Improving the Generalizability of Models of Collaborative Discourse

Chelsea Chandler [1], Rohit Raju [2], Jason G. Reitman [1], William R. Penuel [1, 3], Monica Ko [1],
Jeffrey B. Bush [1], Quentin Biddy [1], Sidney K. D'Mello [1, 2, 4]
[1] Institute of Cognitive Science, University of Colorado Boulder
[2] Department of Computer Science, University of Colorado Boulder
[3] School of Education, University of Colorado Boulder
[4] Department of Psychology and Neuroscience, University of Colorado Boulder
{chelsea.chandler, rohit.raju, jason.reitman, william.penuel, monlin.ko,
jeffrey.bush, quentin.biddy, sidney.dmello}@colorado.edu

ABSTRACT

We investigated methods to enhance the generalizability of large language models (LLMs) designed to classify dimensions of collaborative discourse during small group work. Our research utilized five diverse datasets that spanned various grade levels, demographic groups, collaboration settings, and curriculum units. We explored different model training techniques with RoBERTa and Mistral LLMs, including traditional fine-tuning, data augmentation paired with fine-tuning, and prompting. Our findings revealed that traditionally fine-tuning RoBERTa on a single dataset (serving as our baseline) led to overfitting, with the model failing to generalize beyond the training data's specific curriculum and language patterns. In contrast, fine-tuning RoBERTa with embedding augmented data led to significant improvements in generalization, as did pairing Mistral embeddings with a support vector machine classifier. However, fine-tuning and few-shot prompting Mistral did not yield similar improvements. Our findings highlight scalable alternatives to the resource-intensive process of curating labeled datasets for each new application, offering practical strategies to enhance model adaptability in diverse educational settings.

Keywords

Generalization, Natural language processing, Collaboration analytics

1. INTRODUCTION

Collaboration in small group settings is central to K-12 education, higher education, and professional environments, valued for fostering critical thinking, problem-solving, and social interaction [20, 28, 39]. In K-12 settings, teachers play a pivotal role in orchestrating small group collaboration by monitoring group dynamics, guiding learning activities, and encouraging meaningful discussions to probe deeper thinking and help students develop these skills [41, 54]. However, facilitating such interactions is complex, compounded by substantial gaps in student collaboration skills. For example, in collaborative problem solving (CPS) tasks where students must work together to solve complex problems [16], fewer than 10% of students reach top proficiency levels [34] and fewer than 30% could solve low-complexity problems, highlighting the need for targeted interventions [20, 22, 55].

A key barrier to improving collaboration skills is the lack of consistent assessment and feedback methods [16]. Recent efforts have leveraged Natural Language Processing (NLP) models to analyze student discourse in small groups [47, 59, 3, 17, 49, 37]. For example, [37] fine-tuned RoBERTa to detect CPS skills like constructing shared knowledge, coordination/negotiation, and maintaining team function from a validated framework [53]. These models generate insights that enable immediate, actionable feedback as a means to improve student collaboration skills [12]. By bridging the gap between assessment and intervention, such innovations have the potential to transform how collaboration skills are nurtured and sustained in educational environments.

Generalization - the ability of models to transfer knowledge from their training domain to novel contexts - is a key desideratum in NLP, particularly in collaborative analytics scenarios that span multiple curricula and modes of collaboration (e.g., remote vs. in-person). This ability is typically evaluated by assessing a model's performance on a dataset distinct from the training dataset, taking into account the differences between the distinct sources [27]. Developing generalizable models is essential for broadening their applicability and impact. Further, models that can adapt to diverse educational settings can ostensibly capture the nuanced and dynamic nature of collaboration more effectively, providing meaningful feedback across varied contexts.

However, as we review below, current models of collaborative discourse are optimized on single datasets that have been laboriously annotated for indicators of collaboration skills [53]. Utilizing these models in new contexts entails collecting, annotating, and retraining the models on a new dataset, an endeavor that does not scale to authentic educational settings where curricula and context may vary daily. To address this limitation, we explored approaches to en-

Chelsea Chandler, Rohit Raju, Jason G. Reitman, William R. Penuel, Monica Ko, Jeffrey B. Bush, Quentin Biddy, and Sidney K. D'Mello. Improving the Generalizability of Models of Collaborative Discourse. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 215–227. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870167

hancing the generalizability of NLP models of collaborative discourse trained on a single domain across diverse curricula. Additionally, we conducted a qualitative error analysis to better understand which linguistic features contributed to overfitting and generalizability, addressing the linguistic basis of transfer across collaborative contexts.

2. RELATED WORK

NLP Approaches to Modeling Collaborative Discourse. A growing body of research has explored the use of NLP techniques to model collaborative discourse, leveraging language data from text chats and transcribed speech to model CPS skills like negotiation, regulation, and argumentation [59, 3, 47, 17, 49, 14, 50]. Early NLP approaches relied on extracted features like words, phrases, and part-of-speech tags [18, 24, 42], but recent advances leverage pre-trained neural networks, which effectively capture the complexities of collaboration [31, 37], generally outperforming earlier approaches [59].

Generalizability in Collaborative Discourse. Much of the existing work focuses on data from single domains or curricula, leading to models that are highly specialized but lack the broader, abstract representations necessary for application in varied settings. This narrow focus limits the utility of such models in real-world environments, where diverse contexts and populations are the norm. This issue has received comparatively limited attention in collaboration analytics. One study that investigated this issue in CPS models found that fine-tuned BERT and a dictionary based approach (i.e., Linguistic Inquiry Word Count) generalized from one domain to another better than an an n-gram approach, suggesting methods that learn abstractions beyond literal words enhance generalizability [36].

Similarly, [9] showed that adding contextual features (activity type, time of day, language) to their multimodal learning analytics framework improved the generalizability of their collaborative learning models. While this work offers general insights, the authors relied on random forest machine learning models to conduct all aspects of their analyses, leaving a gap for studies that explore deeper investigations into NLP modeling approaches. Further, the impact of these studies across varied contexts such as age groups, backgrounds, and interaction modalities, as well as qualitative analyses of the models remains unexplored.

Fine-Tuning vs. Prompting for Collaborative Discourse. A fundamental challenge in computational modeling is balancing accuracy with generalizability. Furthermore, NLP models exhibit varying levels of generalizability depending on their design and training. For example, fine-tuned large language models (LLMs) are known for high task-specific accuracy [57, 56], yet often risk overfitting and struggle with generalization across contexts [10, 44]. Instruction-tuned models such as GPT excel in zero-shot and few-shot scenarios, making them effective for applications requiring domain transfer [5, 32, 35]. However, their performance may vary on tasks requiring complex, nuanced interpretations of dialogue, especially in classification tasks grounded in pedagogical frameworks.

Modeling collaborative discourse presents unique challenges for LLMs compared to other NLP task types where they typically excel. Unlike tasks like text summarization or question answering, which rely on language properties and factual reasoning, collaborative discourse involves social, cognitive, and contextual dimensions. These include alignment with pedagogical frameworks, turn-taking dynamics, and reasoning about the problem solving context. Thus, simple prompting approaches may struggle to capture the deeper structures of collaboration, especially in real-world settings.

