participation. Similar discourse features were investigated in Jensen et al. (2020), as part of an effort to bring easy-to-use and high-quality audio recording setups to ordinary classrooms. Suresh and colleagues (Suresh et al., 2019, 2022b) performed a six-way classification of teacher utterances into discursive strategies, called “talk moves” (e.g. “Keeping everyone together”), that promote equitable student participation. Tran et al. (2023) classified student and teacher contributions into ‘talk moves’ such as ‘teacher links student contributions’ and ‘students support claims with evidence’. Nazaretsky et al. (2023) studied ways to evaluate to what extent participants provided meaningful contributions that moved the discussion forward. Most of the prior work, with few exceptions such as Nazaretsky et al. (2023), considered transcripts of live interactions; simulated environments with student avatars aim to extend the practice earlier into the teacher preparation process, before the teacher meets a real classroom (Dalinger et al., 2020). Our work is in the much less explored context of a simulated classroom.

A common theme in research on automated models for high-quality teaching practice is the intended application to providing automated feedback to teachers. Feedback may come in the form of a dashboard summarizing the teacher’s performance. The dashboard may report the (relative) frequency of the target discourse features (Demszky et al., 2023; Jensen et al., 2020). The dashboard may also cite “positive examples” among the teacher’s own utterances to reinforce productive teaching practices (Demszky et al., 2023; Jensen et al., 2020; Nazaretsky et al., 2023). The efficacy of such automated feedback for benefiting classroom outcomes (e.g. proportion of assignments completed by the student) has been demonstrated in a setting with 1:10 teacher-student ratio (Demszky et al., 2023) as well for 1:1 mentoring (Demszky and Liu, 2023).

### 2.2 Modeling Approaches

In terms of modeling approaches, prior work explored pre-trained deep neural embeddings to represent the content of an utterance and either use them directly as features for detecting teachers’ discourse moves of interest (Suresh et al., 2019) or to derive features such as similarity scores between

neighboring teacher and student utterances when modeling uptake (Demšky et al., 2021). Demšky et al. (2021) reported that simpler lexical features quantifying token overlap between student and teacher words were also competitive. Jensen et al. (2020) used a combination of linguistic features such as parts of speech and markers of comparisons or definitions along with features capturing other characteristics of the teacher-student interaction, including utterance length and its normalized position in the session, rate of speech and pauses, in a supervised machine learning setting.

Fine-tuning an LLM-based classifier for the target data and task was also explored. Jensen et al. (2021) found the performance of a BERT-based classifier to be superior to that of feature-based baselines on data of self-recorded classroom interactions from English Language Arts teachers. Nazaretsky et al. (2023) fine-tuned DistilBERT (Sanh et al., 2020) on simulated classroom data in the science domain. Suresh et al. (2021) explored BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) to classify student and teacher utterances into ‘talk moves’ in the domain of mathematics. Tran et al. (2023) used a sequence model BiLSTM-CRF (Huang et al., 2015) with BERT embeddings to classify utterances into a somewhat different set of ‘talk moves’ in the domain of English Language Arts and showed that the sequence model that takes into account neighboring utterances outperformed the BERT-based models that did not utilize sequence information, for detecting some of the talk moves. Suresh et al. (2022b) explored incorporating information from the discourse outside of the teacher’s and neighboring student utterance, showing that taking a much larger discourse context into account helps improve performance; the best-performing models extended the context to seven prior and seven subsequent utterances. Kumaran et al. (2023) explored fine-tuning of DialoGPT (Zhang et al., 2020), a dialog LLM built on GPT-2 (Radford et al., 2019) on the student subset of the ‘talk moves’ data (Suresh et al., 2022a), utilizing a context of nine prior utterances. These approaches tend to include elements of the larger discourse context through incorporation of larger and larger chunks of prior and/or subsequent content into the LLM-based framework.

In the present study, we explore an approach that models discourse dynamics more directly through a set of features that would be used in tandem with

the fine-tuned LLM to provide the overall model with information about relevant aspects of the structure of the discourse. Such hybrid models can also provide some insights into the task, by separating the contribution of the fine-tuned LLM based content models from that of the discourse-dynamic-based model; different aspects may be more or less important for modeling different components of the complex performance task set to the teachers.

### 3 The Mystery Powder Task

#### 3.1 The performance

In the MP task, the PST interacts with five upper elementary student avatars in the simulated classroom (Figure 1). Each avatar is standardized, in terms of their personality (e.g. Will is soft-spoken) and preconceptions related to the MP task (explained below). The human actor, who puppeteers all five avatars, is well-versed in them and is instructed to ensure that they are responsive to the PST’s instructions throughout the discussion.

The scenario is as follows. Prior to the discussion, the class was shown samples of six powders: flour, cornstarch, baking soda, baking powder, sugar, and salt. The class investigated several properties of each sample including texture, color, weight, reaction with vinegar, and outcome when mixed with water. Subsequently, the class was presented a “mystery powder” sample—in fact baking soda, unbeknownst to the students—and the students investigated its properties as well. In small groups, as pre-work to the discussion, the students reflected in writing on their findings and generated evidence-based claims about the mystery powder’s identity and the properties that were useful to identify the mystery powder. See Appendix for a reference table for the powders (Figure 6) and an excerpt from one of the group’s pre-work (Figure 7).

The PST has up to 20 minutes to facilitate a discussion to help the five students arrive at a consensus regarding (1) the identity of the mystery powder, and (2) which properties are important for this identification. As preparation, the PST has access to the students’ written reflections and is provided information about the accuracy of their initial ideas. For instance, the PST must ensure that the discussion rectifies the misconception (held by Mina, Will, Jayla, and Emily) that weight is important for identifying the MP. See Figure 1 for an excerpt from a discussion’s transcript.

