Enhancing Talk Moves Analysis in Mathematics Tutoring through Classroom Teaching Discourse

Jie Cao[1,2]*, Abhijit Suresh[2], Jennifer Jacobs[2], Charis Clevenger[2], Amanda Howard[2], Chelsea Brown[2], Brent Milne[3], Tom Fischaber[3], Tamara Sumner[2], James H. Martin[2]

[1]School of Computer Science, University of Oklahoma,
[2]Institute of Cognitive Science, University of Colorado Boulder,
[3]Saga Education

{jie.cao}@ou.edu, {firstname.lastname}@colorado.edu, {bmilne, tfischaber}@saga.org

Abstract

Human tutoring interventions play a crucial role in supporting student learning, improving academic performance, and promoting personal growth. This paper focuses on analyzing mathematics tutoring discourse using talk moves—a framework of dialogue acts grounded in Accountable Talk theory. However, scaling the collection, annotation, and analysis of extensive tutoring dialogues to develop machine learning models is a challenging and resource-intensive task. To address this, we present SAGA22, a compact dataset, and explore various modeling strategies, including dialogue context, speaker information, pretraining datasets, and further fine-tuning. By leveraging existing datasets and models designed for classroom teaching, our results demonstrate that supplementary pretraining on classroom data enhances model performance in tutoring settings, particularly when incorporating longer context and speaker information. Additionally, we conduct extensive ablation studies to underscore the challenges in talk move modeling.

1 Introduction

Human tutoring has become an essential component in combating learning loss due to the COVID-19 pandemic (Robinson and Loeb, 2021; Zhou et al., 2021; Engzell et al., 2021; Lewis et al., 2021; Patarapichayatham et al., 2021). In addition, expanding the tutoring workforce is critical to addressing teacher shortages. However, novice tutors lack adequate training in both their content area and in current pedagogical approaches and thus require extensive professional development.

Most methods for offering feedback to teachers rely on skilled human observers (Correnti et al., 2015; Wolf et al., 2005), making them costly, time-intensive, and generally inaccessible to paraprofessional tutors. However, recent research has demonstrated automated techniques to reliably detect educationally important discursive features such as productive dialogue, instructional talk, authentic questions, elaborated evaluation, and uptake (Kelly et al., 2018; Suresh et al., 2018; Song et al., 2020; Demszky et al., 2021; Jensen et al., 2020).

Much of this earlier work focuses on traditional classroom settings, not small group tutoring. Here, we address the question of whether models initially created for the classroom can serve as the basis for new models for the tutoring setting. This work mainly focus on the creation of discourse analysis tool for mathematics tutoring. Specifically, we focus on **talk moves** – a set of dialogue acts based on Accountable Talk theory (O'Connor et al., 2015; Resnick et al., 2018; Michaels and O'Connor, 2015), including both teacher and student talk moves (see §3.1). Research has shown that appropriate use of talk moves in the classroom promotes student learning (Resnick et al., 2010; Walshaw and Anthony, 2008; Webb et al., 2019), and ensure that all students have equal access to participation, subject matter content, and developing appropriate habits of mind (Michaels et al., 2008; O'Connor and Michaels, 2019).

To address the mismatch between the classroom and tutoring settings, we developed a new mathematics tutoring dataset with talk move annotations on 121 tutoring sessions. We then examined existing modeling strategies and datasets for classroom mathematics teaching, and explore the best *transfer learning strategies* for our target domain. Our modeling experiments and analyses demonstrate how best to use a supervised pretraining-finetuning framework on tutoring talk move analysis, including dialogue context, speaker information, and training strategies. Our best new models outperform existing baselines by a large margin in the tutoring domain and approach the performance of existing models for the classroom domain. Finally, detailed analyses highlight the challenges

*This work was partially done when Jie Cao was a postdoctoral researcher at the University of Colorado Boulder.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7671–7684
January 19–24, 2025.

and point to future work on discourse modeling for mathematics tutoring.

In short, we (1) introduce a new dataset of talk moves annotated math tutoring sessions, (2) describe talk move models for math tutoring with a thorough comparsion with existing models and datasets, (3) highlight the challenges and future work by extensive ablation studies.