To this point, recent work has shown that in real-world educational tasks, task-specific fine-tuning generally outperforms prompting [19], although performance is highly task-specific. For tasks like determining whether an utterance exhibits evidence of student reasoning (e.g., making sense of concepts, justifying ideas) [11], prompting GPT-4 performed well without any examples, since the task can be explained through a simple descriptive prompt and can leverage cue words like “because”. However, it struggled to classify the more nuanced dimensions of collaborative discourse. Another study compared the accuracy of fine-tuned RoBERTa and GPT-3 on teachers’ use of academically productive talk moves and showed that GPT-3 struggled with recall and underperformed on underrepresented categories [30]. These studies underscore the importance of comparing fine-tuning and prompting methods across tasks and domains [46].

Techniques for Enhancing Model Generalizability. To enhance generalizability without sacrificing accuracy, research has explored augmentation methods such as adversarial training and embedding space perturbations [2]. Adversarial training introduces perturbations into training data to create adversarial examples that are slightly altered, encouraging the model to learn robust representations. It has been shown to produce models capable of adapting to multiple tasks without requiring additional task-specific debiasing steps [15]. Embedding space perturbation specifically involves swapping words with nearby ones in a continuous vector space. Training data augmentation can also involve generating paraphrased utterances, varying discourse structures, or introducing words from different domains, improving the robustness of models [43].

Comparing augmentation methods with a range of NLP methodologies (traditional fine-tuning, prompting, and modeling extracted embeddings) represents a pathway for bridging the gap between traditional NLP tasks and the complex dynamics of collaborative learning, improving models that more accurately support real-world teamwork and problem-solving.

3. CURRENT STUDY, CONTRIBUTION, & NOVELTY

In this study, we investigated the ability of NLP models to classify collaborative discourse across diverse educational contexts. We trained five NLP model types on one dataset called Sensor Immersion (focused on a sensor programming task), and tested them on four held-out datasets: (1) Physics Playground (focused on an educational physics game), (2) Minecraft Hour of Code (focused on block programming), (3) Moderation Unit (focused on gaming system moderation), and (4) Self-Driving Cars (focused on model car assembly and programming).

We considered RoBERTa [33] and Mistral 7B [29] models, leveraging traditional fine-tuning, fine-tuning with data augmentation, embedding extraction paired with traditional machine learning, and few-shot prompting. Fine-tuning involves adjusting the weights of models for specific tasks, while fine-tuning with augmentation mitigates overreliance on context-specific terminology by perturbing training data. Embedding and classifier methods extract utterance embeddings from an LLM for training a traditional machine learning model. Few-shot prompting provides models with context and labeled examples within a prompt. This multifaceted approach allowed for a comprehensive examination of techniques aimed at improving model adaptability. We chose these models and approaches to represent current techniques, but did not conduct a systematic ablation study of all variants.

RoBERTa and Mistral were chosen for their size, open-source accessibility, and specialization capabilities. RoBERTa is an ideal baseline as it is widely used for fine-tuning, and its variants have a proven history in CPS modeling [37]. Mistral 7B, a more recent state-of-the-art LLM, maintains strong performance on specialized tasks and is a more parameter-efficient open-source model than many newer LLMs [29]. Both models offer fast inference and low computational costs. Unlike models like GPT and Gemini which require an API for queries and fine-tuning (a barrier for use with sensitive educational transcripts as in the present case) as well as more resource-intensive fine-tuning processes, RoBERTa and Mistral 7B provide balanced performance and efficiency.

Our task targeted the relationship dimension of collaboration, emphasizing cognitive and social processes that foster positive team dynamics and the development of collaboration skills as an intrinsic goal [16, 20]. We focused on the identification of three so called Community Agreements (CAs), which are mutually-agreed upon norms of behavior that students adopt to facilitate small group work in collaboration with their teachers [1]. These include: Committed to our Community, Moving Thinking Forward, and Being Respectful (hereafter referred to as *Community*, *Thinking*, and *Respect*) [8, 4]. While prior research has shown that NLP models effectively classify CAs within their training domain [8], their ability to generalize remains unknown.

We investigated four research questions: (1) how model selection and training strategies affect generalization across collaborative contexts, (2) whether certain collaboration constructs are more prone to overfitting, (3) which model implementations are robust to automated speech recognition errors, and (4) what linguistic features underlie successful generalization in NLP models of collaboration. We sought to uncover practical strategies for enhancing the robustness and adaptability of LLMs in analyzing classroom discourse, ultimately advancing tools for fostering collaboration in educational environments.

Our research is novel because it systematically investigates the generalizability of NLP models of collaborative discourse across diverse educational contexts. Unlike prior studies that focus on single-domain models prone to overfitting, we explore multiple techniques across five distinct datasets varying in student populations, interaction modalities, contexts, and collaboration tasks. By comparing RoBERTa and Mistral 7B, we provide new insights into balancing model generalizability and domain-specific accuracy, offering scalable alternatives to labor-intensive dataset labeling.

4. DATA

The designated Institutional Review Boards approved all study procedures. All students provided assent or their parents or legal guardians provided consent.

4.1 Data Collection

Five datasets were analyzed as part of this research: *Sensor Immersion*, *Self Driving Cars*, *Moderation Unit*, *Physics Playground*, and *Minecraft Hour of Code* (see Table 1).

Sensor Immersion (Primary Train Dataset) was collected from urban, rural, and suburban public middle school classrooms in the Western United States between 2021-2023. Students worked in small groups programming and wiring sensors to collect environmental data. They explored an interactive display called the Data Sensor Hub [7], constructing scientific models and developing skills to replicate its functionality [13]. Speech was recorded with Yeti Blue, an omnidirectional microphone that was placed at each table. Relying on single microphones in a classroom with several groups interacting concurrently resulted in noisy data [6, 45]. Five-minute segments that met 20-word thresholds were identified from each recording; if none met this criteria, the recording was excluded. The data consisted of 164 students (73 dyads, seven triads, six tetrads) under the guidance of 14 teachers and 91 recordings comprising 8,601 student utterances.

The remaining four datasets were used as held-out test sets to evaluate the generalizability of the models.

Self Driving Cars (Held-out Dataset) was collected from four public middle school classrooms in the Western United States during the 2023-2024 school year. These data were sampled from four lessons of a Self Driving Cars curriculum, where small groups of students worked to assemble and program model cars to follow a path, avoiding obstacles. Speech was recorded and sampled using the same procedure and equipment as in the Sensor Immersion dataset. This dataset consisted of eight students (all dyads) under four teachers and 16 recordings comprising 969 student utterances.

Moderation Unit (Held-out Dataset) was collected from three public middle school classrooms in the Western United States during the 2023-2024 school year. These data were sampled from two lessons in a Moderation Unit curriculum, where small groups worked to program solutions to Minecraft puzzles and evaluate gaming moderation systems. Speech was recorded and sampled in the same manner as the Sensor Immersion dataset. This dataset consisted of 39 students (two dyads, seven triads, seven tetrads) guided by one teacher and 16 recordings comprising 1,568 student utterances.

Due to district policies, demographic details for these data were unavailable, but the schools served diverse student populations.

Physics Playground (Held-out Dataset) and **Minecraft Hour of Code (Held-out Dataset)** were collected from a remote

Table 1: Overview of the five datasets.