### Dimension 3: Encouraging Student-to-Student Interactions

Indicator title	Level 1 Beginning practice	Level 2 Developing practice	Level 3 Well-prepared practice
3a. Peer interaction	The teacher assumes the responsibility for the discussion by rarely promoting peer interaction AND frequently mediates all student contributions.	The teacher occasionally promotes peer interaction, AND the majority of student contributions are mediated through the teacher.	The teacher frequently promotes peer interaction, AND the mediation of student contributions is shared between the teacher and the students.
3b. Engagement with others' ideas	The teacher rarely encourages students to engage with one another's ideas, conceptions, or viewpoints.	The teacher occasionally encourages students to engage with one another's ideas, conceptions, or viewpoints.	The teacher frequently encourages students to engage with one another's ideas, conceptions, or viewpoints.

Table 1: Rubrics for Indicators 3A and 3B (Mikeska et al., 2021).

### 3.2 Rubric and manual evaluation

The MP rubric is made up of several dimension scores, each of which is supported by several more specific indicator scores (Mikeska et al., 2021). The present study focuses on Dimension 3 (“Encouraging student-to-student interactions”) and two of its indicators, Indicator 3A (“Peer interaction”) and Indicator 3B (“Engagement with other’s ideas”).<sup>1</sup> See Table 1 for Indicator score definitions.

After evaluation, the PST expects to see a feedback report that tells their strengths, areas for growth, and recommended next steps in each Dimension. This report must give not only an overall (i.e. transcript-level) evaluation but also supporting evidence (i.e. utterance-level) to reinforce the PST’s desirable practices.

Accordingly, manual evaluation occurs on two levels. First, the human rater cites, for each Indicator, one or more utterances that exemplify the target behavior (positive examples) or its opposite (negative examples). Note that the rater is asked only to provide some examples, not exhaustively label every utterance in the transcript. Second, the human rater scores the transcript, continuous on a scale of 1 to 3 (e.g. a score of 1.40 is possible) on each Indicator and then an integer from 1 to 3 for each Dimension. To calibrate judgments, raters undergo extensive training, which includes completing self-guided webinars and evaluating sample discussions.

### 3.3 Automated evaluation approach

Automated evaluation aims to follow the same two-level process, via classifiers (for utterances) and regressors (for transcripts). Conceptually, regressor

<sup>1</sup>Dimension 3 has a third indicator, “Ideas come from students”, not within the scope of the present study.

features are aggregates of utterance-level information, which include utterance class labels. However, ground-truth labels are not available for new transcripts, so aggregating them is infeasible. Instead, in our approach, after training on the labeled utterances, a classifier predicts positive probabilities for all utterances, labeled and unlabeled. It is then these predicted probabilities that are aggregated into transcript-level features (described in section 4.2). Thus, classifier training and evaluation uses ground-truth labels, for the subset of utterances they are available; but regressor training and evaluation uses only imputed probabilities.

## 4 Data, models, and features

### 4.1 Data

The MP dataset was collected in prior work (Mikeska et al., 2019).<sup>2</sup> A total of 79 PSTs facilitated discussions: 76 engaged in the simulation twice; 3 engaged once. Of the 155 transcripts, 56 were coded by two raters. Reliability was measured with intra-class correlation coefficients (ICCs) and was sufficient (Cicchetti, 1994) for all three constructs: 0.816 for Indicator 3A; 0.679 for Indicator 3B; 0.635 for Dimension 3. For transcripts scored by two raters, the final scores were the average between the raters—thus non-integer scores are also possible for Dimension 3. The MP dataset has a total of 14,558 utterances. For PSTs (6,713 utterances), the interquartile range for utterance length was 8 to 30 tokens; for students (7,845 utterances), it was 4 to 20 tokens. Distributions of transcript-

<sup>2</sup>The Mystery Powder discussion and scoring data used in this study was collected and generated as part of previous grants funded by the National Science Foundation (grant #1621344, #2032179, and #2037983). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

level scores are in Figure 2.

Allocation of transcripts into train and test partitions was done by PST, so that PSTs in the training data would not be again seen in the test set (Nazaretsky et al., 2023). 121 transcripts (from 62 PSTs) were allocated to the training set and 34 transcripts to the test set. For utterance-level analyses, each utterance was allocated to the same partition (train or test) as its parent transcript.

	Indicator 3A		Indicator 3B	
	Train	Test	Train	Test
Class 0	1411	668	558	179
Class 1	267	86	426	144
(unlabeled)	3496	785	4190	1216

Table 2: Breakdown of PST utterances by class label, construct, and train/test.

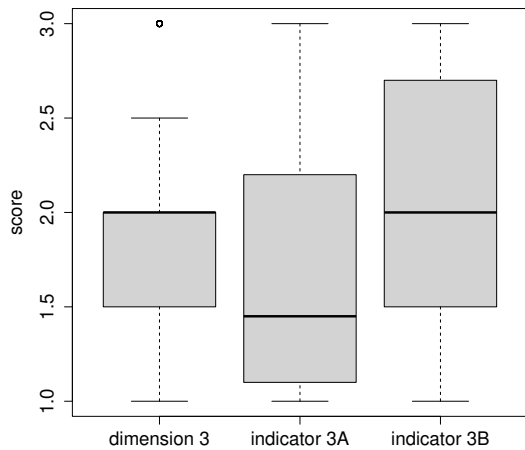


Figure 2: Distributions of transcript scores (training set).

Raters’ citations of positive examples were in free-form text, which was manually coded by the first author. The test set was coded after model selection on the training set. Rater comments were not always timestamps or direct quotes, so some judgment was exercised. The following rules were applied:

1. A PST utterance is labeled “1” (positive) if at least one rater cited it as a positive example.
2. If a PST utterance is not labeled “1”, then it is labeled “0” (nonpositive) if at least one rater

indicated that the transcript had no positive examples in it.

