Cross-Domain Classification of Education Talk-Turns

Achyutarama R. Ganti and **Steven R. Wilson**
University of Michigan-Flint
{arganti,steverw}@umich.edu

Wing-Yue Geoffrey Louie
Oakland University
louie@oakland.edu

Abstract

The study of classroom discourse is essential for enhancing child development and educational outcomes in academic settings. Prior research has focused on the annotation of conversational talk-turns within the classroom, offering a statistical analysis of the various types of discourse prevalent in these environments. In this work, we explore the generalizability and transferability of text classifiers trained to predict these discourse codes across educational domains. We examine two distinct English-language classroom datasets from the domains: literacy and math. Our results show that models exhibit high accuracy and generalizability when the training and test datasets originate from the same or similar domains. In situations where limited training data is available in new domains, few shot and zero shot exhibit more resiliency and are less effected than their supervised counterparts. We also observe that accompanying each talk turn with dialog-level context improves the accuracy of generative models. We conclude by offering suggestions on how to enhance the generalization of these methods to novel domains, proposing directions for future studies to investigate new methods for boosting model adaptability across domains.

1 Introduction

In recent years, computational approaches have increasingly demonstrated their potential to capture and analyze discourse-level features within educational settings (Ganesh et al., 2021). Previous research in this domain has provided valuable insights, particularly in the context of specific educational domains or settings (Wang et al., 2023). However, these studies often limit their focus to a single domain, and the adaptability and effectiveness of these models across varied educational contexts is less well understood. In many cases, it is challenging to obtain large amounts of training data for the exact educational setting in which a system is intended to be deployed. Hence, it is crucial to explore how different approaches perform in scenarios with limited or mismatched training data, to better assess their robustness and transferability across educational domains and contexts.

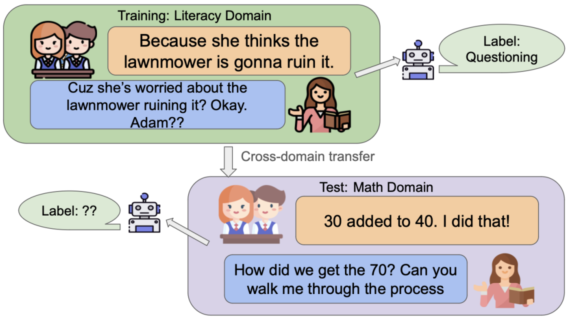

Training: Literacy Domain
Because she thinks the lawnmower is gonna ruin it.
Cuz she's worried about the lawnmower ruining it? Okay. Adam??
Label: Questioning
Cross-domain transfer
Test: Math Domain
Label: ??
30 added to 40. I did that!
How did we get the 70? Can you walk me through the process

Figure 1: Cross Domain Training in an educational context where model trained on classroom discourse from the literacy domain is applied to the math domain to predict talk moves in a new context.

Addressing this gap, we investigate a range of models, from fine-tuned classifiers to in-context learning approaches with generative language models. Our focus is to evaluate how the performance of these models varies as the distance between the contexts of the training and test domains increases. By 'distance,' we refer not only to the difference in academic domains (e.g., English vs. Mathematics) but also to differences in educational materials such as textbooks used and instructional variations among teachers. We explore the generalizability of various computational methods across diverse educational settings as illustrated in Figure 1.

Our primary contributions are (1) an analysis of the generalizability of language models across educational domains, providing insights into the adaptability and limitations of these models when applied in different educational contexts, especially in low-resource settings; (2) experiments evaluating two classes of models, i.e., fine-tuned transformer-based encoder models and in-context

Proceedings of the 31st International Conference on Computational Linguistics, pages 6897–6917
January 19–24, 2025. ©2025 Association for Computational Linguistics

learning approaches, for classifying educational talk-turns with discourse codes; and (3) the creation of a new evaluation set for testing the generalizability and accuracy of various language models for cross-domain classification of discourse codes.

Using data ranging from read-aloud discussions in early education classrooms to interactions from mathematics classes, we investigate how model performance varies across contexts. We aim to shed light on the strengths and limitations of current computational approaches in educational discourse analysis. Our results demonstrate that models exhibit accurate discourse code classification when the training and test datasets originate from the same domains, but as expected, the effectiveness of these models begins to decrease as the training and testing scenarios become more dissimilar. Furthermore, in low-resource settings within new domains, in-context learning models exhibit a degree of resilience, with the performance gap between them and their supervised counterparts being less drastic.

Despite this decline in performance, exploration of fine-tuned and in-context learning models in cross-domain scenarios remains crucial in this area of study in order to precisely quantify the *extent* of this performance dip, especially given recent advancements in large language models (LLMs). Further, we investigate how the choice of model or the inclusion of additional information such as conversational context might help mitigate this drop-off, seeking to highlight which computational approaches might exhibit greater resilience against the challenges posed by domain variance, thereby contributing to the generalizability of these models across diverse educational settings.

2 Related Work

The dynamics of student-teacher classroom discourse play a pivotal role in shaping the experience and outcomes of students. Several papers have studied this phenomenon, particularly in the context of K-12 mathematics education and other childhood learning environments. For example, Suresh et al. (2022a) found that sustained classroom discourse is a critical component of equitable and a rich learning environment. Towards that goal, they built an extensive collection of human-annotated transcripts from K-12 classroom mathematics lessons as they can be effective tools for understanding discourse patterns in classroom instructions. Furthermore, Demszky et al. (2021) argue that teachers' acknowledgement, repetition and reformulation of students' responses has been linked to higher student engagement and achievement. The impact of building upon student contributions in the classroom is explored in studies by Brophy and Good (1984) and Faculty and Michaels (1993). They demonstrate that acknowledgment, repetition, and elaboration of student inputs can significantly enhance student learning and academic achievement. Wright (2019) delves into the significance of read-aloud activities in nurturing children's reading skills and knowledge. They deduce that engaging in interactive read-alouds is beneficial for children in acquiring new vocabulary, understanding textual functions, and developing a diverse set of skills essential for independent reading. Giroir et al. (2015) explore effective methodologies for implementing read-aloud programs. Their research particularly focuses on integrating aspects of second language acquisition and culturally responsive teaching methods, outlining critical steps and applications for an effective read-aloud strategy.

In the context of early childhood education, Christ et al. (2023) study the interactions between teacher and child talk-turns during read-aloud sessions. The statistical discourse analysis conducted in this study provides insights into how certain talk-turns can influence children's comprehension responses, thereby emphasizing the critical role of teacher mediation in shaping learning outcomes. Their findings also demonstrate that when children's talk-turns mediate other children's actions, they act as a predictor for those children's subsequent responses in terms of comprehension.

Given the breadth of actionable findings in this area, a promising direction is to develop tools that assist teachers in refining their instructional strategies. Suresh et al. (2022b) outline the development of the TalkBack application. Their tool leverages deep learning capabilities to provide teachers with automated feedback on their discourse strategies, highlighting the importance of automated feedback to enhance and enrich teacher learning. Specifically, it aids in refining instructional strategies, thereby enhancing the learning environment.

In recent years, advancements in NLP have opened new and effective means of analyzing and enhancing classroom discourse. Ganesh et al. (2021) aims to enhance classroom learning and engagement by developing a system to predict the next talk move (an utterance strategy) in a classroom discussion, based on the academically productive talk (APT) framework. They present a neural network model aimed at predicting the next talk move in a conversation based on its history and associated talk moves potentially leading to more interactive and personalized learning experiences. In this study (Suresh et al.) incorporate enriched contextual cues from previous and subsequent utterances using a RoBERTa model to improve the automated classification of "talk moves" in educational settings. Similarly, (Alic et al., 2022) address the task of creating specific types of questions that promote responsive teaching. The authors created an annotated dataset and employed various supervised and unsupervised learning methods to demonstrate the importance of incorporating computational tools to assist teachers in refining their instructional techniques. Tran et al. (2024) explore how task formulation, context length, and few-shot examples impact the performance of two large language models (LLMs) in assessing classroom discussion quality.

While existing work demonstrated the effectiveness of computational tools to assist in classroom settings, these have typically focused on a single domain. However, in order to use these tools broadly, they must generalize across topical areas and classroom contexts. Therefore, we set out to evaluate the extent to which current methods are able to accurately transfer to new contexts, in terms of both classrooms and teaching domains.

3 Data

In this study, we leveraged existing classroom discourse datasets comprising turn-level student-teacher interactions that were annotated by educational experts. To investigate the generalizability of these codes across different academic domains, we re-annotated a small subset from each dataset using the discourse codes that were originally developed for the other datasets.

3.1 Datasets Used

We used four existing English-language datasets: The MuMo Talk moves dataset (Christ et al., 2023), the National Center for Teacher Effectiveness (NCTE) Transcripts dataset (Demszky and Hill, 2023), and two additional datasets referred to by the names of pseudonymous teachers of their respective classrooms: Mason (Christ and Cho, 2023) and Newman (Cho and Christ, 2022).