Curriculum	Task	Year	Grade	Students	Recordings	Utterances
Sensor Immersion	Programming and wiring sensors	2021-2023	Middle School	164	91	8,601
Self Driving Cars	Assembling and programming model cars	2024	Middle School	8	16	969
Moderation Unit	Evaluating gaming moderation systems	2024	Middle School	39	16	1,658
Physics Playground	Playing an educational physics game	2018-2019	University	285	96	45,550
Minecraft Hour of Code	Playing a block-based programming game	2018-2019	University	96	32	10,816

CPS study involving 288 university students (average age = 22 years) [48]. Speech was recorded with individual headsets and sampled by extracting random 90-second chunks from the first, second, and third five minutes of each 15-minute recording block for transcription and annotation. Participants self-reported as 54% female, 41% male, 1% non-binary/third gender, and 4% did not report. Race was self-reported as 48% Caucasian, 25% Hispanic/Latino, 17% Asian, 3% Black or African American, 1% American Indian or Alaska Native, 3% Other, and 3% did not report.

Physics Playground involved an educational game designed to teach physics concepts (e.g., Newton's laws, energy transfer, properties of torque) through interactive game play. Participants drew objects like ramps, levers, and pendulums to guide a ball toward a target, with all objects adhering to the laws of physics. This dataset comprised 96 recordings with 45,550 student utterances.

Minecraft Hour of Code employed block-based programming to interactively teach programming concepts (e.g., if-else statements). Constructs were represented as interconnecting blocks that assemble into syntactically correct code that controls the actions of a Minecraft character, allowing users to run and preview their code in real-time. This dataset comprised 32 recordings with 10,816 student utterances.

These five datasets highlight diverse instructional contexts and modalities, offering a rich basis for evaluating the adaptability of NLP models across collaborative learning environments. Key differences among the first three datasets (Sensor Immersion, Self-Driving Cars, and Moderation Unit) and the latter two (Physics Playground and Minecraft Hour of Code) are the context of data collection (classroom vs. lab), the age of the participants (K-12 vs. college), interaction modality (in person vs. remote), and microphone (Yeti-Blue tabletop vs. headset microphones).

4.2 Data Processing and Transcription

Recordings from the *Sensor Immersion*, *Moderation Unit*, and *Self Driving Cars* datasets were manually transcribed with annotations for contextual notes (e.g., "[laughter]"), speaker intent (e.g., "[addressing group]"), and inaudible speech ("[inaudible]"). This process captured essential nonverbal and contextual aspects of the data.

Physics Playground and *Minecraft Hour of Code* were historic data sets that had been transcribed using IBM Watson, a popular ASR system at the time of data collection. These datasets did not have human transcripts, and the IBM Watson transcripts were used for initial annotations. We do not provide results on the IBM Watson transcripts.

To ensure consistency across datasets, we transcribed all recordings, including Physics Playground and Minecraft Hour of Code, using Whisper-Large-v2, an open-source ASR model [38]. The word error rate - calculated as *(substitutions + deletions + insertions) / total words* - for Sensor Immersion, Self Driving Cars, and Moderation Unit was 67%, 64%, and 69%, respectively, highlighting the challenges of processing noisy classroom audio with overlapping speech.

Utterances from teachers or individuals outside the group were excluded from the analysis to maintain focus on small group interactions. Transcripts were normalized by removing punctuation, converting text to lowercase, replacing hyphens with spaces, and eliminating transcriber and ASR notes. Finally, utterances were anonymized using the *spaCy* python library, which applied named entity recognition to identify proper names, substituting them with the placeholder "[name redacted]". Reported utterance counts reflect post-processed data.

4.3 Human Coding of CA Labels

CA coding followed work by [8], which involved adapting a CPS framework from [53] to identify CAs. Aligned with competencies defined by the Organisation for Economic Cooperation and Development (OECD) [34] and key social and cognitive collaboration skills [21], the CPS framework encompasses three facets operationalized through 18 validated indicators [52, 53, 58, 51]. The indicators were mapped to the Resepct, Community, and Thinking CAs using OpenSciEd definitions in consultation with collaboration and curriculum experts (Table 2).

Coding was conducted at the utterance level, with coders watching videos to incorporate nonverbal cues and context (e.g., screenshots and camera views). Thus, the coders used information that extends beyond the language itself, posing challenges for the NLP models.

Four coders, including an expert involved the original framework development, annotated *Sensor Immersion* transcripts with the 18 CPS indicators. The process involved iterative refinement to ensure consensus and reliability among coders. Then, coders independently labeled utterances with review

Table 2: Collaborative Problem Solving (CPS) indicator mappings to Community Agreements (CAs) and examples from each of the five datasets. Examples were chosen to highlight context specific wording from each dataset.

Community Agreement CPS Indicators	**Examples**
Being Respectful	
Responds to others' questions or ideas	"[...] Sorry. My bad. Scroll over to where you can see the whole thing." (*SI*)
Asks others for suggestions	"This guy is gonna be winning some drag race challenges." (*SDC*)
Compliments or encourages others	"I actually did it! I like your Minecraft games. They're fun." (*MU*)
Apologizes for one's mistakes	"Sorry because there's one two three underwater online here." (*MHC*)
	"Yes same good job just drop it earlier." (*PP*)
Committed to our Community	
Talk about the challenge situation	"If your pool has a heater, the bar will go up a lot." (*SI*)
Confirms understanding	"Oop, oh, It's moving. Maybe you disconnect it now." (*SDC*)
Discusses the results	"Move toward zombies and then attack." (*MU*)
Provides instructional support	"No we're gonna end up in the water substance." (*MHC*)
Asks others for suggestions	"Okay I think it's stuck in a tree." (*PP*)
Moving Thinking Forward	
Proposes (in)correct solutions	"So now I think we push download and see what happens." (*SI*)
Strategizes to accomplish task goals	"Which one is it? This one? Follow line?" (*SDC*)
Asks others for suggestions	"When spawn what do you want it to do?" (*MU*)
Provides reasons to support a solution	"Maybe we should have them lay the bricks afterwards." (*MHC*)
Questions/corrects others' mistakes	"The water transferred over so should we delete the grey box [...]?" (*PP*)

SI: Sensor Immerison, SDC: Self Driving Cars, MU: Moderation Unit, MHC: Minecraft Hour of Code, PP: Physics Playground

Table 3: Occurrence rate of each CA across the five datasets.

Dataset	Respect	Thinking	Community
Sensor Immersion	12%	12%	20%
Self Driving Cars	16%	15%	34%
Moderation Unit	18%	18%	33%
Minecraft	-	22%	27%
Physics Playground	16%	21%	24%

by the expert to maintain consistency. To assess reliability, the four coders annotated the same 118 utterances across 3 observations, achieving Gwet's AC1 indicator-level agreement [23] of 0.75-1.00.

Since *Self Driving Cars* closely resembled Sensor Immersion, no significant adaptations to the coding scheme were necessary. Coders reviewed observations to discuss and confirm the framework's applicability before proceeding with coding.

Moderation Unit posed unique challenges as the CPS framework had not been applied to this domain before. Annotators analyzed observations from each lesson, identifying areas where adaptations of the codebook were necessary. A few refinements ensured the coding scheme accurately reflected the context.