3. If a PST utterance is not labeled “1”, then it is labeled “0” (nonpositive) if at least one rater indicated that it was a negative example. Since there were only a few negative examples, they were not assigned their own class.
4. If a PST utterance cannot be labeled either “0” or “1” due to the above rules, it is left unlabeled—excluded from training and evaluation of the utterance-level classifiers.

Note that for training and evaluation of classifiers, only the manually-labeled PST utterances are used. But for training and evaluation of regressors, all PST utterances are used, as predicted probabilities are used instead of ground-truth labels.

Since PST performance is the focus of the study, student utterances were used only to generate features pertaining to the adjacent PST utterances, following a process explained in section 4.2. See Table 2 for breakdown of the PST utterance labels in the dataset. Only a small proportion of the utterances are positive examples (in the training set, 5% for Indicator 3A and 8% for Indicator 3B).

## 4.2 Models and features

As we inspected rater justifications and rubric definitions, we decided to hand-craft a number of features as well as leverage neural language models shown to be useful in prior work on teacher discourse analysis (see Section 2). In all, we considered 15 models, summarized in Table 3. Models vary along three factors — level of analysis, target construct, and type of features, as follows:

- A model is either (**C**) an utterance level classifier, or (**R**) a transcript-level regressor.
- A model is concerned with (**A**) Indicator 3A, (**B**) Indicator 3B, or (**D**) Dimension 3.
- A model is (**N**) content-based (via fine-tuning an LLM), (**S**) structure-based (via handcrafted features, some of which involve using LLMs out of the box), or (**X**) a combination of both.

Note that only Indicators have utterance-level analysis, so there are no classifiers for Dimension 3. Also note that models “compete” only in the same cell (e.g. CAN vs. CAS vs. CAX).

Content-only classifiers (CAN and CBN) were constructed by adding a linear classifier head on

	(A) Indicator 3A	(B) Indicator 3B	(D) Dimension 3
(C) Utterance-level classifier	<b>CAN</b> : content only <b>CAS</b> : structure only <b>CAX</b> : combined	<b>CBN</b> : content only <b>CBS</b> : structure only <b>CBX</b> : combined	(none)
(R) Transcript-level regressor	<b>RAN</b> : content only <b>RAS</b> : structure only <b>RAX</b> : combined	<b>RBN</b> : content only <b>RBS</b> : structure only <b>RBX</b> : combined	<b>RDN</b> : content only <b>RDS</b> : structure only <b>RDX</b> : combined

Table 3: All models. See the beginning of section 4.2 for an explanation of the rows and columns.

top of DistilBERT (Sanh et al., 2020) (66M parameters) using the HuggingFace toolkit (Wolf et al., 2020). DistilBERT is a lightweight model that has been used in educational settings (Nazaretsky et al., 2023; Datta et al., 2023; Butt et al., 2022; Pearce et al., 2023). Embedding and transformer layers were frozen. Training was done with learning rate 0.001, batch size 32, and a linear scheduler with no warmup. The number of epochs (between 1 and 10) was a hyperparameter. As inputs to DistilBERT, each utterance was prepended by the speaker (e.g. “Carlos”), and the context for each PST utterance was the student utterance immediately following the teacher’s in the transcript. The intuition is that how students respond is potentially informative for whether the PST utterance is positive or not.

Unlike fine-tuning an LLM, which leverages utterance content, classifiers with handcrafted features mostly use turn-taking dynamics, that is, the structure of the interaction. Utterances (student and PST) are organized in blocks. Each PST utterance begins a block, which spans the subsequent student utterances until the next PST utterance. Figure 1 shows two color-coded blocks of utterances. By computing features per block, features associated with a PST utterance incorporate the turn-taking structure in the subsequent student utterances.

For the structure-only classifier for Indicator 3A (CAS), the following four features were computed

Classifiers	Regressors
(LR) Logistic regression	(LR) Linear regression
(DT) Decision tree	(BR) Bayesian ridge regression
(MP) Multilayer perceptron	(DT) Decision tree
(RF) Random forest	(MP) Multilayer perceptron
	(RF) Random forest

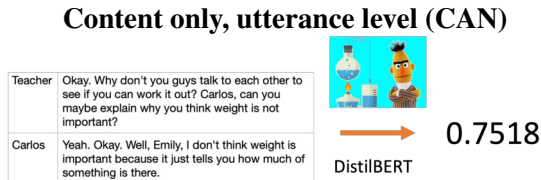
Table 4: Classifiers and regressors to choose from.

per PST utterance based on its block:

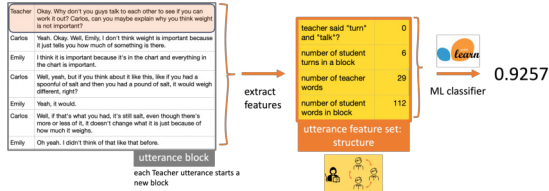
- NUM\_STUDTURNS: Number of student utterances in the block.
- NUM\_TEACHTOKS: Number of tokens in the PST utterance itself.
- NUM\_STUDTOKS: Number of tokens in the students’ utterances in the block.
- NUM\_KW1: “1” if the tokens “turn” and “talk” both appear in the PST utterance; “1” if the token “crosstalk” appears in the PST utterance; and “0” otherwise. (“Turn and talk” is the name of a commonly-used instructional technique where students are put in pairs to discuss an issue (Hindman et al., 2022). In the case of the MP discussion, when this occurs, the avatars produce mumbling sounds often denoted in the transcript as “crosstalk”.)