The **MuMo** Talk moves dataset includes three kindergarten teachers' interactive read-alouds comprising of 736 talk-turns across six video recorded and transcribed sessions. The talk-turns were coded using *a priori* and emergent codes. The authors grouped these codes into higher level categories of the talk-turns. We utilize these high level categories as output labels for our experiments.

In the **Mason** dataset (Christ and Cho, 2023), the authors investigated the engagement of four second-grade emergent bilingual students and their teacher with listening comprehension during interactive read-aloud sessions. These sessions used books with varying levels of cultural relevance. The study aimed to understand how this engagement related to the teacher's implementation of culturally relevant and sustaining pedagogical practices. To conduct the analysis, the researchers collected data through cultural relevance ratings of the books, video recordings, and transcripts of nine 20-minute lessons, resulting in a total of 2781 talk-turns.

The **Newman** dataset investigates how two emergent bilingual student groups from refugee families interacted with the same culturally relevant book as used in the Mason dataset, though with a different teacher and students, in a third-grade class in the Midwest U.S. Using video recordings and transcripts of 12 read-aloud discussions, interviews, and cultural relevance ratings, this study analyzed the students' inference-making processes and examined their use of text information, background knowledge, and the coherence of their inferences. This dataset had a total of 2470 talk-turns.

The **NCTE** transcripts are the largest dataset of mathematics classroom transcripts available. The dataset consists of 580408 anonymous transcripts of whole lessons collected as part of the National Center for Teacher Effectiveness, NCTE study, spanning across the K-12 math classrooms across four districts serving largely historically marginalized students. However, 2348 transcripts were annotated for the classification experiments and analysis.

Initially, we had three distinct codebooks that had been used to annotate the source datasets: one for MuMo (Christ et al., 2023), and second that was used for both Newman and Mason, and NCTE, being a math dataset, had a third. In assembling our datasets for this study, we adopted the codebook which was developed for MuMo to use for both Newman and Mason after recognizing the similar-

Variable	MuMo	Mason	Newman
Response Evaluation	67	345	656
Providing Information	115	290	344
Revoicing	113	330	335
Strategy Related	62	206	302
Questioning	219	563	503
Behavior Management	55	326	354
Turn Management	207	283	325
Total	**736**	**2550**	**2467**

Table 1: Number of talk turns with each label across MuMo, Mason, and Newman datasets. **Total** indicates the number of talk turns in the dataset. Note that each talk turn may have more than one label.

ity in the codes across the three datasets. While the datasets have distinct sets of fine-grained codes, in this work, we experiment only with higher-level categories rather than these fine-grained codes. This makes the process of matching the high-level categories across codebooks fesible, and allows for a comprehensive cross-dataset analysis. Refer Table 20 for details on how the MuMo codebook was adopted and aligned to the Mason and Newman datasets. Table 1 shows the class distribution of variables belonging to Class 1 across the MuMo, Mason, and Newman datasets using our merged codebook, and Table 2 shows the class distribution of variables belonging to Class 1 across the NCTE dataset.

3.2 Annotation Process

To investigate the cross-domain generalizability of our classifiers, we sampled data from each domain to re-annotate according to the codebook from the other domain as a new evaluation set. Specifically, we chose 140 data points from the MuMo, Mason, and Newman datasets combined, selecting 10 talk-turns from each session. MuMo contributed data from 6 sessions, while Mason and Newman each had 4 sessions, collectively providing the 140 data points. From the NCTE dataset, which is comprised of a single extensive session, we sampled a total of 100 data points. Once the annotation guidelines (see Appendix A) were established, five trained annotators, including the first two authors, re-annotated these subsets. With the goal of creating a ground truth for evaluating the language models, the annotators applied the discourse codes from the math domain dataset (i.e., NCTE) to the literacy domain datasets (i.e., MuMo, Mason, and Newman) and vice versa. Each talk-turn was annotated by at least three annotators to ensure reliable accuracy and consistency. The inter-annotator agreement was quantitatively measured using Krippendorff's alpha (Krippendorff, 2011). Annotators had high agreement (average Krippendorff's alpha = 0.883, per category results in Appendix D). In case of discrepancies among annotators, the label that received the majority consensus among the three annotators was chosen as the final label for each talk-turn in our test set.

Variable	Count
Student on Task	1964
Teacher on Task	2004
High Uptake	813
Focusing Question	359
Total	**2348**

Table 2: Number of talk turns with each label in the NCTE dataset. **Total** indicates the number of talk turns in the dataset. Note that each talk turn may have more than one label.

4 Experimental Methodology

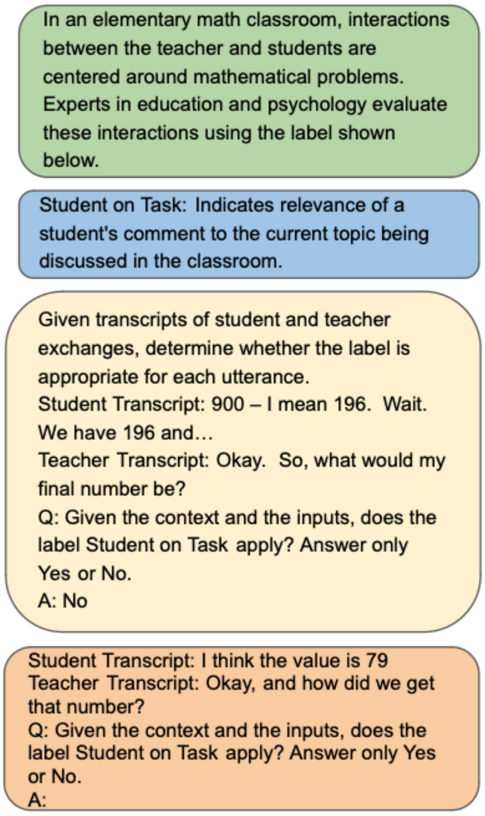
For in-domain experiments, the NCTE dataset was partitioned using an 80-10-10 split for training, validation, and test data. In the case of the MuMo, Mason, and Newman datasets (those from the literacy domain), our experimental design included two setups, both of which relied on splitting across entire sessions rather than utterances. First, for within-dataset experiments, we held out one entire session to function as the test set. Secondly, for experiments within the same domain but across different datasets, the same held-out session from the target domain was used as the test set. This approach allowed us to examine both the domain-specific and cross-domain efficacy of our models.

For cross-domain experiments, we use one dataset as both the training and validation set, while a dataset from a different domain is designated as the test set. This strategy was applied to explore the adaptability of models across varied educational contexts without any prior data available in the target context.

We investigated both fine-tuned transformer encoder models and auto-regressive generative models focusing on in-context learning.[1] The model hyperparameters are specified in Appendix C.

For transformer-based deep learning models, we chose BERT (Devlin et al., 2019) and RoBERTa

[1] We also explored classical machine learning approaches using bag-of-words features, but found these to always underperform the transformer-based approaches.


In an elementary math classroom, interactions between the teacher and students are centered around mathematical problems. Experts in education and psychology evaluate these interactions using the label shown below.

Student on Task: Indicates relevance of a student's comment to the current topic being discussed in the classroom.

Given transcripts of student and teacher exchanges, determine whether the label is appropriate for each utterance.
Student Transcript: 900 – I mean 196. Wait. We have 196 and…
Teacher Transcript: Okay. So, what would my final number be?
Q: Given the context and the inputs, does the label Student on Task apply? Answer only Yes or No.
A: No

Student Transcript: I think the value is 79
Teacher Transcript: Okay, and how did we get that number?
Q: Given the context and the inputs, does the label Student on Task apply? Answer only Yes or No.
A:

Figure 2: Prompt components for generative models for the math setting (used for NCTE). From top to bottom, the blocks display the background information (green), labels (blue), few-shot examples (yellow).

(Liu et al., 2019), using pre-trained weights and fine-tuning code from the HuggingFace transformers Wolf et al. (2019) library. We utilized the `bert-base-uncased` and `roberta-base` checkpoints along with their default tokenizers. The output from the [CLS] input token was then used as the input for a trainable classification layer.

For the generative models, we opted to use the Llama2-7B (Touvron et al., 2023) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)[2] models specifically due to their open weights[3] availability. We use open-weights models to increase transparency and reproducibility, and also to avoid leakage of datasets via the use of online APIs without the consent of the participants of the original studies whose interactions led to the creation of the datasets (Balloccu et al., 2024). The generative models operate by receiving an instruction or a prompt as input and generating a response that aligns with the given context or question. By using in-context learning, we investiage whether these models are capable of predicting talk-turn labels, particularly in scenarios where there is limited data availability.

The experimental setup for the auto-regressive models was conducted in both a zero-shot and a few-shot learning context. In the zero-shot setup, the background information and label description are prepended to the prompt followed by a section of the transcript to be classified. Finally we ask the model if the given label is appropriate for the transcript. The model is constrained to answer only in a Yes or No format for calculating accuracy, F1 score and other metrics. We repeated the experiments with same prompt three times to check for any variability in the model's outputs.