Minecraft Hour of Code and *Physics Playground* annotation followed Sun et al. [52]. Three experts coded utterances using IBM Watson transcripts and their associated videos, achieving Gwet's AC1 indicator-level agreement between 0.88 and 1.00 on ten 90-second videos (406 utterances). After achieving adequate reliability, videos were randomly assigned to the three coders for independent coding.

Table 3 provides base rates for the three CAs by dataset. The occurrence of the Respect CA was very low for Minecraft so we excluded it from the analyses.

5. METHODS

We developed five models for each of the three CAs using two open-source transformer-based architectures: RoBERTa and Mistral. RoBERTa was selected due to its robust fine-tuning performance across various NLP domains, including in related tasks such as CPS prediction. Mistral, a more recent model, was chosen for its efficiency and capacity to produce high-quality embeddings, making it an ideal candidate for few-shot learning and traditional machine learning pipelines. For each approach, we followed previous research indicating that training models with a combination of human and ASR transcripts (ASR augmented training) improved accuracy in noisy classroom scenarios [8, 6].

The five model implementations included (1) fine-tuning RoBERTa, (2) fine-tuning RoBERTa with embedding space data augmentation, (3) training a support vector machine (SVM) classifier with Mistral embeddings, (4) prompting Mistral with few-shot examples, and (5) fine-tuning Mistral.

Fine-Tuned RoBERTa (Baseline). The baseline model was a fine-tuned RoBERTa language model, a BERT variant with a multilayer bidirectional transformer architecture. Following previous research [8], RoBERTa was fine-tuned using binary labels for each CA. Utterances were tokenized using the RoBERTa tokenizer to maintain uniform sequence lengths through padding and truncation. Fine-tuning involved a batch size of 32, a learning rate of 5e-6, 50 training epochs, and 50 warmup steps. Hyperparameters were guided by prior research [36] and only minimally adjusted.

Augmented RoBERTa. To mitigate overfitting, we perturbed text sequences with the EmbeddingAugmenter class from the python package *textattack* which simulates variations in utterances by transforming individual words with replacements that are close to the original word in an embedding space. For example, the utterance “Maybe we still don’t have sensors.” was augmented to “Maybe we still don’t have *detectors*.” For each utterance, up to five augmented examples were generated and incorporated into the training dataset, culminating in a total of 70,836 utterances from the Sensor Immersion dataset. This total includes the original human-transcribed utterances and their augmentations, as well as the ASR-transcribed utterances and their corresponding augmentations. In some cases, fewer than five new utterances were generated, as the method enforces a minimum cosine similarity threshold of 0.8. If no suitable word replacements could be found within this threshold, a swap was not made, thereby limiting the number of generated utterances. This model was trained identically to the baseline RoBERTa model but included the augmented data.

Mistral Embeddings + SVM. We used the Mistral model *mistral-7B-v0.1* as an embedding extractor for a downstream SVM classifier. Transcripts were processed through Mistral to generate dense vector embeddings for each utterance, capturing contextual representations of the speech. We then trained the SVM classifier with each embedding, paired with corresponding CA labels, using the parameters $C = 1$, $gamma = scale$, and $kernel = rbf$.

Prompting Mistral. Few-shot prompting was implemented with Mistral, using five labeled examples from the Sensor Immersion dataset for each test utterance. The prompt included instructions, labeled examples, and a test utterance:

> Scenario: You’re observing students working collaboratively. Your task is to assess the student utterance after <<< and determine if the utterance exhibits any of the following indicators of Moving Thinking Forward. If it does, respond with ‘Yes’, otherwise, respond with ‘No’.
>
> 1. Providing Reasons or Evidence: The student offers reasons or evidence (e.g., from past experience or investigation) to support their action, suggestion, or conjecture.
> 2. Realization or Insight: Look for markers such as ‘so’ and ‘oh’ as evidence of a student realizing something or gaining insight.
> 3. Conjecture: The student proposes an idea for the group to consider, often using hedges like ‘maybe’. These proposals invite responses and are about claims or assertions.
>
> Here are some examples:
> Student Utterance: “Let’s all take turns”
> Exhibits moving thinking forward: Yes
>
> [...]
>
> <<<
>
> Test Utterance: “How about we split into pairs?”
> Exhibits moving thinking forward:

The five labeled few-shot examples were selected randomly but stratified to ensure representation across the dataset. To ensure a fair assessment, we repeated this process five times per test utterance with different examples, averaging the accuracy from each of the five rounds for a final metric.

Fine-Tuned Mistral. Finally, we employed Low-Rank Adaptation (LoRA) [25] to fine-tune the Mistral model. LoRA is a parameter-efficient method that injects trainable low-rank matrices into the model’s attention layers, allowing fine-tuning without updating all parameters, making the process more memory-efficient. The training involved a learning rate of 2e-4, a batch size of 32, and nine training epochs.

Cross Validation. We adopted stratified 10-fold cross validation for evaluation within the Sensor Immersion dataset. The data were split into ten subsets with approximately equal occurrence rates, ensuring that observations (classroom sessions) did not span multiple folds. We iteratively trained the models on nine folds and tested on the one held-out, using the same folds across all five implementations. After evaluating performance within the Sensor Immersion dataset, each model was fully trained on Sensor Immersion data and tested on transfer datasets to assess generalization.

Evaluation Metrics. The primary evaluation metric we used was the Area Under the Receiver Operating Characteristic curve (AUROC), which is a widely used metric for evaluating the performance of binary classifiers, representing the model’s ability to distinguish between positive and negative classes. We chose to use AUROC to compare results from different models on the same dataset because it provides a comprehensive assessment of performance across various thresholds, offering insight into the trade-offs between sensitivity and specificity. It is independent of thresholds and class imbalance, an important feature when transferring models to new datasets where the distributions varied.

To compare transfer results with the RoBERTa baseline, we calculated the percent change with the following formula:

$$PercentChange = 100 \times \frac{TransferAUROC - BaselineAUROC}{BaselineAUROC}$$

which allowed us to quantify the improvement or decline in performance relative to the baseline RoBERTa model, facilitating a clear understanding of how each model compared with respect to generalization accuracy.

6. RESULTS

6.1 Baseline Model: Sensor Immersion

We first investigated within-domain performance on the Sensor Immersion dataset. When averaged across CAs and transcript type (human and Whisper), the fine-tuned RoBERTa model (mean AUROC = 0.71) tied with the Mistral + SVM approach (0.71) as the best model and both models outperformed the other three approaches (Table 4). Whereas the fine-tuned RoBERTa model demonstrated the best performance for two of the three CAs (Thinking and Respect), Mistral embeddings paired with the SVM classifier outperformed other models for the Community CA.

The RoBERTa model with data augmentation (mean AUROC = 0.68) and LoRA Mistral (0.67) achieved lower average AUROC scores compared to the traditionally fine-tuned RoBERTa model (0.71), but may still be valuable for generalization to other datasets. In contrast, the few-

Table 4: Average AUROC from 10-fold cross validation within the Sensor Immersion dataset for the five model implementations. Results from each of the test CAs and transcript types are shown.