For the structure-only classifier for Indicator 3B (CBS), the handcrafted features capture student-to-student uptake. Each student utterance  $u_1$  is paired with the previous student utterance  $u_0$  in the transcript. For every such pairing, the following five features are computed:

- PROP\_IN\_LEFT: Proportion of tokens in  $u_0$  also found in  $u_1$ , range: [0,1].
- PROP\_IN\_RIGHT: Proportion of tokens in  $u_1$  also found in  $u_0$ , range: [0,1].
- JACCARD: Jaccard coefficient between the two sets of tokens, range: [0,1].
- BLEU: BLEU (Papineni et al., 2002) score for reference  $u_0$  and hypothesis  $u_1$ , range: [0,1].
- SENTBERT: Cosine similarity between the sentence-BERT (Reimers and Gurevych, 2019) embeddings of  $u_0$  and  $u_1$ , range: [-1,1].



**Structure only, utterance level (CAS)**



**Content+structure, utterance level (CAX)**

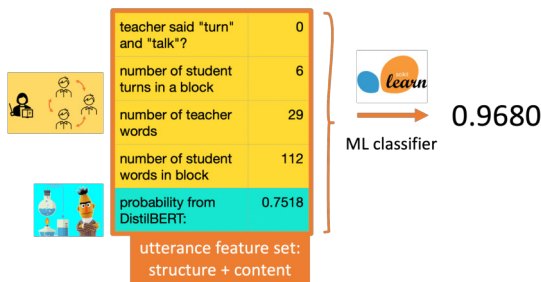


Figure 3: Illustration of utterance-level modeling in Indicator 3A, for a single PST utterance. Refer to Table 3 for model acronyms. Indicator 3B models (CBN, CBS, CBX) proceed analogously. Structure features are highlighted in yellow; content feature is highlighted in turquoise.

Snowball stemming, as implemented in NLTK (Loper and Bird, 2002), was used prior to computing word overlap. Each pair yields a 5-dimensional feature vector. The feature vector of a PST utterance is the mean-aggregate vector using the pairs in its block, skipping over utterances with fewer than 5 tokens. For PST utterances with no eligible student utterances in the block, we use the lowest possible value of the feature (e.g. 0 for JACCARD).

Combined classifiers (CAX and CBX) used both types of features. For Indicator 3A (CAX), the features were all the structure-only features (e.g. NUM\_TEACHTOKS from CAS) as well as the DistilBERT-predicted positive probability (from CAN). Indicator 3B (CBX) followed analogously. Figure 3 is a cartoon summarizing which features appear in which classifier.

For structure-only classifiers (CAS and CBS) and combined classifiers (CAX and CBX), we used shallow learning models as implemented in the Scikit-learn toolkit (Buitinck et al., 2013). See Table 4 for the classifiers considered and Table 7

**Content+structure, transcript level (RAX)**

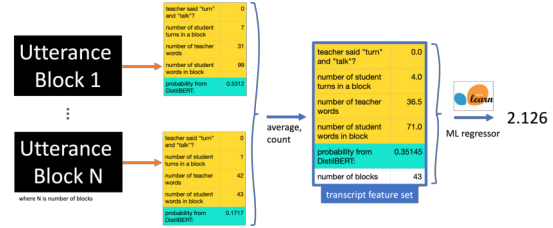


Figure 4: Illustration of the transcript-level combined model on Indicator 3A, for a single PST utterance. Refer to Table 3 for model acronyms. The Indicator 3B model (RBX) proceeds analogously.

(in the Appendix) for the hyperparameter grid.

At the transcript level, Indicator regressor features are constructed by simple aggregates of utterance-level information. For content-only Indicator regressors (RAN and RBN), there are only two features: the relevant average DistilBERT-predicted probability (from CAN or CBN); and the count of PST utterances (or utterance blocks). For structure-only Indicator regressors (RAS and RBS), the features are the averages of the relevant structure-only features (e.g. NUM\_TEACHTOKS from CAS, or JACCARD from CBS) and the count of PST utterances. For combined Indicator regressors (RAX and RBX), the features are the averages from both types of features and the count of PST utterances. Figure 4 in the Appendix is a cartoon summarizing how utterances are aggregated. See Table 4 for the regressors considered. See Table 8 (in the Appendix) for the hyperparameter grid.

As for Dimension 3 regressors (RDS, RDN, and RDX), features are simply the union of the features of the Indicator regressors. RDN inherits features from RAN and RBN; RDS inherits features from RAS and RBS; and RDX inherits features from RAX and RBX.

All experiments were carried out on a MacBook Pro laptop, with Apple M1 Pro chip. Computations did not use GPU.

**4.3 Model selection and evaluation**

For model selection, we performed a 5-fold cross-validation (CV) on the training set. Folds were split by PST, as described for the train/test partition (section 4.1).

For classifiers, the metric for model selection was  $\kappa$  (Cohen, 1960); higher values are better. For regressors, the metric was mean squared error (MSE); lower values are better. Since manual scores range from 1 to 3, the predicted score was

truncated to this range. For each of the 15 models in Table 3, the final number of epochs (for LLMs) or final estimator (for shallow learning models) was selected using cross-validation in order to advance to test set evaluation. For choosing the numbers of epochs, the one-standard-error rule (Hastie et al., 2017) was used.

Models that used the predicted probability from DistilBERT as feature (i.e. all except CAS, CBS, RAS, RBS, and RDS) used the best-performing number of epochs from the corresponding fine-tuned LLM classifier.

## 5 Results

Table 5 shows the test set results for classifiers. For Indicator 3A, structure-based models dominated content-based models. For Indicator 3B, the trend was the opposite. For both Indicators, the combined models had the best performance. See Appendix for examples of positive-predicted utterances.

Table 6 shows the test set results for regressors. For Indicator 3A, structure and content models show similar performance. For Indicator 3B, the fine-tuned LLM dominated. For both Indicators, as well as the Dimension 3 score, the combined models showed the best performance.

## 6 Discussion

### 6.1 Modeling approach

Our results show that classifiers focused on the content of the PST utterance perform better for Indicator 3B, while those focused on the structure of the discourse perform better for Indicator 3A. Thus, the results suggest that it is quite difficult to get out of the *content* of a PST's utterance whether or not the utterance encouraged peer interaction. However, since the simulated students (a) do not tend to spontaneously engage in a multi-party discussion, yet (b) are compliant with the teacher's instructions, whether or not multiple students speak following the teacher is a fairly strong signal of whether the teacher encouraged them to do so.