2 Related Work

Our contributions on new tutoring datasets and transfer learning from existing models build on two lines of research: existing classroom datasets and talk move models for mathematics education.

2.1 Dialogue Datasets on Mathematics Education

Most publicly available datasets are based on mathematics classroom instruction. The **TalkMoves** (Suresh et al., 2022a) and **NCTE** (Demszky and Hill, 2022) datasets are annotated dialogue corpora collected from real-world classrooms. TalkMoves is derived from three collections of transcripts: Inside Mathematics [1];the Third International Mathematics and Science Study (TIMSS) 1999 video study [2]; Video Mosaic [3]. National Center for Teacher Effectiveness (NCTE) [4] conducted a systematic collection of recorded mathematics classroom observations, from 2010 to 2013, over 300 classrooms were filmed, resulting in 1,660 lessons for elementary math. Both the TalkMoves and NCTE datasets have been used extensively to create models of classroom discourse.

Creating a tutoring dataset as the size and quality of earlier TalkMoves and NCTE efforts is time-consuming and expensive. Limited resources are available for authentic, high quality, tutoring sessions. CIMA (Stasaski et al., 2020), TSCC (Caines et al., 2022) are one-to-one corpora for tutoring for language learning, either through crowdsourced role-playing or online private chatroom. MathDial (Macina et al., 2023) collect one-to-one dialogue between an expert annotator as teacher and an LLM that simulates the student. Our study addresses this need by providing a small real-world math tutoring dataset annotated in a manner that is consistent with existing resources.[5]

[1] https://www.insidemathematics.org
[2] http://www.timssvideo.com
[3] https://videomosaic.org
[4] https://cepr.harvard.edu/ncte
[5] Please contact the first author for the code and datasets.

2.2 Automatic Talk Moves Analysis

Suresh et al. (2022a) report on a set of pre-trained transformer-based (Vaswani et al., 2017) models to provide automatic, personalized feedback on the use of this limited set of talk moves. They fine-tuned BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), Electra (Clark et al., 2020) in a classification setting (with one previous utterances as context, and the target utterance, without speaker information). In later work, Suresh et al. (2022b) employed longer contexts by concatenating previous utterances, target utterance, and subsequent utterances into a single input sequence with ordered utterances. A longitudinal pilot study points to the utility value of this tool for teachers, including its positive impact on their discourse practices over time (Jacobs et al., 2022; Scornavacco et al., 2021).

Recently, Cao et al. (2023) extended talk moves modeling to collaborative learning setting, focusing on how the noisy speech in real-world small-group classroom impacts the student talk move modeling. By providing a description and an example utterance for each talk move type, Wang et al. (2023) introduced an instruction-based method to jointly predict the student talk move label with an explanation. Due to consent issues, we re-implement their work as an in-context learning baseline by using Mistral-0.2-instruct-7B model to replace ChatGPT. We leave more LLM studies as future work.

3 Datasets

3.1 Talk Move Categories

Accountable Talk theory includes a large number of talk move types with varying frequency of use and likelihood of application. For tractability, the existing TalkMoves dataset focuses on **7 Teacher Talk Moves**, including keeping everyone together (KPTG), getting student related (GSTUR), restating (RESTAT), revoicing (REVOIC), press for reasoning (PRSREA) or accuracy (PRSACC), none of the above (NONE) and **5 Student Talk Moves**, such as making claims (MCLAIM), providing evidence and reasoning (PRSEVI), reacting to others ideas (RELTO), asking for more information (ASKMI), and none of the above (NONE). In this paper, we focus on the above talk moves for data annotation and discourse analysis.

3.2 Data Collection on Math Tutoring

Saga Education is a non-profit organization that has forged partnerships with school districts across

the U.S. with significant low-income and historically marginalized communities. Saga's tutoring model operates on a hybrid framework, wherein students physically attend sessions within a traditional school classroom. Tutors work remotely, leveraging technology to engage with students effectively. Both tutors and students are equipped with individual computers, facilitating interaction through a virtual workspace. This shared environment integrates video conferencing capabilities with speech, chat messages, digital whiteboards, and other essential tools. These features enable detailed mathematical representations, including charts, graphs, tables, and equations.