In the few-shot setup, several example interactions between teachers and students, along with their correct labels, were prepended to the prompt in a question-answer format, along with the background information and label description. Similar to the zero-shot setup, experiments were conducted using three different prompts for both Llama2 7B and Mixtral 8x7B. In case of the generative models, the average of the three turns of the best performing prompt was reported. A summary of the various components utilized in this setup can be found in Figures 2. Also refer to Figure 5 in Appendix F.

We also experimented with varying the number of prior talk turns provided as context. This setup was only applied to the generative models due to the input size limitations of 512 tokens for the BERT-based models, which was typically not large enough to include additional context. When incorporating context, each talk turn was accompanied by the preceding one, three, or five interaction(s) and the speaker tag (whether the talk turn was uttered by a teacher or a student).

5 Results

Figure 3 and Tables 16, 17, 18, and 19 present the performance of different classes of models averaged across all output labels for the domains of literacy and mathematics. Please refer to appendix G for a detailed breakdown of results for each label and each model. Among these, fine-tuned models, *i.e.*, BERT and RoBERTa, demonstrated better performance in most of the experiments over generative models with in-context learning. But the generative models outperformed supervised models when added context for certain variables.

[2] Henceforth referred to simply as "Mixtral."

[3] We distinguish between "open source" and "open weights", where the former includes cases where all code to fully reproduce the model is available, while the latter refers to the open availability of the trained model's parameters.

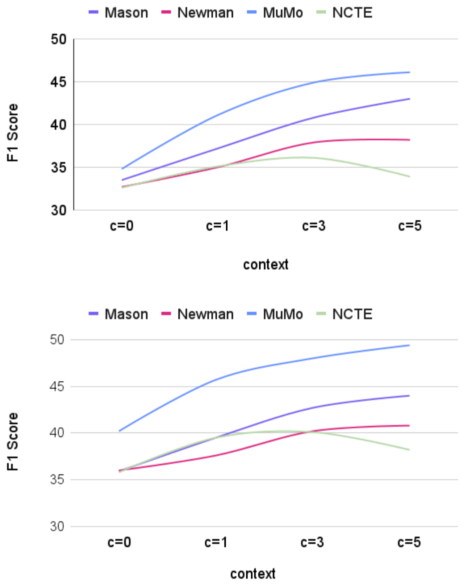


Figure 3: F1 scores for Llama2 (**top**) and Mixtral (**bottom**) models averaged across different training sets for the various test sets Mason, Newman, MuMo and NCTE. **c** denotes the number of prior interactions provided as context to the generative models as part of in-context learning during classification.

Test Set	Model	Student on Task	Teacher on Task	High Uptake	Focusing Question
NCTE	Mixtral	0.702	0.545	0.241	0.584
	Llama2	0.666	0.524	0.217	0.492

Table 3: Zero-shot performance of Mixtral and Llama2 models on the NCTE test set across four categories.

Among the generative models, Mixtral outperformed Llama2 in most scenarios. Interestingly, Mixtral, when prompted with prior interactions, outperformed BERT and RoBERTa models for certain variables in the binary classification tasks, refer 16. Despite their overall lower performance compared to fine-tuned models, the fact that these generative models utilized far fewer training data (few shots with $n = 3$ and context $c = \{3, 5\}$) while learning highlights their potential in specific contexts. Tables 5 and 6 show a decline in model performance when tested on a new domain however, generative models showed resilience.

Another takeaway is within the generative models, when it comes to variables like turn management, behavior management and questioning, the zero-shot experiments achieved closer results to that of the few shot models indicating that the models do not need much data for classification. The performance metrics, *i.e.*, the F1 scores, showcased a common trend across all models: a higher degree of accuracy when both the training and test sets originated from the same domain. However, we observed a decline as the contextual distance between training and testing data increased. This reflects the challenge of applying machine learning models to diverse educational content due to their varying subject matter and teaching methodology.

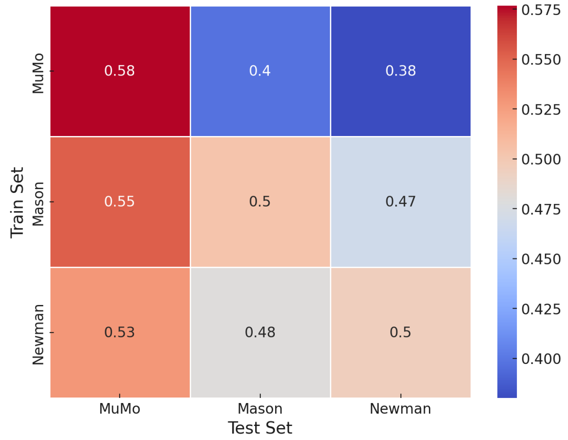


Figure 4: Cross-Dataset Performance Heatmap between MuMo, Mason, and Newman Datasets. The heatmap visualizes the performance (F1 scores) of models trained and tested across different datasets. The diagonal shows the results of models evaluated on the same dataset used for training, while the off-diagonal elements represent the transfer performance between different datasets. **red** indicates a higher average transfer learning performance.

A critical observation from our experiments is the distribution of classes within our datasets. Notably, several variables have only a very small number of positive examples in the data. The lack of sufficient support for the majority class in these datasets likely contributed to the lower F1 scores observed for those variables in many scenarios.

5.1 Error Analysis

In this section we investigate the discrepancies between the ground truth labels and the model predictions on the test set. We used the best performing model, *i.e.* BERT out of all the various experiments on a specific test set for this analysis. Table 7 shows paraphrased examples of interactions where the model failed to accurately predict the output label. We found that the model's predictions were mostly incorrect due to the lack of prior interactions as context. Classroom discourse is inherently continuous and time-series, meaning that understanding any given talk-turn often depends on the

Test set	Model	Questioning	Response Evaluation	Providing Information	Revoicing	Strategy Related	Behavior Management	Turn Management
MuMo	Mixtral	0.713	0.274	0.201	0.373	0.284	0.267	0.424
	Llama2	0.628	0.253	0.266	0.341	0.223	0.248	0.395
Mason	Mixtral	0.497	0.327	0.263	0.364	0.301	0.304	0.375
	Llama2	0.504	0.281	0.247	0.333	0.278	0.279	0.363
Newman	Mixtral	0.471	0.344	0.321	0.362	0.285	0.321	0.364
	Llama2	0.443	0.333	0.299	0.313	0.275	0.299	0.354

Table 4: Zero-shot results from MuMo, Mason, and Newman datasets across various categories.

Train set	Model	Questioning	Response Evaluation	Providing Information	Revoicing	Strategy Related	Behavior Management	Turn Management
Mu+Nw+Ms	BERT	0.547	0.346	0.322	0.323	0.281	0.304	0.344
	RoBERTa	0.539	0.323	0.334	0.312	0.267	0.319	0.331
	Mixtral	0.478	0.271	0.256	0.224	0.222	0.235	0.258
	Mixtral(c=1)	0.501	0.272	0.238	0.233	0.238	0.245	0.267
	Mixtral(c=3)	0.514	0.287	0.299	0.256	0.248	0.251	0.243
	Mixtral(c=5)	0.509	0.282	0.297	0.264	0.233	0.278	0.241
	Llama2	0.475	0.258	0.230	0.236	0.201	0.214	0.231
	Llama2(c=1)	0.491	0.268	0.242	0.241	0.215	0.219	0.242
	Llama2(c=3)	0.499	0.274	0.274	0.247	0.233	0.223	0.237
	Llama2(c=5)	0.504	0.285	0.263	0.251	0.238	0.246	0.233

Table 5: Transfer learning when the training set is MuMo, Newman, and Mason, but the test set is NCTE labeled with MuMo, Mason, Newman labels.

Train Data	Model	Student on Task	Teacher on Task	High Uptake	Focusing Question
NCTE	BERT	0.568	0.499	0.264	0.392
	RoBERTa	0.569	0.523	0.263	0.371
	Mixtral	0.439	0.411	0.198	0.236
	Mixtral(c=1)	0.444	0.435	0.206	0.264
	Mixtral(c=3)	0.469	0.448	0.231	0.273
	Mixtral(c=5)	0.470	0.442	0.256	0.251
	Llama2	0.421	0.376	0.163	0.202
	Llama2(c=1)	0.437	0.391	0.167	0.227
	Llama2(c=3)	0.472	0.427	0.199	0.239
	Llama2(c=5)	0.465	0.436	0.208	0.242

Table 6: Transfer learning when the train set is from NCTE and the test set is from MuMo, Mason, and Newman annotated with NCTE labels.

preceding turns, and the output labels rely heavily on the interactional context to be accurately classified. For example, in the table below under the MuMo test set, the model mislabels the talk turn “Rocks?" as Class 1 (Questioning) because the context was insufficient and the ground truth wasn’t Class 1 in this specific scenario. The students were supposed to simply repeat the utterance and the question mark doesn’t indicate a question being asked. Similarly, according to the codebook used for this paper, a compliment given by a teacher can be considered Providing Information, but the model struggled to label some of those interactions accurately and providing context could have eliminated that confusion.