	Human Transcripts			Whisper Transcripts		
	Community	Thinking	Respect	Community	Thinking	Respect
Baseline RoBERTa	0.68	**0.76**	**0.81**	0.62	**0.70**	**0.71**
Augmented RoBERTa	0.64	0.73	0.80	0.59	0.65	0.69
Mistral + SVM	**0.71**	0.75	0.80	**0.64**	0.67	0.69
Few-shot Mistral	0.57	0.62	0.47	0.57	0.59	0.47
LoRA Mistral	0.64	0.68	0.76	0.60	0.63	0.69

Table 5: Test set AUROC for each held out dataset and model implementation. Results are broken down by CA and transcript type (top section contains human transcript results and the bottom contains Whisper transcript results).

Human Transcripts

	Moderation Unit			Self Driving Cars		
	C	T	R	C	T	R
Baseline RoBERTa	0.55	0.61	0.77	0.60	0.70	0.78
Augmented RoBERTa	**0.69**	**0.68**	**0.79**	**0.69**	**0.71**	0.80
Mistral + SVM	0.65	0.67	0.74	0.68	0.70	**0.82**
Few-shot Mistral	0.56	0.57	0.51	0.52	0.55	0.43
LoRA Mistral	0.56	0.61	0.50	0.58	0.58	0.45

Whisper Transcripts

	Moderation Unit			Self Driving Cars			Minecraft		Physics Playground		
	C	T	R	C	T	R	C	T	C	T	R
Baseline RoBERTa	0.51	0.57	0.56	0.55	0.59	0.64	0.67	0.74	0.60	0.59	0.62
Augmented RoBERTa	0.54	**0.60**	**0.59**	**0.58**	**0.61**	**0.65**	**0.75**	**0.79**	0.62	**0.63**	**0.63**
Mistral + SVM	0.52	0.57	0.58	**0.58**	0.59	0.62	0.73	0.75	**0.63**	**0.63**	0.60
Few-shot Mistral	0.53	0.49	0.50	0.49	0.52	0.49	0.64	0.62	0.50	0.57	0.48
LoRA Mistral	**0.55**	0.55	0.47	0.52	0.56	0.50	0.65	0.71	0.58	0.60	0.46

C: Community, T: Thinking, R: Respect

shot prompting (0.55) barely outperformed chance within the Sensor Immersion dataset for all three CAs. This indicates that while few-shot prompting techniques hold promise in other contexts, they may not yet be optimized for this specific qualitative coding task or dataset.

Given that the fine-tuned RoBERTa model tied for the best within-domain result with the Mistral + SVM approach and outperformed it for two of the three CAs, we adopted it as our baseline model to test generalizablity.

6.2 Testing Generalizability

Overall Results. After fully training each of the five models on the Sensor Immersion dataset, we evaluated their performance on four held-out test sets to assess their generalizability. The AUROCs are given in Table 5. Density plots of the test set AUROCs pooled over CA, dataset, and transcript type (human or Whisper) are shown in Figure 1. These plots provide a high-level view of the overall trends in generalizability. The baseline RoBERTa model exhibited a substantial drop in accuracy when applied to the held-out datasets, whereas the augmented RoBERTa and Mistral + SVM and approaches maintained more consistent performance across datasets.

The general performance pattern (with average AUROC in parantheses) was as follows:

$$Augmented\,RoBERTa\,(0.67) > Mistral + SVM\,(0.65) > Baseline\,RoBERTa\,(0.63) > LoRA\,Mistral\,(0.56) > Few-Shot\,Mistral\,(0.53)$$

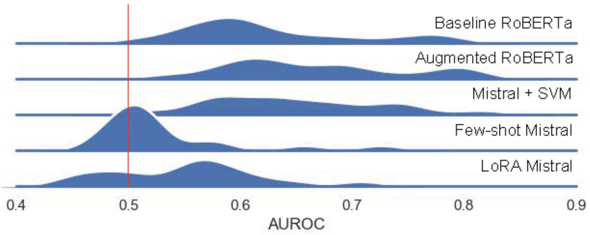

Figure 1: Distributions of test set AUROCs for each of the five models, aggregated across dataset, CA and transcript type (human and Whisper). The red line indicates an AUROC of 0.5, or random guessing.

Both the augmented RoBERTa model and the Mistral + SVM approach showed greater success in generalizability by collectively outperforming the other three models in all but one case. While the baseline RoBERTa model occasionally matched the performance of the augmented RoBERTa and Mistral + SVM models, it more frequently lagged behind.

The few-shot prompting and fine-tuning approaches for Mistral did not produce robust classification models. Despite being a large-scale language model designed with extensive

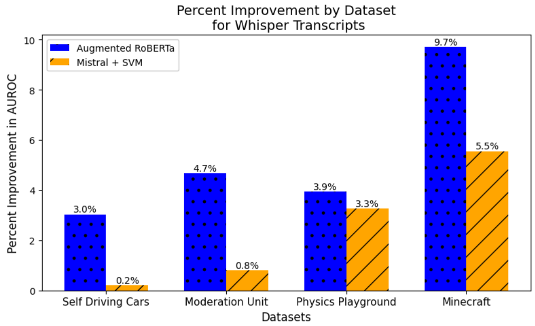

Figure 2: Bar chart showing the percent improvement in AUROC over the baseline RoBERTa model for augmented RoBERTa (blue) and Mistral + SVM (orange) across the 4 transfer datasets. Improvements are shown for Whisper transcripts only and averaged over the three CAs.

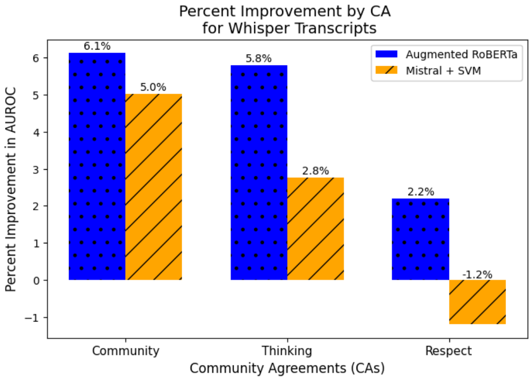

Figure 3: Bar chart showing the percent improvement in AUROC over the baseline RoBERTa model for augmented RoBERTa (blue) and Mistral + SVM (orange) across the 3 CAs. Improvements are shown for Whisper transcripts only and averaged over test dataset.

general knowledge, Mistral struggled to learn and apply the more qualitative patterns required by the CA framework. This limitation may stem from the mismatch between the model's size and complexity and the relatively small, domain-specific training dataset. The fine-tuning process likely failed to adjust the model effectively to the nuances of the CAs, resulting in low performance.

The results showed variance across models and datasets, emphasizing that model selection and training methodologies greatly influence generalizability. The variability also suggests that differences in linguistic and contextual features across datasets may impact model performance, especially for curriculum-specific patterns. Next we examined these patterns from different points of view, with an emphasis on the augmented RoBERTa and Mistral + SVM models, which yielded the best generalization results.

Results by Dataset. Figure 2 highlights the variation in improvement by dataset within the Whisper transcript results only (due to the absence of comparative human transcripts for the Physics Playground and Minecraft Hour of Code datasets). For each dataset, we see larger improvements from the augmented RoBERTa model than the Mistral + SVM model, however in each case there is positive improvement from both types. Interestingly, the models applied to the Minecraft Hour of Code dataset showed the most substantial improvements over baseline. The other three datasets showed approximately equal improvements with the augmented RoBERTa approach. These results may reflect the fact that the Minecraft dataset had greater linguistic and contextual differences from the source Sensor Immersion dataset, allowing for greater gains from augmentation, whereas datasets like Self Driving Cars that are more closely aligned with the original training data, resulted in smaller relative improvements.