In contrast, whether or not the teacher encouraged the students to engage with each others' ideas is easier to recover from the actual PST utterance than from evidence of lexical overlap or semantic similarity between subsequent student utterances. This may be because, given the highly constrained topic of the conversation (properties of the six powers), on the one hand, consecutive student utterances generally tend to have substantial textual

overlap, whether or not the teacher encouraged that; on the other hand, overlap or semantic similarity as captured in pre-trained models may not be sufficiently nuanced to distinguish between actual uptake and mere accidental, topic-induced, semantic similarity or lexical overlap.

We observe that modeling the discourse dynamic explicitly and separately from the fine-tuned-LLM-based model of the content yields more explainable models than models where the content of a large surrounding context is used within the LLM-based model. Thus, our design and results allow us to see clearly the extent to which the fact of the within-block students' utterances, irrespective of what is said, can predict the score on Indicator 3A, as well as to observe the complementarity of the content and structure as sources of information.

### 6.2 Generalization based on select examples

We observed previously that the design of the human evaluation campaign conducted prior and independently from the computational modeling was such that raters were asked to provide justifications for their scores in the form of specific utterances that could serve as positive examples of the target behavior; only 5–8% of the PST utterances were picked as positive examples. The general prevalence of utterances that exhibit the target behavior was not known a-priori, nor was it obvious that better performance, based on holistic scores, would clearly correspond to having more utterances that exhibit such behaviors.

Figure 5 shows boxplots of the proportion of automatically predicted positive examples for either Indicator by human-assigned holistic proficiency levels according to Dimension 3 scores. First, we observe that the system was able to detect many more positive examples than were provided – even at the lowest level of performance, most PSTs exhibited the target behavior in more than 10% of their utterances, while most of the best-performing PSTs did it in more than 40% of theirs.

Second, we observe a strong differentiation between proficiency levels – boxes containing middle 50% of the performances per level have almost no overlap. This provides validity evidence not only for the automated modeling but for the human holistic scores as well, showing that they correspond to explicit, quantifiable transcript-level aggregation of relevant evidence.

Third, the emergent differentiation enables easily



Construct	Model	Number of epochs or estimator	Accuracy	Cohen $\kappa$	F1
Indicator 3A	CAN	3 epochs	0.899	0.283	0.321
	CAS	MP	0.924	0.604	0.646
	CAX	RF	0.931	0.641	0.679
Indicator 3B	CBN	6 epochs	0.774	0.531	0.711
	CBS	RF	0.632	0.260	0.602
	CBX	MP	0.793	0.571	0.739

Table 5: Test set evaluation results for utterance-level classifiers.

Construct	Model	Estimator	MSE	Pearson correlation
Indicator 3A	RAN	MP	0.343	0.468
	RAS	RF	0.354	0.480
	RAX	RF	0.335	0.513
Indicator 3B	RBN	LR	0.238	0.705
	RBS	MP	0.325	0.530
	RBX	LR	0.215	0.724
Dimension 3	RDN	BR	0.242	0.547
	RDS	MP	0.202	0.631
	RDX	BR	0.183	0.678

Table 6: Test set results for transcript-level regressors. Lower MSE is better; higher correlation is better.

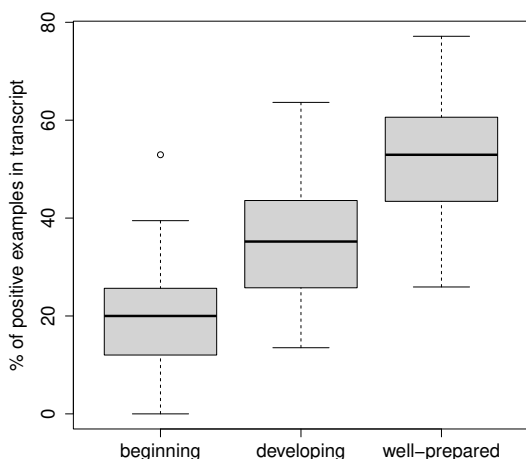


Figure 5: Boxplot of human-assigned Dimension 3 proficiency level vs. percentage of model-predicted positive examples (either Indicator) in transcript.

explainable and visually clear feedback whereby a PST’s performance could be mapped against teachers at various levels of proficiency, to communicate not only current performance level but also how much more frequently one needs to im-

plement the target behavior in order to move to the next level. Taken together, our results suggest that having humans provide select evidence for the score could be a viable alternative to a more comprehensive utterance-level annotation that is the prevalent approach in the literature on automation of the detailed evaluation of teacher discourse.

## 7 Conclusion

The goal of the current study was automated evaluation of teacher discourse when facilitating a discussion in a simulated elementary science classroom. We showed that models focused on the content of the teacher’s utterances using fine-tuned large language models and models focused on the structure of the discourse modeled using handcrafted features captured complementary aspects of the target construct and could be fruitfully combined into hybrid models that outperformed both content and structure models. Our results also demonstrated strong generalization from a small number of “score justifications” provided by expert human raters, suggesting a potentially more efficient data generation paradigm than an exhaustive annotation of discourse moves.

## 8 Limitations

A limitation of the current study is the use of only one scenario for a simulated discussion, namely, the Mystery Powder task for an elementary science classroom and so it is not clear to what extent the type of models discussed in this paper will generalize to other scenarios. To address this limitation, we are developing additional scenarios, collecting discussion transcripts, and conducting human evaluations to generate data for additional studies that would examine the generalization of the technique proposed in this paper to new scenarios in both science and mathematics contexts.