SAGA22 Dataset Our study is based on a high school dataset collected in 2022 and provided to us by Saga (denoted as SAGA22, using the year to distinguish with future version of data collections). [6] Institutional Research Boards approved all data collection procedures, and data were only collected from students who provided both personal assent and parent's consent. From this dataset, we selected 148 videos for analysis. The videos were manually transcribed and three annotators annotated the transcribed conversations with talk move labels with annotation guidelines adapted for tutoring sessions. On a subset of 10 videos, our inter-annotator agreement on all labels reaches more than 80 Cohen's kappa on most of the talk move labels, with for a slightly lower score of 75 on one of the labels. Within the 148 transcribed videos, we annotated 121 sessions, resulting in 69.7 hours of videos with 33695 teacher utterances and 11115 student utterances with talk moves labels.

3.3 Talk Move Datasets for Teaching

In addition to the SAGA22 dataset, we reuse two previously published classroom teaching datasets: the TALKMOVES and NCTE-119 datasets described earlier in the related work section.

TALKMOVES The original TALKMOVES dataset (Suresh et al., 2022a) contains 567 mathematics classroom sessions covering a broad array of topics from elementary school to high school. All transcripts in the dataset are human-annotated for 7 teacher and 5 student talk moves. Because the previous work didn't release a validation set, we keep the same 63 sessions in the original test set, and re-split the original training set into 441 training, 63 validation, thus denoting the resulting dataset as TALKMOVES.

	Overall				Per Session	
	sess	T-utt	S-utt	domain	S-num	len
TALKMOVES	567	174168	59823	Mix-Teach	20	30-55
NCTE-119	119	27523	7241	E-Teach	20	50
SAGA22	121	33695	11115	H-Tutor	2-5	35

Table 1: Datasets Summerization

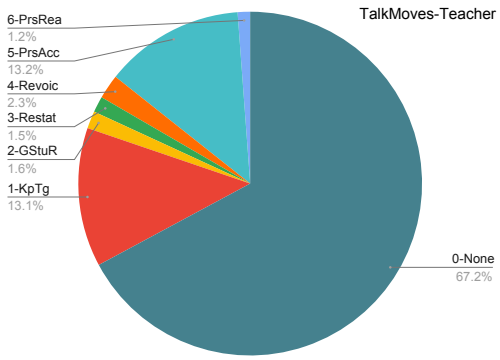
NCTE-119 The original NCTE dataset (Demszky and Hill, 2022) has 1660 sessions in total, however, without any talk move annotation on that. We randomly selected 119 sessions to annotate with talk move labels (thus denoting as NCTE-119 with the total number of 119 to distinguish with future annotation releases), which are mainly for elementary school math.

Table 1 summarizes the overall statistics for the three datasets including the total sessions (sess), Total number of teacher or tutor utterances (T-utt), the total number of student utterances (S-utt), the domain (E-Teach means Elementary Teaching, H-Tutor means High School Tutotring, while Mix-Teaching means math classroom teaching from elementary school to high schools), the average student number (S-num) and average session length in minutes (mins) for all the sessions.

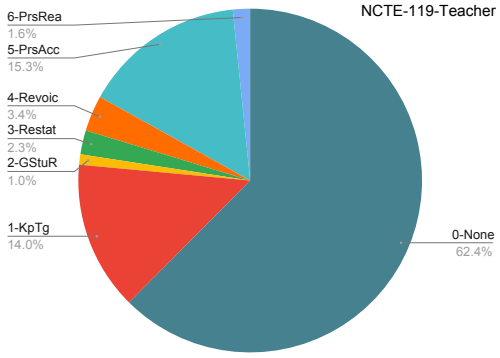
3.4 Teaching vs. Tutoring

Figure 1 indicates that tutor talk moves in the SAGA22 datasets have the similar distribution with teacher talk moves in TALKMOVES and NCTE-119. However, SAGA22 tutoring setting are slightly more in NONE labels, and less in every other talk moves. One possible explanation is that teachers in classroom teaching (TALKMOVES and NCTE-119) might receive more training and engage in more proactive pedagogical practices, Alternatively, the grade distribution, with a higher proportion of high school recordings, could result in reduced communication levels (Muhonen et al., 2024). In Figure 2, students talk moves in tutoring setting are also less than the classroom teaching, which could be indirectly influenced by the reduced use of talk moves by tutors. However, more ASKMI indicates that small group tutoring provides closer interactions, allowing for more opportunities to ask questions. Overall, teaching and tutoring share similar talk move distribution while differ in various amount.In this paper, we primarily focus on transfer learning, leaving a more in-depth analysis for future work. This will involve