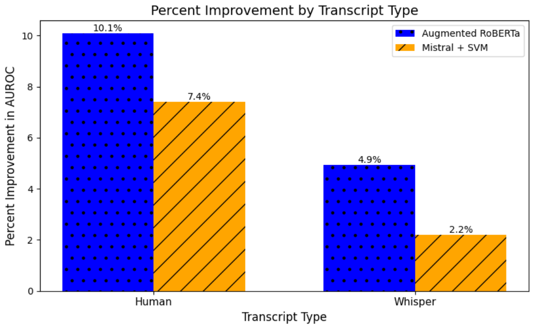

Figure 4: Bar charts showing the percent improvement in AUROC over the baseline RoBERTa model for augmented RoBERTa (blue) and Mistral + SVM (orange) across transcript type. Improvements are averaged over CA and test dataset.

Results by CA. Figure 3 shows that, when the percent improvement in AUROC for Whisper transcripts is split by CA and averaged across datasets, the augmented RoBERTa model consistently outperforms the baseline RoBERTa model, with the Mistral + SVM model generally achieving smaller improvements and showing an average negative improvement for the Respect CA. The largest improvement over baseline is observed in the Community CA, followed by Thinking and finally Respect. The larger improvements for Community and Thinking compared to Respect likely come from the fact that the baseline RoBERTa models for these categories were more prone to overfitting to domain-specific words as we elaborate below (e.g., Figure 5). Since Respect exhibited less overfitting in the baseline model, there was less room for improvement when applying advanced techniques.

Results by Transcript Type. Figure 4 presents the results for the classroom datasets (Self Driving Cars and Moderation Unit) only, disaggregated by transcript type. These datasets allow us to directly compare performance differences between human and Whisper transcripts. The results reveal discrepancies in model performance based on transcript type, with human transcripts mostly exhibiting greater improvements. Whisper transcripts also showed improved results, albeit to a lesser extent. This trend suggests that noise introduced by mistranscriptions from Whisper may hinder model performance more generally.

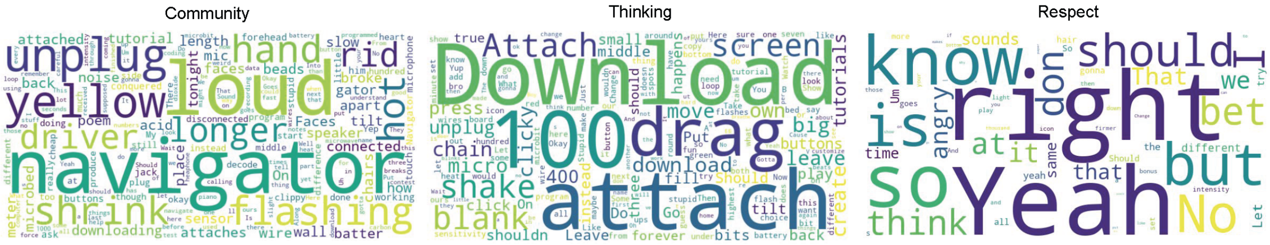

Figure 5: Word clouds generated from the most important words from the Sensor Immersion dataset for the Community, Thinking, and Respect RoBERTa models, as ascertained by the LIME technique.

CA	Example Utterances (Domain)
Community	"Forever do move towards the sheep" (*MU*)
	"Also, wait, mine was artificial intelligence and paid moderators" (*MU*)
	"So first one is using an if else statement" (*SDC*)
	"So I have no idea how to put these wheels on" (*SDC*)
	"...because we only have like fifteen blocks so" (*MHC*)
	"Place bricks, place bricks or something" (*MHC*)
	"I think you should keep doing that place bedrock ahead" (*PP*)
	"I'm just gonna guess the yellow one goes on, yep exactly" (*PP*)
Thinking	"Click space to use them to drop the move" (*MU*)
	"Player and then run" (*MU*)
	"We should do that one following a line" (*SDC*)
	"We need the connector thingy, right?" (*SDC*)
	"You would go like turn nine forward, forward, destroy forward" (*MHC*)
	"Okay so then place bedrock flattened to" (*MHC*)
	"I think with all that we can always just put the repeat three times" (*PP*)
	"I think for this one so press the water, move forward and then go to the loft" (*PP*)
Respect	"I actually did it! I like your Minecraft games, they're fun" (*MU*)
	"Right there, right there, right there. You're good" (*MU*)
	"I thought it was fine. Slow and steady wins the race" (*SDC*)
	"Can you look up IR sensor please? What does it do?" (*SDC*)
	"Yeah, [...] I think it'll work if it's the right dimensions" (*PP*)
	"No I don't think so but I can delete this" (*PP*)

SDC: Self Driving Cars, MU: Moderation Unit, MHC: Minecraft Hour of Code, PP: Physics Playground

Table 6: True positive examples where the baseline RoBERTa model predicted false, but the augmented RoBERTa and Mistral + SVM models successfully predicted true. Two utterances were chosen per domain to illustrate the generalizability issues.

6.3 Qualitative Results

Sensor Immersion Dataset. To better understand the performance discrepancies, we also conducted qualitative error analyses. Using the Local Interpretable Model-agnostic Explanations (LIME) technique [40], we analyzed the importance of individual words in the Sensor Immersion dataset for the baseline RoBERTa model and visualized them as word clouds in Figure 5. These word clouds depict the words most frequently identified as critical by LIME, weighted by the frequency with which they were *the most* important word in a positively classified utterance. The analysis revealed that the RoBERTa model was overfitting to curriculum-specific language, particularly for the Community and Thinking CAs. Words and phrases unique to the Sensor Immersion curriculum such as "navigator", "download", and "attach" disproportionately influenced the model's predictions. This overfitting was less pronounced for the Respect CA, where the model appeared to rely on broader contextual signals rather than curriculum-specific terms. This suggests that the linguistic features of Respect are more generalizable, while Community and Thinking may require additional intervention to mitigate overfitting.

Transfer Datasets. In several cases where domain-specific wording was used in positive examples of CAs, the baseline RoBERTa model failed to correctly classify utterances, while the augmented RoBERTa and Mistral + SVM models demonstrated improved performance. Table 6 presents examples from the transfer datasets where the baseline model misclassified positive examples, but either one of the two improved models identified them accurately. For instance, in the Physics Playground dataset, the baseline model misclassified the phrase "I think you should keep doing that. Place bedrock ahead" as a negative example of Community, but the other two models identified it positively, as it is an example of talking about the challenge situation and providing instructional support. A positive example of Thinking from the Minecraft Hour of Code dataset, "You would go like turn nine forward, forward, destroy forward", was incorrectly predicted by the baseline RoBERTa but correctly identified by both improved techniques. This example satisfied the proposing (in)correct solutions and strategizing to accomplish task goals indicators. These examples illustrate the ability of the augmented RoBERTa and Mistral + SVM models to capture nuanced behaviors across domains.