Furthermore, in the Mason dataset, we noticed that the model failed to label the talk turn “I understand, I get that you are sleepy, but that also means you need to go to bed earlier" as Revoicing. In the interaction prior to the above talk turn, the student told the teacher that they were sleepy, and missing this context led to a misclassification. Another issue arose when the teacher’s instructions were mixed with reading sections of a book. For example, the teacher’s phrase “Can you sit down please?" accompanied by the teacher reading a small paragraph from the textbook in the same interaction led to prediction errors. This problem was observed in both the Mason and Newman datasets. Additionally, quite a few errors in the student’s talk turns were a result of short sentences and the lack of context provided to the models. These short, isolated statements were often misclassified because the model couldn’t access the surrounding interactions that would clarify their meaning. While the fine-tuned models gave the best performance, they are limited in their ability to incorporate the necessary context for precise predictions in certain scenarios. Therefore future work might explore the fine-tuning of models with large context windows to reap the benefits of both additional context and fine-tuning.

Test Set	Category	Speaker	Transcript	Actual Label	Pred Label
MuMo	Questioning	Student	Rocks?	0	1
	Turn Management	Teacher	Very good. And let me see what this is down here. A mallow.	1	0
	Questioning	Teacher	Rocks?	1	0
	Providing Information	Teacher	Very good. And a home for everyone.	1	0
Mason	Providing Information	Teacher	We're going to listen to that story now, talk to each other if we notice certain things. We can discuss more tomorrow as well.	1	0
	Revoicing	Teacher	I understand, I get that you are sleepy. But that also means you need to go to bed earlier.	1	0
	Literal Responses	Student	Then we can have two weddings	0	1
	Behavior Management	Teacher	Can you sit down please? [T reading: After the cake was served... We are doing the flower girl] They are pretending to be a flower girl while dancing.	1	0
	Questioning	Teacher	Can you sit down please? [T reading: After the cake was served... We are doing the flower girl] They are pretending to be a flower girl while dancing.	0	1
Newman	Literal Responses	Student	I was attached last winter, everybody hit me. James is their boss. I was very upset. I threw Quincy on accident.	0	1
	Questioning	Student	You know the paper? I mixed the two up. I was gonna write sad and the word said how they feel and then how they feel. How they feel	0	1
	Reading	Teacher	[reads from the book]. So was it okay for Jack to go to the library since there was no book to read from?	1	0
	Questioning	Student	Can I see? I cannot see the book.	0	1
NCTE	Focusing Question	Teacher	Oops, I bet, you know what? I made something, you know what happens?	0	1
	Focusing Question	Teacher	Where would you line up those Xs?	1	0
	Student on Task	Student	I am going to do a shout out.	0	1

Table 7: Error Analysis of Model Predictions Across Different Test Sets

6 Conclusion

Understanding classroom discourse is pivotal for improving educational outcomes and child development. In this study, we assess the generalizability of discourse codes across distinct educational domains of literacy and mathematics using automatic text classifiers such as transformer based models and in context learning based open weights generative models. We utilized several datasets from prior studies both from literacy and mathematics disciplines; annotated a subset of those data sets to generate ground truths for cross domain classification of educational classifiers. Our findings suggest show that transformer-based models, particularly BERT, and RoBERTa were better at classifying classroom discourse compared to open weights generative models. However, in-context models display resilience when tested on a new domain with limited training data.

In addition to these findings, we conducted error analysis using the best performing model, providing a fresh perspective on the model failures. We also experimented with providing context to the generative models in the form of prior interactions, and found out that such context could significantly impact the models' ability to understand and classify discourse accurately. The cross-domain experiments involving the Mason, Momo, and Newman datasets, labeled with the NCTE labels, achieved decent scores, except for the high uptake variable. This indicates a potential for these models to understand and classify discourse in educational settings to some extent. However, the experiments relating to NCTE data labeled for the literacy discourse codes did rather poorly, highlighting the difficulties in accurately capturing and generalizing discourse patterns within this domain. Given these challenges, we recommend future directions in this area of study to enhance the effectiveness of these models in the field of education. Enhancing the collection and annotation of classroom discourse data across a wider range of educational settings could improve the representation within training datasets. Implementing novel cross-domain techniques could help with better transfer learning and adaptation. Using better architectures and state-of-the-art (SOTA) models to help generalize the discourse codes across domains more effectively.

7 Implications and Future Work

While our research provides insights into the cross-domain generalizability of educational classifier models, future work could explore the application of these models across other educational domains such as social studies, science or other languages. Also, studying non-english classroom settings could help us understand unique challenges of applying these models. Or work focused on the higher level categories of discourse codes but examining the impact of fine grained labels on the efficacy of these models could be a worthwhile research direction. Our work considers the past few classroom interactions as context, but a potential next step wold be to embed information about the textbook being taught, teacher's instructional styles, child's background or prior academic performances etc. Future work could explore how the broadened contextual feature impact the generalizability or performance of these models.

The direction of this research could have potential for the development of educational tools for educators such as offering real time suggestions in classroom discussions, helps teachers to come up with better strategies to improve discourse codes such as questioning or behavior management etc. These models, especially the few shot and zero shot could reduce the need for extensive data collection and annotation, thus lower human labor. This could help economically backward regions and schools. The development of these educational classifier tools could also aid in the certain of ai driven agents, such as classroom robots that could actively participate in classroom discussions, demonstrate effective learning behaviors, support teachers as assistants.

Limitations

In case of generative models, we used only open weights models and local data processing strictly adhering to our data privacy and ethical standards. While this approach aligns with our ethical stance and ensures data confidentiality, it also narrows our selection of computational tools. Potentially more sophisticated and proprietary models with higher performance metrics were not considered in this study due to these constraints. The nature of our datasets presents another potential limitation. Some of the datasets utilized in our analysis (i.e., Mason, Neuman, and MuMo) were shared by the original authors of the work on the condition that we do not make them publicly available. This restriction could impose a barrier to the reproducibility of our study for future research. Our research primarily concentrates on specific subject domains like mathematics and English literacy. These subjects represent only a fraction of the diverse disciplines within the educational field, which our current paper does not account for.

Ethics Statement

The study utilized existing datasets derived from prior research that were shared with us by the authors of that work. In alignment with our commitment to confidentiality, we have anonymized all personal information. Names and other identifying details of students and teachers have been replaced with pseudonyms, thereby protecting their identities. Furthermore, the tools and models applied in our research such as Mixtral and Llama2 7B are open-weights generative models. The decision to use open-weights models supports transparency of our methods and further protects privacy by eliminating the need to transfer sensitive data to external servers. The use of open-weights models can also facilitate reproducibility in the research, allowing other researchers to validate and build upon the findings in our paper. The primary goal of the research was to investigate the efficacy of cross domain classification of educational discourse, particularly doctors within the classroom setting. We recognize the implications of applying AI in analyzing children's classroom interactions. It is important to approach the application of our research with the understanding of the potential impacts of AI application, making sure that it serves to enhance the educational experience rather than compromising it.

Acknowledgments

We would like to thank Evan Dallas, Souha Ben Hassine and Yosra Abdessamed for their assistance with data annotation, and Prof. Tanya Christ for providing access to the anonymized transcripts used for the analysis of the elementary classroom read-alouds in the literacy domain.

References

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Jere E Brophy and Thomas L. Good. 1984. Teacher behavior and student achievement microform jere brophy and thomas l. good.

Hyonsuk Cho and Tanya Christ. 2022. How two emergent bilingual students from refugee families make inferences with more and less culturally relevant texts during read-alouds. *TESOL Quarterly*, 56(4):1112–1135.

Tanya Christ, Iman Bakhoda, Ming Ming Chiu, X. Christine Wang, Alexa Schindel, and Yu Liu. 2023. Mediating learning in the zones of development: Role of teacher and kindergartner talk-turns during read-aloud discussions. *Journal of Research in Childhood Education*, 37(4):519–549.

Tanya Christ and Hyonsuk Cho. 2023. Emergent bilingual students' small group read-aloud discussions. *Literacy Research and Instruction*, 62(3):203–232.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

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

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.

Mary Faculty and Sarah Michaels. 1993. Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly*, 24:318 – 335.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What would a teacher do? predicting future talk moves.

Shannon Giroir, Leticia Grimaldo, Sharon Vaughn, and Greg Roberts. 2015. Interactive read-alouds for english learners in the elementary grades. *The Reading Teacher*, 68.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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. *arXiv preprint arXiv:1907.11692*.

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. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment.

Deliang Wang, Dapeng Shan, Yaqian Zheng, and Gaowei Chen. 2023. Teacher talk moves in k12 mathematics lessons: Automatic identification, prediction explanation, and characteristic exploration. In *Artificial Intelligence in Education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings*, page 651–664, Berlin, Heidelberg. Springer-Verlag.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Tanya Wright. 2019. Reading to learn from the start: The power of interactive read-alouds. *American Federation of Teachers*.