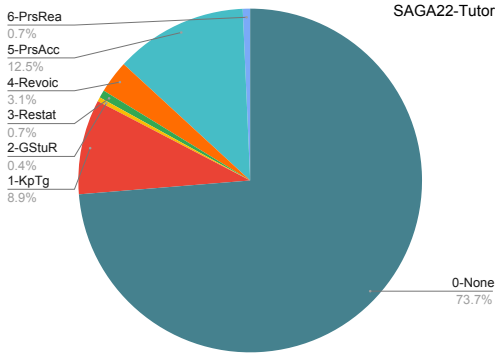
[6] We denote this underlined and uppercase text format (e.g.,SAGA22) to indicate a talk move dataset.



(a) Teacher Talk Moves in TALKMOVES

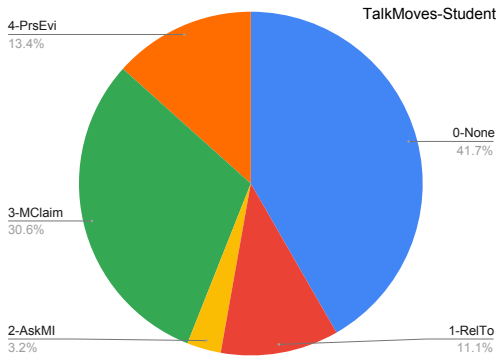


(b) Teacher Talk Moves in NCTE-119

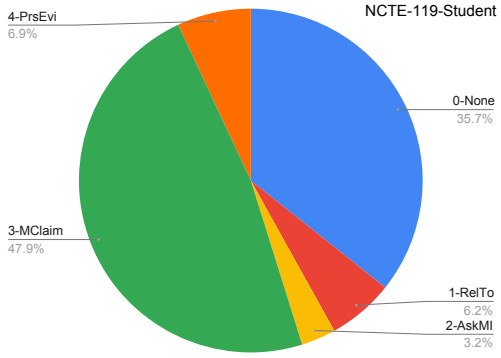


(c) Tutor Talk Moves in SAGA22

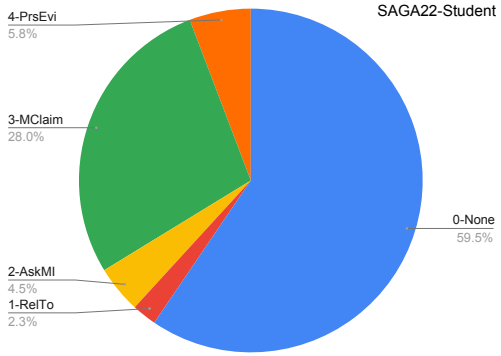
Figure 1: Comparison on Teacher/Tutor Talk Moves



(a) Student Talk Moves in TALKMOVES



(b) Student Talk Moves in NCTE-119



(c) Student Talk Moves in SAGA22

Figure 2: Comparison on Student Talk Moves

exploring latent factors such as class information, grade level, tutor/teacher background, and additional dialogue and discourse analyses (Jurafsky, 1997; Mann and Thompson, 1988; Asher and Lascarides, 2003; Cai et al., 2022).

4 Models

Existing models on talk moves analysis could be categorized into two paradigms: pretrain-finetuning (e.g., Suresh et al., 2018, 2022b), and in context learning (e.g, Wang et al., 2023). We focus on pretrain-finetuning paradigm with "RoBERTa-base" model (Liu et al., 2019) as our backbone foundational model [7]. In this paper, our final target task is talk moves analysis for tutoring. Thus, we can use either the original foundational model or the intermediate talk move models designed for teaching as bases for further fine-tuning. When using raw foundational models as bases for target finetuning, we denote it as **regular pretraining** or **pretraining from-scratch**. When using intermediate models as bases for target finetuning, we denote the secondary pretraining to build the intermediate models as **supplementary pretraining**.