7. DISCUSSION

Building scalable models for collaborative discourse analysis requires a balance between domain-specific accuracy and cross-context generalizability. While fine-tuned models excel within their training domains, their transfer to new contexts remains a significant challenge. This study explored strategies to mitigate overfitting and enhance model generalizability, revealing key insights into scalable approaches to improve generalization across diverse educational settings.

7.1 Main Findings

Model overfitting highlights a challenge in achieving broad generalization, but this work suggests that models fine-tuned on specific datasets can be highly effective in specialized or localized educational settings with high consistency in language and tasks. These models provide benefits in such contexts as they can deliver precise, context-sensitive insights. On the other hand, the success of the augmented RoBERTa approach, which leveraged contextual embeddings and data augmentation to contextualize and diversify training data, demonstrated robust classification across various domains with minimal overfitting. Practically, these findings underscore the potential of lightweight, adaptable methods for scenarios where large, diverse datasets are unavailable.

The observed differences between the model types reveal practical challenges of generalizing models that are trained on curriculum-specific datasets. While large pre-trained models like Mistral hold potential due to their vast knowledge base, our results suggest that more focused methods, such as embeddings combined with simple classifiers or augmentation strategies, may be more effective for tasks requiring adaptation. Our results align with previous research suggesting that more traditional language models may outperform LLMs in classifying short sequences of text, especially in more theoretically grounded scenarios [26, 19].

Some aspects of collaboration, particularly those tied to domain-specific language and problem solving, appear more susceptible to overfitting, while others that rely on more universal social and affective signals generalize more readily. These findings suggest that collaboration is not a homogenous construct from a modeling perspective; rather, it encompasses dimensions with varying levels of linguistic and contextual dependency. This work underscores the implications for how we design, train, and evaluate models for diverse educational applications.

7.2 Limitations & Future Work

This study is not without its limitations. A significant challenge was the high word error rate encountered in Whisper transcripts, which negatively affected model performance. Noise and inaccuracies during both the training and testing phases compounded the difficulty of drawing reliable conclusions, particularly for ASR-generated transcripts. Additionally, while the five datasets used in this study represent a range of contexts, they do not encompass the full diversity of classroom environments and demographics. Expanding the scope of datasets to include broader cultural, socioeconomic, and instructional diversity remains a priority for future research. Finally, due to privacy of the student data, we were restricted to the use of open source models and unable to harness more recent state-of-the-art models.

Several domain-specific factors complicated cross-context generalization. For instance, the Minecraft Hour of Code and Physics Playground datasets involved older students whose discourse reflected more structured syntax, colloquialisms, and specialized task jargon compared to middle school classrooms. Similarly, transfer datasets introduced technical language and task-specific elements, such as a reset button in the Moderation Unit or references to levers and pendulums in Physics Playground. These variations underscore the linguistic and contextual diversity models must navigate to achieve broad generalization. To mitigate these challenges, we refined the coding process and adapted our annotation framework to account for domain-specific features, ensuring the robustness of CA labels and providing a strong foundation for accurate downstream analysis.

Future work should explore advanced techniques, such as adversarial models and multi-task learning, to further enhance transferability. These approaches could enable models to adapt more effectively to diverse classroom contexts. We aim to investigate the application of these models in real-time classroom settings, integrating them with classroom technologies to facilitate formative assessments and feedback loops. Such efforts would provide direct evaluations of the practical impact of these tools on teaching and learning outcomes.

8. CONCLUSION

This study underscores the challenges of fine-tuning large language models for domain-specific tasks, and their susceptibility to overfitting when transitioning between distinct curricula and contexts. While fine-tuned RoBERTa performed well within its training domain, its decline on held-out datasets highlights the fragility of traditional fine-tuning approaches. In contrast, augmenting training data and leveraging Mistral embeddings with an SVM classifier resulted in more robust performance across diverse environments. These methods balanced generalizability and domain-specific adaptability, demonstrating their potential as scalable solutions for real-world applications. The underperformance of prompting and fine-tuning a Mistral model (despite its sophisticated architecture) highlights the difficulty of training large-scale models on specialized qualitative datasets. Their inability to learn the nuanced patterns of a collaborative framework suggests that careful alignment between training data size and methodologies is crucial.

Our work highlights strategies for enhancing the robustness of LLMs and stresses the importance of methodological rigor in adapting models for educational applications. By emphasizing training data augmentation and lightweight classifiers, we provide actionable insights for future research and for practitioners aiming to implement automated collaboration analysis tools across varied educational contexts.

9. ACKNOWLEDGMENTS

This research was jointly supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and Learning Engineering Virtual Institute (LEVI). The opinions expressed in this paper are those of the authors only and do not reflect the views of the NSF or LEVI.

10. REFERENCES

[1] OpenSciEd. http://www.openscied.org/. Accessed: 2025-02-12.

[2] E. Altinisik, H. Sajjad, H. Sencar, S. Messaoud, and S. Chawla. Impact of adversarial training on robustness and generalizability of language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7828–7840, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] J. Andrews-Todd, J. Steinberg, M. Flor, and C. M. Forsyth. Exploring automated classification approaches to advance the assessment of collaborative problem solving skills. *Journal of Intelligence*, 10(3), 2022.

[4] T. Breideband, J. Bush, C. Chandler, M. Chang, R. Dickler, P. Foltz, A. Ganesh, R. Lieber, W. R. Penuel, J. G. Reitman, J. Weatherley, and S. D'Mello. The community builder (cobi): Helping students to develop better small group collaborative learning skills. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 376–380, New York, NY, USA, 2023. Association for Computing Machinery.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and A. Neelakantan. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[6] J. Cao, A. Ganesh, J. Cai, R. Southwell, M. E. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 250–262, New York, NY, USA, 2023. Association for Computing Machinery.

[7] A. G. Chakarov, Q. Biddy, C. H. Elliott, and M. Recker. The data sensor hub (dash): A physical computing system to support middle school inquiry science instruction. *Sensors*, 21(18), 2021.

[8] C. Chandler, T. Breideband, J. G. Reitman, M. Chitwood, J. B. Bush, A. Howard, S. Leonhart, P. W. Foltz, W. R. Penuel, and S. K. D'Mello. Computational modeling of collaborative discourse to enable feedback and reflection in middle school classrooms. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 576–586, New York, NY, USA, 2024. Association for Computing Machinery.

[9] P. Chejara, L. P. Prieto, M. J. Rodriguez-Triana, R. Kasepalu, A. Ruiz-Calleja, and S. K. Shankar. How to build more generalizable models for collaboration quality? lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, page 111–121, New York, NY, USA, 2023. Association for Computing Machinery.

[10] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online, Nov. 2020. Association for Computational Linguistics.

[11] D. Demszky and H. Hill. The NCTE transcripts: A dataset of elementary math classroom transcripts. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] S. K. D'Mello, N. D. Duran, A. Michaels, and A. E. Stewart. Improving collaborative problem solving skills via automated feedback and scaffolding: A quasi-experimental study with cpscoach 2.0. In *User Modeling and User-Adapted Interaction*, 2024.

[13] C. H. Elliott, J. Nixon, J. B. Bush, A. G. Chakarov, and M. Recker. "do i need to know what i am doing if i am the teacher?" developing teachers' debugging pedagogies with physical computing. *International Conference of the Learning Sciences*, 2021.