Another limitation is that the current data come from pre-service teachers only; an online simulation could also be useful for early career in-service teachers. We are in the process of collecting data from in-service teachers and will be able to examine generalization to a different user population as the project progresses and more data become available for computational analysis.

Our experiments did not vary the size of the context window for DistilBERT. In line with [Suresh et al. \(2022b\)](#), it is possible that larger windows might substantially improve the performance of the fine-tuned-LLM-based models. That said, larger windows can potentially “encroach” on the structure-based models’ territory making the distinction between what is due to the structure and what is due to the content harder to maintain, and with it, the explainability that comes from being able to point to the distinct aspects of the simulated discussions as information sources for the models. The explainability of the models is important not only for the PST buy-in, but also for the interdisciplinary team that is working on creating feedback reports based on the models’ output. An explanation connecting the focus of the rubric to the performance of models with different types of information, as in section 6.1, helps the science teacher educators on the team appreciate the alignment between the rubric and the automated models.

Another limitation of the current study is using only DistilBERT. This model was picked for its efficiency and prior successful use in educational settings ([Butt et al., 2022](#); [Pearce et al., 2023](#)); however, larger and more powerful models may support stronger performance, especially for Indicator 3A, where there is substantial room for improvement, with the current best performance of  $r = 0.513$ . Having established the baselines in this study, we

intend to explore additional LLMs, resources permitting.

The data used in the study comes from predominantly White and female PSTs, reflecting the demographic at the data collection sites and in the teacher population in the USA. In the ongoing data collection, we are making an effort to reach out to more diverse demographics. Demographic information about the expert raters who provided scores and justifications was not collected; this will be rectified in future studies.

All current data come from pre-service teachers in the USA and all simulated discussions are conducted in English. In principle, actors who speak other languages could be trained to provide online practice to pre-service teachers in other cultural and linguistic environments; however, the detail and nuance of culturally appropriate teacher-student and student-student interactions might differ. At the moment, the scope and funding of the ongoing project do not allow addressing this limitation.

## 9 Ethics statement

The transcripts, scores and score justifications used in this study were collected with the approval of our Institutional Review Board with informed consent of the participants as part of previous studies. Participants were provided information about the purpose of the study, the risks and benefits to participating in the study, and details about what participation entailed. The raters were paid for the time they contributed to generating the scores and score justifications and the PSTs were paid for being research participants. The PSTs were enrolled in an elementary methods course at their university and were recruited based on their professor’s participation in the study. Each PST could voluntarily consent to participate (or not) in the research study to have their transcript data used for research purposes. The consent form for participants included the following statement about risks: “Some participants may experience a small degree of discomfort when facilitating the discussions in the simulated classroom environments.” All transcript data is de-identified, and a PST is represented by a numerical ID in each transcript. The data does not contain offensive content. The collected data is used in compliance with the consent. The consent form contained an explanation of the intended use: “The video recordings and transcripts of your sessions will be used for research purposes . . . anonymized

data and recordings may be used in future research studies.”

Since the ultimate goal of the project is to enable automated feedback to PSTs that would replace human feedback, there is a risk of incorrect feedback, since it is unlikely that an automated system will be accurate 100% of the time. First, human raters also sometimes make mistakes. Second, at least some of the use cases of the tool with feedback are within teacher training programs led by teacher educators; any feedback that surprised the PST or seemed unclear or incorrect can be discussed with the teacher educator. Third, every PST has access to the video recording of their own simulated discussion from Mursion; they can review the video to verify that the feedback makes sense with respect to their performance. Finally, a PST can engage in the simulation multiple times and it is possible that some of the feedback mistakes will be rectified in successive simulations.

Our use of the toolkits is in accordance with their licensing terms: Apache 2.0 license for HuggingFace transformers<sup>3</sup> and BSD 3.0 license for scikit-learn.<sup>4</sup>

## 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 the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 224–233, Seattle, Washington. Association for Computational Linguistics.
- 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.
- Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.
- Ahmed Ashraf Butt, Saira Anwar, Ahmed Magooda, and Muhsin Menekse. 2022. Comparative analysis of the rule-based and machine learning approach for assessing student reflections. In *Proceedings of the 16th International Conference of the Learning Sciences – ICLS 2022*.
- Courtney B. Cazden. 1988. *Classroom Discourse: The Language of Teaching and Learning*. Heinemann, Portsmouth, NH, USA.
- Domenic V. Cicchetti. 1994. [Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology](#). *Psychological Assessment*, 6(4):284–290.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- 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 B. Thomas, Susan Stansberry, and Ying Xu. 2020. [A mixed reality simulation offers strategic practice for pre-service teachers](#). *Computers & Education*, 144(103696).
- Debajyoti Datta, James P Bywater, Sarah Lilly, Jennifer L Chiu, Ginger S Watson, and Donald E Brown. 2023. [Classifying mathematics teacher questions to support mathematical discourse](#). In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2023. Communications in Computer and Information Science*, volume 1831, Cham. Springer.
- 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 and Jing Liu. 2023. [M-Powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes](#). In *Proceedings of the Tenth ACM Conference on Learning @ Scale, L@S '23*, page 59–69, Copenhagen, Denmark. Association for Computing Machinery.
- Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. [Can automated feedback improve teachers’ uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course](#). *Educational Evaluation and Policy Analysis*.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. [Measuring conversational uptake:](#)