A Annotation Guidelines for the Classroom interaction dataset

These guidelines are designed to assist coders to accurately annotate classroom interactions between students and the teacher based on four specific labels from Demszky and Hill (2023). The output labels are: student on task, teacher on task, high uptake and focusing question.

The dataset consists of turn level utterances (paired annotations) between students and the teacher.

The table provided is set up to display a dialogue between students and their teacher, captured as turn-level utterances. Each row in the table represents a pair of exchanges, with the left column titled Student Transcript showing what the student said, and the right column titled Teacher Transcript presenting the teacher's response.

When annotating, it's important to note that the student's utterance comes first, followed by the teacher's response in the same row. So, this sequential flow indicates that the teacher's comment is a direct response to the student's immediately preceding utterance.

For example, if a student makes an observation or asks a question, the corresponding teacher's utterance in the same row will be a response or follow-up to that particular student's input.

A.1 Labels and Definitions

1. **Student on Task:** This label indicates whether a student's utterance is relevant to the current topic being discussed in the classroom.

2. **Teacher on Task:** This label reflects whether the teacher's utterance pertains to the topic of the current classroom session.

3. **High Uptake:** This label identifies instances where a speaker (teacher or student) builds upon what their interlocutor has said, demonstrating an understanding and extension of the conversation.

4. **Focusing Question:** This label is used when a teacher asks a question that prompts students to articulate, clarify, or reflect upon their own thoughts or those of their classmates.

A.2 Labeling Process

1. Student on Task

   (a) Label as 1 (On Task): If a student's utterance directly relates to the topic of the lecture or session. For example, discussing a specific math problem when the topic is math. Or if the classroom session is discussing the NLP textbook, then the topic would be NLP or anything related to it.

   (b) Label as 0 (Off Task): If a student's utterance is unrelated to the topic of the lecture. Such as talking about the weather or making a joke unrelated to the topic at hand.

2. Teacher on Task

   (a) Label as 1 (On Task): If the teacher's utterance is directly related to the subject matter of the current session similar to student on task label.

   (b) Label as 0 (Off Task): If the teacher's utterance is not related to the topic of the session.

3. High Uptake

   (a) Label as 1 (High Uptake): When a teacher acknowledges, repeats, or reformulates what the student has said, thereby extending the conversation.

   (b) Label as 0 (Low Uptake): When the response does not build upon the previous speaker's (student's) contribution.

4. Focusing Question

(a) Label as 1 (Focusing Question): If a teacher's question prompts the student to think deeply, articulate their understanding, or engage in reflection about their own thoughts or those of other students.

(b) Label as 0 (Funneling Question): If the teacher's question or teacher's set of questions to lead students to a desired procedure or conclusion, while giving limited attention to student responses that veer from the desired path

A.3 Examples for Each Category

1. Student on Task / Teacher on Task

   (a) Example (Label 1): Topic is English textbook "My friend Jamal". S: "It is because Jamal was a friend of joseph and they lived nearby." T: "yes! They were friends and what does that mean for joseph??"

   (b) Example (Label 0): Topic is English textbook "My friend Jamal. S: "I played soccer yesterday." T: "Shhh! Sit down quietly. We have 15 minutes left."

2. High Uptake

   (a) Example (Label 1): S: "Cause you took away 10 and 70 minus 10 is 60". T: "Why did we take away 10?".

   (b) Example (Label 1): S: "There's not enough seeds". T: "There's not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn't make sense?".

   (c) Example (Label 0): S: "Because the base of it is a hexagon". T: "Student K?".

3. Focusing Question

   (a) Example (Label 1): S: "I disagree with Student A because if you skip count by 100 ten times, that will get you to 1,000". T: "Let's try it. You ready? Let's start right here with Student F". S: "A hundred."

   (b) Example (Label 1): S: I first got 32 and then I got 48. T: And how did you find that? S: "Because I did 16 times two is 32".

   (c) Example (Label 0): S: "Do we eat pizza today". T: "Student K? What are you doing there???".

B Annotation Guidelines to label NCTE dataset

These guidelines are designed to assist annotators in labeling the classroom interactions between students and the teacher based on the categories defined in the research conducted by Christ, T., et al (2022). Note that this is a multi-label classification task and each individual interaction can have one or more possible output labels.

B.1 Labels and Definitions

1. **Response Evaluation:** When the teacher either compliments a child or expresses uncertainty about an incorrect response.

2. **Providing Information:** Extending or elaborating what was said either by the teacher or the student, building background knowledge, defining, using target words that are utilized in the context.

3. **Misinformation:** Either by providing misinformation or verifying an incorrect response.

4. **Revoicing:** When the teacher acknowledges and repeats what the student has said earlier.

5. **Strategy related:** Teacher directs a child to look at or think about text clues, or asks children to check their prediction.

6. **Questioning:** When a teacher questions a child to get a more detailed response, or elicit noticing text clues, or to define a target vocabulary etc.

7. **Behavior Management:** Gives children or a particular child a behavioral directive.

8. **Turn Management:** Teacher calls on particular child to respond or acknowledges or rejects a child's initiative to talk.

B.2 Labeling Process

To annotate this dataset:

(a) We read a transcript, and identify the possible codes that apply to that utterance using the codebook provided in the original paper.

(b) Look up the category that those particular codes fall under, and label either 1 or 0 on the spreadsheet. NOTE: The idea is to map the codes back to their categories and use them as labels instead.

(c) For example, when the teacher says “He is mowing the grass. Good!! What will the mower do to the flower if the dad gets closer? What would the mower do? Student K??”, the authors of the paper identified that the teacher repeated the student’s response. Then proceeded to compliment the child, acknowledging that the child has given the correct answer. Then proceeds to ask a question while directing that question to a particular child. Therefore we ended up with 4 possible codes for that one teacher utterance. Now we map those codes back to their categories.

C Model Hyperparameters

For the transformer-based deep learning models, we initialize each from the model checkpoint and fine-tune on our training data for 5 epochs with a batch size of 16, weight decay of 0.01, and a learning rate of 2e-5.

Details of the hyperparameter tuning for the generative models, Mixtral and Llama2:

1. **Do sample:** Set to false, this parameter ensures deterministic outputs by selecting tokens based on their probability distribution rather than introducing variability. This aligns with the experiment’s objective of restricting outputs to only "yes" and "no" tokens.

2. **Max new tokens:** With a value of 1, this parameter confines the model to generate exactly one token after the input prompt. Given our experiment’s focus on producing either "yes" or "no," a value of one facilitates the desired output format of one token per response.

3. **Temperature:** Set to 0, indicating no randomness in output selection. By eliminating randomness, the model consistently chooses the same sequence of tokens from the input prompt, thereby ensuring deterministic output.

4. **Top k:** Set to 2, this parameter limits consideration to the top two tokens with the highest probabilities. Since the objective of this experiment is binary output ("yes" or "no"), setting Top k to 2 effectively restricts the model’s outputs to these two options.

5. **Num_return_sequence:** set to 1

D Krippendorf’s Alpha

Label	Alpha	Source Dataset	Target Dataset
Response Evaluation	0.849	MMN	NCTE
Providing Information	0.958		
Revoicing	0.899		
Strategy-related	0.818		
Questioning	0.909		
Behavior Management	0.801		
Turn Management	0.936		
Misinformation	N/A		
Student on Task	0.912	NCTE	MMN
Teacher on Task	0.943		
High Uptake	0.847		
Focusing Question	0.851		

Table 8: Krippendorff’s α intercoder agreement scores for the combined datasets of MuMo, Mason, and Newman (MMN) using labels created for the NCTE dataset and vice versa. The label “Misinformation” was never assigned to any text.

E Dataset Statistics

In order to evaluate the generalization performance of the models, we annotated new data as described in subsection 3.2. The number of datapoints assigned each label are presented in Tables 9 and 10.

Variable	Class 0	Class 1
Student on Task	21	119
Teacher on Task	13	127
High Uptake	74	66
Focusing Question	83	57

Table 9: Class distribution of MuMo/Mason/Newman data annotated with NCTE labels. Class 0 indicates the label does not apply and Class 1 indicates that it does.

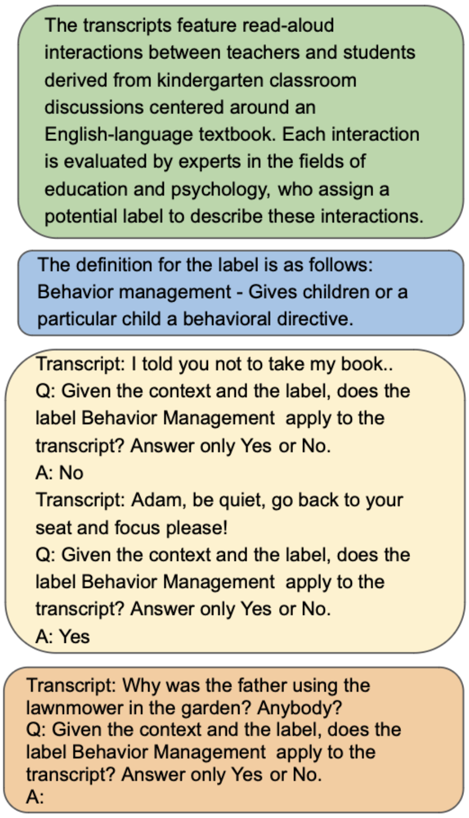
F Prompt Components for Generative Models

G Experimental Details

H Mapping the MuMo Codebook to Mason and Newman Datasets

Variable	Class 0	Class 1
Response Evaluation	55	45
Providing Information	58	42
Revoicing	74	26
Strategy Related	69	31
Questioning	32	68
Behavior Management	90	10
Turn Management	70	30

Table 10: Class distribution of NCTE data when annotated with MuMo/Mason/Newman label set. Class 0 indicates the label does not apply and Class 1 indicates that it does.