We define a unified model search space $X^{\{\mathbf{P},\mathbf{F}\}}_{\{\mathbf{C},\mathbf{SI}\}}$

[7] Please refer to the appendix Appendix B for more results and analysis on "RoBERTa-large" models.

to cover both teacher/tutor (when $X = T$), and student (when $X = S$) models [8], and their potential improvements. We use **subscripts** to represent the fundamental modeling settings **that are consistent** when transiting from pretraining to finetuning, such as (1) dialogue context (variable **C**), (2) speaker information (variable **SI**). We use **superscripts** to represent training strategies such as (3) the combinations of supplementary pretraining datasets (variable **P**) (4) whether further finetuning on SAGA22 (variable **F**). Hence, assigning values to all or a subset of 5 variables in $X^{\mathbf{P},\mathbf{F}}_{\mathbf{C},\mathbf{SI}}$ will lead to a single specific model ($X^{\{\cdot,\cdot\}}_{\{\cdot,\cdot\}}$, with **a dot** $\cdot$ to represent a specific value) or a set of models ($X^{\{,1\}}_{\{\pm7,\}}$, with **an empty variable value** to represent all possible values for that variable).

4.1 Dialogue Context: C $\in \{-1, \pm7\}$

We follow two settings used in previous works:

1. **Previous-One-Utterance**, is used in Suresh et al. (2018, 2022a), denoted as -1. More specific, for teacher models, they use the previous student utterance as context; While, for student models, they use the previous utterance no matter it is from teacher or student. We follow the same sentence pair modeling as the original papers for this context setting, where the context as sequence 1, and the current utterance as sequence 2.

2. **Previous 7 and Subsequent 7**, is used in (Suresh et al., 2022b), denoted as ±7. We concatenate previous 7 utterances, current utterance, and the subsequent 7 utterances into a single sequence and wrap each utterance as special sentence boundary tokens, thus we keep the original order of the dialogue. Then we force to learn the first special token [CLS] as the context-aware utterance representation of our talk move analysis task [9].

4.2 Speaker Information: SI $\in \{spk, nospk\}$

In previous models (Suresh et al., 2022a,b), the dialogue context didn't use any speaker information during the talk move modeling, which is problematic. For example, without speaker info, when the target utterance simply restating the previous utterance, this could be hard to decide whether it is "Relating to another student" or simply follow the teacher's talk for more information. This could be even worse for longer context settings. Hence, in this paper, we prepend a prefix "T: " or "S: " in front of each utterance to indicate it is said by a teacher/tutor or a student, respectively.

1. **Teacher/Tutor Prefix "T:"** is used for both teacher and tutor utterances to make the model easier transferable for the encodings of "T:" from teaching datasets to tutoring datasets.

2. **Student Prefix "S:"** is also applied to each students' utterances without distinguishing which student that is. We could only do this for TALKMOVES and NCTE-119 datasets, because 20 students are **not distinguishable** in the classroom session, the transcripts always de-identify the different student speakers as the same student "S: ". Noticing this deficit, we make sure that our SAGA22 transcripts **explicitly distinguish** different students as "Student-1", "Student-2". However, to be consistent with previous setting, we still use the single student prefix "S:" to model our tutoring dialogue without distinguishing. We leave the transfer learning from bi-party to multi-party as future work.

When naming the models, we use "spk" and "nospk" as a subscript to indicate with or without speaker prefixes, e.g., $S^{\{,\}}_{\{,spk\}}$ means a set of student models trained with speaker information.

4.3 Supplementary Pretraining Datasets: P $\in \{\varnothing, \text{“}t\text{”}, \text{“}t+n\text{”}, \text{“}t+s\text{”}, \text{“}t+n+s\text{”}\}$

We have three available talkmove datasets, TALKMOVES and NCTE-119 for teaching, and SAGA22 for tutoring. When describing the model name, we use the lower-cased first letter of each dataset name to indicate the pretraining datasets. Since the TALKMOVES has the largest amount of data, we always involve "t" in our combinations of pretraining datasets, resulting in 4 non-empty combinations and 1 empty pretraining set $\varnothing$, as $\mathbf{P} \in \{\varnothing, \text{“}t\text{”}, \text{“}t+n\text{”}, \text{“}t+s\text{”}, \text{“}t+n+s\text{”}\}$. We investigate the best combinations for our supplementary pretraining, denoted as a **superscript**,

[8] Following previous work, we only consider two separate models for tutor and students respectively, leaving the joint model as future work.