<sup>3</sup><https://github.com/huggingface/transformers/blob/main/LICENSE>

<sup>4</sup><https://github.com/scikit-learn/scikit-learn/blob/main/COPYING>

- A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.
- 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 A. Dieker, Jacqueline A. Rodriguez, Benjamin Lignugaris/Kraft, Michael C. Hynes, and Charles E. Hughes. 2014. [The potential of simulated environments in teacher education: Current and future possibilities](#). *Teacher Education and Special Education*, 37(1):21–33.
- Zara Ersozlu, Susan Ledger, Alpay Ersozlu, Fiona Mayne, and Helen Wildy. 2021. [Mixed-reality learning environments in teacher education: An analysis of TeachLive™ research](#). *SAGE Open*, 11(3).
- Evan J. Fishman, Hilda Borko, Jonathan Osborne, Florencia Gomez, Stephanie Rafanelli, Emily Reigh, Anita Tseng, Susan Million, and Eric Berson. 2017. [A practice-based professional development program to support scientific argumentation from evidence in the elementary classroom](#). *Journal of Science Teacher Education*, 28(3):222–249.
- GO Discuss Project. 2021. [Scoring](#). Qualitative Data Repository.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2017. *Elements of Statistical Learning*, 2nd edition. Springer.
- Annemarie H Hindman, Barbara A Wasik, and Kate Anderson. 2022. [Using turn and talk to develop language: Observations in early classrooms](#). *Reading Teacher*, 76:6–13.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. [Bidirectional LSTM-CRF models for sequence tagging](#). *CoRR*, abs/1508.01991.
- 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*, pages 1–13, Honolulu, HI, USA. Association for Computing Machinery.
- Emily Jensen, Samuel L. Pugh, and Sidney K. D’Mello. 2021. [A deep transfer learning approach to modeling teacher discourse in the classroom](#). In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21*, page 302–312, Irvine, CA, USA. Association for Computing Machinery.
- Vikram Kumaran, Jonathan Rowe, Bradford Mott, Snigdha Chaturvedi, and James Lester. 2023. [Improving classroom dialogue act recognition from limited labeled data with self-supervised contrastive learning classifiers](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10978–10992, Toronto, Canada. Association for Computational Linguistics.
- Jay L. Lemke. 1990. *Talking Science: Language, Learning, and Values*. Ablex Publishing, Norwood, NJ, USA.
- 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](#). *CoRR*, abs/1907.11692.
- Malinda Ann Hoskins Lloyd, Nancy J. Kolodziej, and Kathy Brashears. 2016. [Classroom discourse: An essential component in building a classroom community](#). *School Community Journal*, 26(2):291–304.
- Edward Loper and Steven Bird. 2002. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Jamie N. Mikeska, Heather Howell, Joseph Ciofalo, Adam Devitt, Elizabeth Orlandi, Kenneth King, Michelle Lipari, 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 science: The mystery powder task](#). ETS Research Memorandum RM-21-06.
- Jamie N. Mikeska, Heather Howell, and Carrie Straub. 2019. [Using performance tasks within simulated environments to assess teachers’ ability to engage in coordinated, accumulated, and dynamic \(CAD\) competencies](#). *International Journal of Testing*, 19(2):128–147.
- Tanya Nazaretsky, Jamie N Mikeska, and Beata Beigman Klebanov. 2023. [Empowering teacher learning with AI: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion](#). In *LAK23: 13th International Learning Analytics and Knowledge Conference*, page 122–132, Arlington, TX, USA. Association for Computing Machinery.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

- Pennsylvania, USA. Association for Computational Linguistics.
- Kate Pearce, Sharifa Alghowinem, and Cynthia Breazeal. 2023. [Build-a-bot: Teaching conversational AI using a transformer-based intent recognition and question answering architecture](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16025–16032.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Victor Sampson and Margaret R. Blanchard. 2012. [Science teachers and scientific argumentation: Trends in views and practice](#). *Journal of Research in Science Teaching*, 9:1122–1148.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). arXiv:1910.01108.
- Jonathan T. Shemwell and Erin Marie Furtak. 2010. [Science classroom discussion as scientific argumentation: A study of conceptually rich \(and poor\) student talk](#). *Educational Assessment*, 15(3-4):222–250.
- Abhijit Suresh, Jennifer Jacobs, Charis Clevenger, Vivian Lai, James H Martin, and Tamara Sumner. 2021. [Using AI to promote equitable classroom discussions: The TalkMoves application](#). In *AIED 2021: Artificial Intelligence in Education*, volume 12749 of *Lecture Notes in Computer Science*, Cham. Springer.
- Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022a. [The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.
- Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022b. [Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.
- Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. [Automating analysis and feedback to improve mathematics teachers’ classroom discourse](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9721–9728.
- Harriet R. Tenenbaum, Naomi E. Winstone, Patrick J. Leman, and Rachel E. Avery. 2020. [How effective is peer interaction in facilitating learning? A meta-analysis](#). *Journal of Educational Psychology*, 112(7):1303–1319.
- Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2023. [Utilizing natural language processing for automated assessment of classroom discussion](#). In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, volume 1831 of *Computer and Information Science*, pages 490–496, Cham. Springer Nature Switzerland.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

## A Combined utterance-level classifier predictions

Here are some utterance blocks whose PST utterance was predicted positive for Indicator 3A, for the combined model (CAX).

**TEACHER**  
Okay. I'm trying to figure out how to explain this the best way possible. Actually, Carlos, do you want to explain it because sometimes hearing it from a friend is easier.

**CARLOS**  
Yeah. I didn't want to look at the weight because it just tells you how much there is, not what it is.

**WILL**  
Hmm. Well that's a little confusing.

**CARLOS**  
Well, what I mean is if you have a slice of pizza or you have a whole pizza, it's still pizza, right? It's just a different size.

**WILL**  
Well, yeah, I guess so. A slice of pizza is still pizza.

**CARLOS**  
Yeah, exactly and so it's the same with this. It doesn't matter how much you have, It's still the same thing.

**WILL**  
I guess the weight doesn't actually tell you what it is.

**CARLOS**  
Yeah, exactly.

**TEACHER**  
So in your small groups, and Carlos, you can join with Jayla and Emily. Talk about how you feel about the way that you went about the experiment and how you feel that you could've changed it.