The transcripts feature read-aloud interactions between teachers and students derived from kindergarten classroom discussions centered around an English-language textbook. Each interaction is evaluated by experts in the fields of education and psychology, who assign a potential label to describe these interactions.

The definition for the label is as follows:
Behavior management - Gives children or a particular child a behavioral directive.

Transcript: I told you not to take my book..
Q: Given the context and the label, does the label Behavior Management apply to the transcript? Answer only Yes or No.
A: No
Transcript: Adam, be quiet, go back to your seat and focus please!
Q: Given the context and the label, does the label Behavior Management apply to the transcript? Answer only Yes or No.
A: Yes

Transcript: Why was the father using the lawnmower in the garden? Anybody?
Q: Given the context and the label, does the label Behavior Management apply to the transcript? Answer only Yes or No.
A:

Figure 5: Prompt components for generative models for the read-aloud setting (used for MuMo/Mason/Newman). From top to bottom, the blocks display the background information (green), labels (blue), few-shot examples (yellow).

Train Set	**Models**	**Questioning**	**Response Evaluation**	**Providing Information**	**Revoicing**	**Strategy Related**	**Behavior Management**	**Turn Management**
MuMo	*Baseline*	*0.241*	*0.414*	*0.151*	*0.178*	*0.230*	*0.194*	*0.198*
	BERT	**0.641**	**0.331**	**0.320**	**0.444**	0.233	**0.323**	**0.269**
	RoBERTa	0.612	0.266	0.319	0.415	**0.279**	0.303	0.253
	Mixtral	0.450	0.285	0.294	0.371	0.320	0.307	0.250
	Mixtral (c=1)	0.462	0.270	0.299	0.364	0.303	0.310	0.248
	Mixtral (c=3)	0.525	0.299	0.316	0.386	0.288	0.313	0.252
	Mixtral (c=5)	0.537	0.283	0.333	0.405	0.284	0.322	0.255
	Llama2	0.438	0.276	0.239	0.310	0.244	0.279	0.259
	Llama2 (c=1)	0.445	0.240	0.240	0.313	0.241	0.290	0.250
	Llama2 (c=3)	0.484	0.279	0.282	0.333	0.274	0.300	0.251
	Llama2 (c=5)	0.487	0.270	0.311	0.347	0.266	0.304	0.251
Mason	BERT	**0.699**	**0.460**	**0.324**	**0.538**	**0.333**	**0.460**	**0.458**
	RoBERTa	0.699	0.422	0.315	0.530	0.329	0.459	0.457
	Mixtral	0.535	0.340	0.330	0.397	0.332	0.350	0.420
	Mixtral (c=1)	0.601	0.360	0.315	0.455	0.327	0.365	0.421
	Mixtral (c=3)	0.637	0.407	0.315	0.473	0.331	0.384	0.443
	Mixtral (c=5)	0.644	0.436	0.318	0.470	0.333	0.401	0.444
	Llama2	0.488	0.281	0.284	0.395	0.275	0.334	0.351
	Llama2 (c=1)	0.503	0.312	0.281	0.423	0.292	0.365	0.369
	Llama2 (c=3)	0.585	0.303	0.293	0.488	0.307	0.372	0.397
	Llama2 (c=5)	0.599	0.326	0.305	**0.544**	0.330	0.351	0.405
Newman	BERT	**0.710**	0.489	0.339	**0.535**	**0.396**	**0.473**	**0.510**
	RoBERTa	0.702	0.460	0.325	0.530	0.383	0.472	0.497
	Mixtral	0.472	0.384	0.349	0.420	0.313	0.370	0.400
	Mixtral (c=1)	0.510	0.412	**0.352**	0.429	0.327	0.384	0.428
	Mixtral (c=3)	0.666	0.440	0.343	0.481	0.367	0.386	0.459
	Mixtral (c=5)	0.669	**0.490**	0.336	0.493	0.365	0.368	0.450
	Llama2	0.444	0.340	0.287	0.359	0.304	0.344	0.401
	Llama2 (c=1)	0.592	0.365	0.295	0.382	0.324	0.359	0.403
	Llama2 (c=3)	0.594	0.425	0.316	0.436	0.315	0.387	0.418
	Llama2 (c=5)	0.571	0.484	0.344	0.443	0.339	0.368	0.435
Ms+Nw	BERT	**0.689**	**0.479**	**0.353**	**0.530**	**0.388**	0.462	**0.495**
	RoBERTa	0.676	0.472	0.353	0.492	0.364	**0.444**	0.453
	Mixtral	0.478	0.367	0.311	0.400	0.345	0.369	0.381
	Mixtral (c=1)	0.614	0.384	0.325	0.427	0.353	0.392	0.409
	Mixtral (c=3)	0.617	0.427	0.335	0.438	0.370	0.429	0.403
	Mixtral (c=5)	0.582	0.401	0.352	0.427	0.374	0.457	0.419
	Llama2	0.413	0.352	0.279	0.333	0.326	0.388	0.325
	Llama2 (c=1)	0.497	0.336	0.291	0.369	0.344	0.402	0.334
	Llama2 (c=3)	0.500	0.361	0.330	0.437	0.360	0.440	0.353
	Llama2 (c=5)	0.490	0.379	0.339	0.428	0.341	**0.460**	0.333
Mu+Ms	BERT	**0.666**	**0.457**	**0.350**	**0.466**	**0.378**	**0.446**	0.494
	RoBERTa	0.641	0.446	0.327	0.428	0.354	0.429	**0.507**
	Mixtral	0.430	0.360	0.241	0.279	0.292	0.314	0.377
	Mixtral (c=1)	0.473	0.384	0.252	0.315	0.315	0.317	0.401
	Mixtral (c=3)	0.577	0.412	0.304	0.379	0.345	0.344	0.402
	Mixtral (c=5)	0.654	0.449	0.300	0.353	0.373	0.354	0.396
	Llama2	0.414	0.345	0.248	0.245	0.250	0.279	0.383
	Llama2 (c=1)	0.479	0.349	0.270	0.286	0.271	0.307	0.401
	Llama2 (c=3)	0.567	0.404	0.308	0.379	0.327	0.336	0.444
	Llama2 (c=5)	0.562	0.392	0.299	0.369	0.375	0.343	0.419
Mu+Nw	BERT	0.691	0.463	0.347	**0.419**	0.396	0.430	**0.500**
	RoBERTa	0.676	0.460	0.324	0.412	0.376	**0.423**	0.471
	Mixtral	0.468	0.301	0.350	0.344	0.319	0.355	0.306
	Mixtral (c=1)	0.504	0.333	**0.356**	0.366	0.329	0.366	0.342
	Mixtral (c=3)	0.612	0.404	0.353	0.387	0.360	0.405	0.346
	Mixtral (c=5)	**0.694**	**0.472**	0.349	0.411	**0.401**	0.427	0.343
	Llama2	0.384	0.300	0.275	0.286	0.325	0.287	0.295
	Llama2 (c=1)	0.453	0.328	0.285	0.308	0.340	0.313	0.337
	Llama2 (c=3)	0.573	0.393	0.320	0.362	0.367	0.379	0.324
	Llama2 (c=5)	0.680	0.462	0.344	0.411	0.400	**0.432**	0.310

Table 11: F1-score when training on various train sets and evaluating on the test set from *Newman*. **Bold** indicates the best score for each column for each training set, underline indicates the best overall score for each column.

Train Set	Model	Questioning	Response Evaluation	Providing Information	Revoicing	Strategy Related	Behavior Management	Turn Management
NCTE	*Baseline*	*0.706*	*0.308*	*0.625*	*0.625*	*0.308*	*0.533*	*0.308*
	BERT	**0.821**	0.480	0.525	0.000	0.333	0.000	0.800
	RoBERTa	0.7724	**0.666**	**0.649**	0.000	**0.495**	0.000	**0.813**
	Mixtral	0.692	0.389	0.363	**0.241**	0.337	0.000	0.525
	Llama2	0.614	0.321	0.381	0.219	0.238	**0.190**	0.316
	Mixtral c=1	0.686	0.359	0.370	0.258	0.322	0.200	0.569
	Mixtral c=3	0.721	0.378	0.365	0.255	0.317	0.214	0.555
	Mixtral c=5	0.714	0.377	0.333	0.263	0.309	0.179	0.565
	Llama c=1	0.604	0.338	0.383	0.222	0.281	0.222	0.407
	Llama c=3	0.628	0.334	0.385	0.246	0.325	0.222	0.411
	Llama c=5	0.624	0.313	0.342	0.210	0.287	0.213	0.405

Table 12: Generalization performance on NCTE data labeled with MuMo, Mason and Newman dataset labels.