[14] M. Emara, N. M. Hutchins, S. Grover, C. Snyder, and G. Biswas. Examining student regulation of collaborative, computational, problem-solving processes in open-ended learning environments. *Journal of Learning Analytics*, 8(1):49 – 74, 2021.

[15] J. S. Ernst, S. Marton, J. Brinkmann, E. Vellasques, D. Foucard, M. Kraemer, and M. Lambert. Bias mitigation for large language models using adversarial learning. In *Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023*, Krakow, Poland, 2023.

[16] S. M. Fiore, A. Graesser, and S. Greiff. Collaborative problem-solving education for the twenty-first-century workforce. *Nature human behaviour*, 2(6):367–369, 2018.

[17] M. Flor and J. Andrews-Todd. Towards automatic annotation of collaborative problem-solving skills in technology-enhanced environments. *Journal of Computer Assisted Learning*, 38(5):1434–1447, 2022.

[18] M. Flor, S.-Y. Yoon, J. Hao, L. Liu, and A. von Davier. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 31–41, San Diego, CA, June 2016. Association for Computational Linguistics.

[19] A. Ganesh, C. Chandler, S. D'Mello, M. Palmer, and

K. Kann. Prompting as panacea? a case study of in-context learning performance for qualitative coding of classroom dialog. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 835–843, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[20] A. C. Graesser, S. M. Fiore, S. Greiff, J. Andrews-Todd, P. W. Foltz, and F. W. Hesse. Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2):59–92, 2018. PMID: 30497346.

[21] P. Griffin and E. Care. *The ATC21S Method*, pages 3–33. Springer Dordrecht, 09 2015.

[22] A. Gurria. Pisa 2015 results in focus. *PISA in Focus*, (67):1, 2016.

[23] K. L. Gwet. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[24] J. Hao, L. Chen, M. Flor, L. Liu, and A. A. von Davier. Cps-rater: Automated sequential annotation for conversations in collaborative problem-solving activities. *ETS Research Report Series*, 2017(1):1–9, 2017.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[26] Huggingface. Comparing the performance of llms: A deep dive into roberta, llama 2, and mistral for disaster tweets analysis with lora, November 2023.

[27] D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmer, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, and Z. Jin. A taxonomy and review of generalization research in nlp. *Nat Mach Intell*, 5:1161–1174, 2023.

[28] H. Jeong, C. E. Hmelo-Silver, and K. Jo. Ten years of computer-supported collaborative learning: A meta-analysis of cscl in stem education during 2005–2014. *Educational Research Review*, 28:100284, 2019.

[29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[30] A. Kupor, C. Morgan, and D. Demszky. Measuring five accountable talk moves to improve instruction at scale, 2023.

[31] J. Lämsä, P. Uribe, A. Jiménez, D. Caballero, R. Hämäläinen, and R. Araya. Deep networks for collaboration analytics: Promoting automatic analysis of face-to-face interaction in the context of inquiry-based learning. *Journal of Learning Analytics*, 8(1):113 – 125, 2021.

[32] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), Jan. 2023.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[34] OECD. *PISA 2015 Assessment and Analytical Framework*. 2017.

[35] OpenAI, J. Achiam, and S. A. et al. Gpt-4 technical report, 2023.

[36] S. L. Pugh, A. Rao, A. E. Stewart, and S. K. D'Mello. Do speech-based collaboration analytics generalize across task contexts? In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 208–218, New York, NY, USA, 2022. Association for Computing Machinery.

[37] S. L. Pugh, S. K. Subburaj, A. R. Rao, A. E. Stewart, J. Andrews-Todd, and S. K. D'Mello. Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)*, EDM21, page 55–67, 2021.

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. Technical report, OpenAI, 2022.

[39] D. Ramdani, H. Susilo, S. Suhadi, and S. Sueb. The effectiveness of collaborative learning on critical thinking, creative thinking, and metacognitive skill ability: Meta-analysis on biological learning. *European Journal of Educational Research*, 11(3):1607–1628, 2022.

[40] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[41] J. Roschelle, Y. Dimitriadis, and U. Hoppe. Classroom orchestration: Synthesis. *Computers and Education*, 69:523–526, 2013.

[42] C. P. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3:237–271, 2008.

[43] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4323–4330. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[44] C. Shi, Y. Su, C. Yang, Y. Yang, and D. Cai. Specialist or generalist? instruction tuning for specific NLP tasks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15336–15348, Singapore, Dec. 2023. Association

for Computational Linguistics.

[45] R. Southwell, S. Pugh, E. M. Perkoff, C. Clevenger, J. Bush, R. Lieber, W. Ward, P. Foltz, and S. D'Mello. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 302–315, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[46] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[47] J. Steinberg, C. Forsyth, and J. Andrews-Todd. An exploratory approach to predicting performance in an online electronics collaborative problem-solving task. *ETS Research Report Series*, 2024(1):1–12, 2024.

[48] A. E. Stewart, M. J. Amon, N. D. Duran, and S. K. D'Mello. Beyond team makeup: Diversity in teams predicts valued outcomes in computer-mediated collaborations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[49] A. E. Stewart, Z. Keirn, and S. K. D'Mello. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 31(4):713–751, 2021.

[50] A. E. Stewart, H. Vrzakova, C. Sun, J. Yonehiro, C. A. Stone, N. D. Duran, V. Shute, and S. K. D'Mello. I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[51] C. Sun, V. J. Shute, A. Stewart, and S. D'Mello. The relationship between collaborative problem-solving skills and group-to individual learning transfer in a game-based learning environment. In *In Proceedings of the 2025 International Conference on Learning Analytics Knowledge (LAK25) ACM.*, in press.

[52] C. Sun, V. J. Shute, A. Stewart, J. Yonehiro, N. Duran, and S. D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers and Education*, 143:103672, 2020.

[53] C. Sun, V. J. Shute, A. E. Stewart, Q. Beck-White, C. R. Reinhardt, G. Zhou, N. Duran, and S. K. D'Mello. The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, 128:107120, 2022.

[54] M. Tissenbaum and J. D. Slotta. *Scripting and Orchestration of Learning Across Contexts: A Role for Intelligent Agents and Data Mining*, pages 223–257. Springer Singapore, Singapore, 2015.

[55] S. Vayssettes et al. *PISA 2015 assessment and analytical framework: science, reading, mathematic and financial literacy.* OECD publishing, 2016.

[56] Y. Wang, S. Si, D. Li, M. Lukasik, F. Yu, C.-J. Hsieh, I. S. Dhillon, , and S. Kumar. Two-stage llm fine-tuning with less specialization and more generalization. 2023.

[57] H. Yang, Y. Zhang, J. Xu, H. Lu, P.-A. Heng, and W. Lam. Unveiling the generalization power of fine-tuned large language models. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 884–899, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[58] G. Zhou, R. Moulder, C. Sun, and S. D'Mello. Investigating temporal dynamics underlying successful collaborative problem solving behaviors with multilevel vector autoregression. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 290–301, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[59] M. Zhu, X. Wang, X. Wang, Z. Chen, and W. Huang. Application of prompt learning models in identifying the collaborative problem solving skills in an online task. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), Nov. 2024.