**EMILY**  
You know, I guess problem has enough properties, just not enough of the right properties.

**CARLOS**  
Yeah. I thought I was on the right track with only using the weight, but I guess I didn't see it or realize that what the color in this is also the same one.

**JAYLA**  
Yeah. I think because we were trying to be like "Lets test all the properties" but I guess now that we know that.

Here are some utterance blocks whose PST utterance was predicted positive for Indicator 3B, for the combined model (CBX).

**TEACHER**  
Okay. Does anybody think that they should have looked at more properties or less? And why?

**CARLOS**  
Well, I think that they should have looked at more properties because they only looked at a couple. And they were also talking about how they looked at weight and they didn't need to look at that one.

**TEACHER**  
Okay. What does everyone else think about what Jayla just said?

**WILL**  
Well, what she said about why weight is an important property I didn't think that. I thought it was an important property, because you could measure it. But what Carlos said makes sense.

## B Some figures and tables

Powder	Properties				
	Texture	Color	Weight	Reaction with Vinegar	Mix with Water
Flour	Smooth	White	24 grams	No reaction	Looks cloudy
Cornstarch	Smooth	White	20 grams	No reaction	Looks cloudy
Baking Soda	Smooth	White	24 grams	Bubbles	Looks clear
Baking Powder	Smooth	White	24 grams	Bubbles	Looks cloudy
Sugar	Rough	White	26 grams	No reaction	Looks clear
Salt	Rough	White	22 grams	No reaction	Looks clear

Figure 6: Reference table of powders and properties for the Mystery Powder Task (Mikeska et al., 2021, p. 30).

Names(s): Jayla and Emily

We think the mystery powder is baking soda because it matches all of the properties on the chart that baking soda has listed. Baking soda is white, smooth, bubbles when mixed with vinegar, looks clear when mixed with water, and weighs 24 grams. The mystery powder has all the same properties.

All of the properties are useful and needed to decide what the unknown powder is, we can't just use one or two. Some of the powders have some of the same properties. Like all the powders we looked at were white so we wouldn't be able to identify the mystery powder if we just looked at the color. We had to find out which powder had the most matching properties as the mystery powder.

Figure 7: Pre-work by Jayla and Emily (Mikeska et al., 2021, p. 30).

```

cv_dict_classifier = {
"LR": (LogisticRegression(), {"C": [1,2,3,4,5,10,20], \
"class_weight": [None, "balanced"]}), \
"MLP": (MLPClassifier(random_state=42, max_iter=int(3e3)), {"hidden_layer_sizes": \
[1*(5,), 2*(5,), 3*(5,), 1*(10,), 2*(10,), 3*(10,), 1*(20,), 2*(20,), 3*(20,), 1*(30,)
, 2*(30,), 3*(30,)], \
"activation": ["logistic", "tanh", "relu"], \
"solver": ["lbfgs", "sgd", "adam"], "alpha": [0.00005, 0.0005]}), \
"DT": (DecisionTreeClassifier(random_state=42), { "splitter": ["best", "random"], \
"max_depth": np.arange(3, 15), "max_features": ["log2", "sqrt", None], "class_weight": [
None, "balanced"]}),
"RF": (RandomForestClassifier(random_state=42), {"max_depth": [5,10,20,30, None],
"max_features": [1, "sqrt"], "min_samples_leaf": [1, 2, 4], "min_samples_split":
[2, 5, 10], \
"class_weight": [None, "balanced"]})
}

```

Table 7: Classifier hyperparameter grids, for use with Scikit-learn.

```

cv_dict_regressor = {
"LR": (LinearRegression(), {"fit_intercept": [False, True]}), \
"MLP": (MLPRegressor(random_state=42, max_iter=int(3e3)), {"hidden_layer_sizes":
[1*(5,), 2*(5,), 3*(5,), 1*(10,), 2*(10,), 3*(10,), 1*(20,), 2*(20,), 3*(20,), 1*(30,)
, 2*(30,), 3*(30,)],
"activation": ["logistic", "tanh", "relu"],
"solver": ["lbfgs", "sgd", "adam"], "alpha": [0.00005, 0.0005]}), \
"DT": (DecisionTreeRegressor(random_state=42), { "splitter": ["best", "random"],
"max_depth": np.arange(3, 15), "max_features": ["log2", "sqrt", None]}), \
"BR": (BayesianRidge(), {"tol": [1e-4, 1e-3, 1e-2],
"alpha_1": [1e-7, 1e-6, 1e-5, 1e-4, 1e-3], "lambda_1": [1e-7, 1e-6, 1e-5, 1e-4, 1e
-3],
"fit_intercept": [False, True]}), \
"RF": (RandomForestRegressor(random_state=42), {"max_depth": [5,10,20,30, None],
"max_features": [1, "sqrt"], \
"min_samples_leaf": [1, 2, 4], \
"min_samples_split": [2, 5, 10]}),
}

```

Table 8: Regressor hyperparameter grids, for use with Scikit-learn.



Model	Selected	Cohen's $\kappa$ mean (SE)
CAN	3 epochs	0.475 (0.001)
CAS	MP	0.653 (0.026)
CAX	RF	0.717 (0.026)
CBN	7 epochs	0.491 (0.007)
CBS	RF	0.324 (0.043)
CBX	MP	0.622 (0.033)

Table 9: 5-fold cross-validation results for classifiers, with Cohen's  $\kappa$  as metric. Higher values are better.

Model	Selected	MSE mean (SE)
RAN	MP	0.332 (0.046)
RAS	RF	0.250 (0.035)
RAX	RF	0.245 (0.042)
RBN	LR	0.242 (0.015)
RBS	MP	0.387 (0.017)
RBX	LR	0.236 (0.012)
RDN	BR	0.251 (0.044)
RDS	MP	0.233 (0.034)
RDX	BR	0.219 (0.034)

Table 10: 5-fold cross-validation results for regressors, with mean squared error (MSE) as metric. Lower values are better.