Train Set	Model	Student on Task	Teacher on Task	High Uptake	Focusing Question
Mu+Ms+Nw	*Baseline*	*1.000*	*1.000*	*0.929*	*0.636*
	BERT	**0.962**	**0.800**	0.333	**0.694**
	RoBERTa	0.941	0.785	**0.369**	0.656
	Mixtral	0.784	0.601	0.303	0.666
	Mixtral (c=1)	0.740	0.600	0.300	0.661
	Mixtral (c=3)	0.767	0.637	0.297	0.628
	Mixtral (c=5)	0.749	0.615	0.270	0.612
	Llama2	0.678	0.610	0.263	0.537
	Mixtral (c=1)	0.684	0.580	0.284	0.550
	Mixtral (c=3)	0.647	0.613	0.289	0.523
	Mixtral (c=5)	0.666	0.600	0.251	0.501

Table 13: Generalization performance on subset of MuMo Data labeled using NCTE's labels.

Train Set	Models	Questioning	Response Evaluation	Providing Information	Revoicing	Strategy Related	Behavior Management	Turn Management
MuMo	*Baseline*	*0.470*	*0.230*	*0.200*	*0.364*	*0.105*	*0.036*	*0.364*
	BERT	**0.904**	**0.375**	**0.564**	**0.522**	**0.362**	**0.533**	**0.737**
	RoBERTa	0.871	0.344	0.555	0.501	0.370	0.555	0.747
	Mixtral	0.790	0.318	0.231	0.444	0.333	0.378	0.478
	Mixtral (c=1)	0.881	0.320	0.324	0.451	0.333	0.415	0.538
	Mixtral (c=3)	0.888	0.337	0.362	0.484	0.345	0.467	0.594
	Mixtral (c=5)	0.865	0.330	0.363	0.476	0.358	0.525	0.575
	Llama2	0.643	0.275	0.286	0.310	0.219	0.344	0.404
	Llama2 (c=1)	0.726	0.294	0.323	0.371	0.252	0.387	0.480
	Llama2 (c=3)	0.810	0.297	0.404	0.375	0.303	0.465	0.511
	Llama2 (c=5)	0.807	0.280	0.417	0.380	0.306	0.541	0.518
Mason	BERT	**0.722**	**0.345**	**0.542**	**0.500**	**0.500**	**0.557**	**0.688**
	RoBERTa	0.718	0.333	0.514	0.500	0.500	0.548	0.680
	Mixtral	0.523	0.327	0.289	0.346	0.334	0.694	0.675
	Mixtral (c=1)	0.657	0.335	0.439	0.451	0.402	0.690	0.689
	Mixtral (c=3)	0.685	0.335	0.447	0.447	0.430	0.734	0.680
	Mixtral (c=5)	0.636	0.347	0.452	0.492	0.439	0.697	**0.701**
	Llama2	0.551	0.287	0.282	0.322	0.321	0.474	0.595
	Llama2 (c=1)	0.580	0.294	0.353	0.381	0.354	0.493	0.614
	Llama2 (c=3)	0.651	0.323	0.434	0.422	0.420	0.533	0.652
	Llama2 (c=5)	0.658	0.348	0.436	0.417	0.419	0.521	0.653
Newman	BERT	0.741	**0.460**	0.461	0.470	0.330	0.500	**0.595**
	RoBERTa	**0.742**	0.440	**0.562**	**0.476**	**0.388**	**0.528**	0.555
	Mixtral	0.464	0.349	0.387	0.341	0.307	0.444	0.463
	Mixtral (c=1)	0.696	0.376	0.405	0.364	0.309	0.454	0.487
	Mixtral (c=3)	0.695	0.409	0.430	0.426	0.323	0.485	0.545
	Mixtral (c=5)	0.735	0.461	0.452	0.466	0.329	0.502	0.599
	Llama2	0.400	0.350	0.378	0.330	0.240	0.295	0.301
	Llama2 (c=1)	0.568	0.371	0.390	0.358	0.262	0.340	0.355
	Llama2 (c=3)	0.655	0.416	0.430	0.411	0.298	0.422	0.481
	Llama2 (c=5)	0.663	0.454	0.452	0.465	0.330	0.507	0.585
Ms+Nw	BERT	**0.620**	**0.333**	0.499	**0.458**	**0.333**	0.430	0.467
	RoBERTa	0.611	0.333	**0.504**	0.443	0.327	**0.442**	**0.476**
	Mixtral	0.525	0.298	0.395	0.354	0.311	0.380	0.294
	Mixtral (c=1)	0.542	0.307	0.411	0.378	0.315	0.386	0.331
	Mixtral (c=3)	0.570	0.322	0.462	0.410	0.323	0.405	0.405
	Mixtral (c=5)	0.620	0.327	0.494	0.464	0.332	0.435	0.470
	Llama2	0.485	0.290	0.371	0.350	0.279	0.320	0.344
	Llama2 (c=1)	0.522	0.300	0.404	0.377	0.291	0.345	0.369
	Llama2 (c=3)	0.573	0.316	0.455	0.419	0.311	0.382	0.414
	Llama2 (c=5)	0.623	0.337	0.499	0.460	0.330	0.423	0.463
Mu+Ms	BERT	**0.840**	**0.369**	**0.542**	**0.492**	**0.351**	**0.511**	0.651
	RoBERTa	0.822	0.338	0.530	0.464	0.351	0.500	**0.666**
	Mixtral	0.622	0.295	0.389	0.400	0.317	0.381	0.471
	Mixtral (c=1)	0.776	0.308	0.420	0.412	0.320	0.450	0.571
	Mixtral (c=3)	0.750	0.378	0.422	0.474	0.333	0.458	0.588
	Mixtral (c=5)	0.743	0.342	0.450	0.457	0.309	0.501	0.547
	Llama2	0.595	0.244	0.352	0.373	0.308	0.331	0.386
	Llama2 (c=1)	0.651	0.273	0.397	0.396	0.330	0.408	0.441
	Llama2 (c=3)	0.734	0.295	0.465	0.436	0.332	0.442	0.443
	Llama2 (c=5)	0.729	0.269	0.434	0.414	0.341	0.469	0.413
Mu+Nw	BERT	0.832	**0.364**	**0.518**	**0.490**	**0.323**	**0.509**	**0.668**
	RoBERTa	**0.833**	0.349	0.516	0.489	0.330	0.510	0.640
	Mixtral	0.643	0.313	0.334	0.387	0.307	0.363	0.444
	Mixtral (c=1)	0.858	0.325	0.422	0.447	0.315	0.430	0.563
	Mixtral (c=3)	0.801	0.347	0.442	0.493	0.321	0.515	0.609
	Mixtral (c=5)	0.786	0.345	0.428	0.487	0.325	0.486	0.622
	Llama2	0.594	0.238	0.248	0.278	0.231	0.321	0.367
	Llama2 (c=1)	0.748	0.268	0.298	0.374	0.246	0.353	0.430
	Llama2 (c=3)	0.739	0.360	0.413	0.410	0.282	0.436	0.473
	Llama2 (c=5)	0.819	0.326	0.408	0.402	0.317	0.405	0.573

Table 14: F1-score when training on various train sets and evaluating on the test set from *MuMo*. **Bold** indicates the best score for each column for each training set, underline indicates the best overall score for each column.

Train Set	Models	Questioning	Response Evaluation	Providing Information	Revoicing	Strategy Related	Behavior Management	Turn Management
MuMo	*Baseline*	*0.420*	*0.120*	*0.033*	*0.275*	*0.160*	*0.065*	*0.128*
	BERT	**0.644**	**0.305**	**0.327**	**0.458**	0.255	**0.341**	**0.294**
	RoBERTa	0.638	0.284	0.306	0.443	**0.306**	0.322	0.273
	Mixtral	0.472	0.295	0.277	0.380	0.303	0.288	0.269
	Mixtral (c=1)	0.556	0.303	0.289	0.389	0.296	0.300	0.357
	Mixtral (c=3)	0.581	0.344	0.308	0.420	0.280	0.323	0.370
	Mixtral (c=5)	0.552	0.303	0.327	0.455	0.259	0.345	0.303
	Llama2	0.444	0.300	0.264	0.339	0.263	0.306	0.279
	Llama2 (c=1)	0.493	0.303	0.280	0.366	0.261	0.316	0.310
	Llama2 (c=3)	0.477	0.327	0.302	0.450	0.258	0.330	0.324
	Llama2 (c=5)	0.453	0.309	0.322	0.466	0.252	0.341	0.295
Mason	BERT	0.729	**0.462**	0.334	**0.610**	**0.365**	**0.480**	**0.501**
	RoBERTa	0.700	0.445	**0.337**	0.608	0.325	0.434	0.478
	Mixtral	0.560	0.380	0.309	0.399	0.316	0.367	0.400
	Mixtral (c=1)	0.704	0.389	0.318	0.436	0.322	0.394	0.418
	Mixtral (c=3)	**0.748**	0.433	0.321	0.515	0.352	0.428	0.465
	Mixtral (c=5)	0.727	0.467	0.337	0.614	0.365	0.486	0.493
	Llama2	0.562	0.291	0.275	0.401	0.269	0.318	0.383
	Llama2 (c=1)	0.689	0.328	0.285	0.444	0.288	0.351	0.405
	Llama2 (c=3)	0.669	0.391	0.313	0.522	0.323	0.409	0.448
	Llama2 (c=5)	0.718	0.463	0.338	0.602	0.361	0.480	**0.506**
Newman	BERT	**0.711**	**0.447**	**0.339**	**0.541**	**0.328**	0.449	0.495
	RoBERTa	0.693	0.444	0.309	0.517	0.326	**0.461**	**0.512**
	Mixtral	0.500	0.422	0.300	0.381	0.279	0.344	0.425
	Mixtral (c=1)	0.643	0.427	0.305	0.406	0.285	0.361	0.433
	Mixtral (c=3)	0.714	0.438	0.320	0.486	0.309	0.410	0.460
	Mixtral (c=5)	0.721	0.447	0.344	0.543	0.328	0.448	0.489
	Llama2	0.469	0.375	0.238	0.366	0.278	0.343	0.366
	Llama2 (c=1)	0.608	0.396	0.258	0.406	0.287	0.358	0.394
	Llama2 (c=3)	0.633	0.412	0.299	0.467	0.314	0.400	0.448
	Llama2 (c=5)	0.614	0.452	0.335	0.541	0.327	0.448	0.502
Ms+Nw	BERT	**0.716**	**0.434**	**0.329**	**0.563**	**0.348**	0.444	0.478
	RoBERTa	0.707	0.432	0.313	0.520	0.341	**0.400**	**0.463**
	Mixtral	0.475	0.353	0.279	0.342	0.268	0.332	0.384
	Mixtral (c=1)	0.633	0.366	0.290	0.388	0.287	0.361	0.438
	Mixtral (c=3)	0.692	0.404	0.305	0.468	0.311	0.391	0.440
	Mixtral (c=5)	0.719	0.400	0.325	0.568	0.351	0.443	0.482
	Llama2	0.527	0.301	0.233	0.344	0.287	0.300	0.348
	Llama2 (c=1)	0.654	0.341	0.253	0.390	0.304	0.328	0.375
	Llama2 (c=3)	0.718	0.337	0.294	0.467	0.319	0.383	0.430
	Llama2 (c=5)	0.680	0.326	0.324	0.567	0.351	0.452	0.448
Mu+Ms	BERT	**0.703**	**0.440**	**0.323**	**0.587**	**0.349**	**0.444**	**0.455**
	RoBERTa	0.694	0.438	0.319	0.560	0.347	0.429	0.454
	Mixtral	0.512	0.373	0.304	0.331	0.296	0.259	0.382
	Mixtral (c=1)	0.650	0.380	0.302	0.481	0.304	0.293	0.390
	Mixtral (c=3)	0.698	0.419	0.316	0.493	0.334	0.273	0.422
	Mixtral (c=5)	0.630	0.435	0.323	0.501	0.355	0.338	0.407
	Llama2	0.499	0.334	0.292	0.306	0.297	0.251	0.340
	Llama2 (c=1)	0.593	0.355	0.293	0.427	0.305	0.292	0.368
	Llama2 (c=3)	0.617	0.401	0.308	0.453	0.332	0.368	0.408
	Llama2 (c=5)	0.628	0.437	**0.325**	0.445	0.349	0.353	0.388
Mu+Nw	BERT	**0.683**	**0.436**	**0.310**	0.540	**0.316**	0.405	**0.444**
	RoBERTa	0.665	0.428	0.307	0.527	0.300	**0.421**	0.421
	Mixtral	0.489	0.366	0.264	0.460	0.247	0.301	0.399
	Mixtral (c=1)	0.569	0.379	0.270	0.473	0.263	0.323	0.412
	Mixtral (c=3)	0.598	0.409	0.297	**0.547**	0.268	0.380	0.434
	Mixtral (c=5)	0.616	0.401	0.267	0.521	0.261	0.354	0.443
	Llama2	0.420	0.341	0.251	0.414	0.233	0.258	0.384
	Llama2 (c=1)	0.540	0.359	0.259	0.445	0.250	0.290	0.396
	Llama2 (c=3)	0.573	0.403	0.292	0.485	0.283	0.345	0.427
	Llama2 (c=5)	0.543	0.403	0.305	0.535	0.249	0.400	0.412

Table 15: F1-score when training on various train sets and evaluating on the test set from *Mason*. **Bold** indicates the best score for each column for each training set, underline indicates the best overall score for each column.

Train Set	BERT	RoBERTa	Mixtral				Llama2			
			c=0	c=1	c=3	c=5	c=0	c=1	c=3	c=5
MuMo	**0.375**	0.367	0.326	0.355	0.375	0.363	0.313	0.332	0.351	0.348
Mason	0.497	0.475	0.390	0.426	0.466	**0.499**	0.357	0.398	0.439	0.495
Newman	0.473	0.466	0.379	0.409	0.448	**0.474**	0.348	0.387	0.425	0.460
Ms+Nw	**0.473**	0.453	0.348	0.395	0.430	0.470	0.334	0.378	0.421	0.450
Mu+Ms	**0.471**	0.463	0.351	0.400	0.422	0.427	0.331	0.376	0.413	0.418
Mu+Nw	**0.448**	0.438	0.361	0.384	0.419	0.409	0.329	0.362	0.401	0.407

Table 16: Average F1-score for each model across different training sets for test set Mason. **c** denotes the number of prior interactions provided as context to the generative models during classification.

Train Set	BERT	RoBERTa	Mixtral				Llama2			
			c=0	c=1	c=3	c=5	c=0	c=1	c=3	c=5
MuMo	**0.350**	0.336	0.325	0.321	0.341	0.345	0.306	0.306	0.326	0.316
Mason	**0.527**	0.471	0.384	0.407	0.423	0.435	0.337	0.364	0.392	0.400
Newman	**0.494**	0.462	0.387	0.407	0.439	0.447	0.341	0.373	0.399	0.415
Ms+Nw	**0.486**	0.468	0.379	0.401	0.420	0.421	0.337	0.368	0.392	0.386
Mu+Ms	**0.464**	0.453	0.342	0.361	0.399	0.395	0.320	0.351	0.378	0.371
Mu+Nw	**0.469**	0.431	0.343	0.357	0.388	0.407	0.318	0.338	0.388	0.401

Table 17: Average F1-score for each model across different training sets for test set Newman. **c** denotes the number of prior interactions provided as context to the generative models during classification.

Train Set	BERT	RoBERTa	Mixtral				Llama2			
			c=0	c=1	c=3	c=5	c=0	c=1	c=3	c=5
MuMo	0.542	0.517	0.419	0.480	0.485	0.484	0.348	0.420	0.451	0.466
Mason	0.541	0.536	0.455	0.519	0.532	**0.570**	0.404	0.463	0.516	0.523
Newman	0.515	**0.529**	0.393	0.455	0.470	0.497	0.342	0.407	0.448	0.471
Ms+Nw	**0.488**	0.488	0.335	0.367	0.396	0.436	0.306	0.348	0.390	0.406
Mu+Ms	**0.601**	0.582	0.405	0.448	0.479	0.473	0.352	0.413	0.447	0.445
Mu+Nw	**0.577**	0.561	0.406	0.475	0.518	0.506	0.338	0.413	0.444	0.454

Table 18: Average F1-score for each model across different training sets for test set MuMo. **c** denotes the number of prior interactions provided as context to the generative models during classification.

Train Set	BERT	RoBERTa	Mixtral				Llama2			
			c=0	c=1	c=3	c=5	c=0	c=1	c=3	c=5
NCTE	0.437	**0.488**	0.358	0.395	0.401	0.382	0.326	0.351	0.361	0.339

Table 19: Average F1-score for each model across different training sets for NCTE data. **c** denotes the number of prior interactions provided as context to the generative models during classification.

MuMo	**Mason/Newman**
Questioning	Questioning
Providing Information	Providing correct definitions/examples
Response Evaluation	Text based responses
Revoicing	Restates
Strategy Related	Interconnected thinking
Behavior Management	Cultural norms and expectations for behavior
Turn Management	Identification of expected behaviors/invitation to participate

Table 20: Adoption of MuMo codebook for Mason and Newman Datasets. The high-level categories from MuMo were used as a standardized framework to maintain a consistent higher level labels and facilitate cross-domain analysis