[9] Empty utterances will be prepended and padded to make sure there are 15 utterances as inputs; 15 utterance is the longest window size, given we fixed roberta-base model. We decide to keep the two typical and extreme settings -1 and ±7 to show the overall trend across a broad range of options.

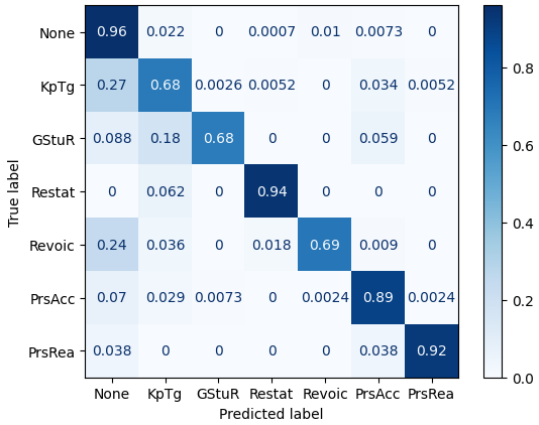


Figure 3: Confusion matrix for the best tutor model $T^{\{t+n,1\}}_{\{\pm 7,spk\}}$ on SAGA22

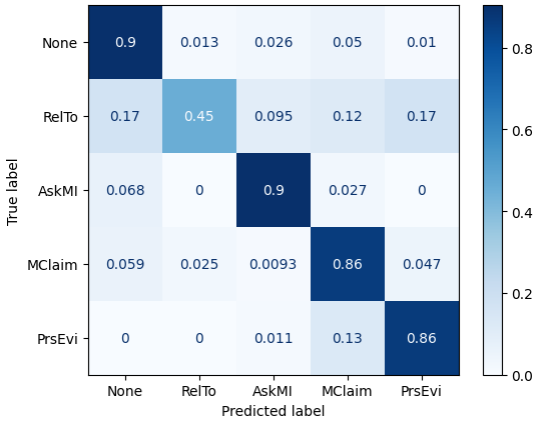


Figure 4: Confusion matrix for the best student model $S^{\{t+n+s,1\}}_{\{\pm 7,spk\}}$ on SAGA22.

e.g., $T^{\{t+n+s,\}}_{\{,\}}$ means a family of tutor models pretrained on the combination of all three datasets.

4.4 Fine-tuning on SAGA22: F $\in \{0, 1\}$

After the above supplementary pertaining on the combination of datasets, the resultant models could be inferred on our SAGA22 dataset with or without any further fine-tuning. It is unknown which is better. When describing the model name, we indicate this further fine-tuning as a superscript on the model tag ('T' or 'S)', e.g., $T^{\{,1\}}_{\{,\}}$ is a family of models eventually fine-tuned on SAGA22.

Model Search With the above 5 variables, we first fixed the variable **P** = "t" (given that large amount TALKMOVES will be necessary in high probability), and only searched over the rest 16 models assigned with the other four binary modeling choices (**X,F,C,SI**), to prioritize the investigation on more interesting modeling factors, such as the dialogue context (**C**) and speaker information (**SI**)). Then we performed extensive search over all other 4 combinations of pretraining datasets **P**. In total, we discovered the best models and conducted the ablation studies by searching over 80 experimental settings.

5 Results

With extensive model search, we summarize our main results in Table 2 and 3 for tutor's and students' talk move analysis respectively. Each table contains 5 categories, the majority baseline, and 2 existing supervised learning work (Suresh et al., 2022a,b), 1 ICL baseline work (Wang et al., 2023), and the last 2 rows are the best training from-scratch model and best-of-all model. For the two existing supervised learning categories, each work can be extended to three equivalent baseline variants in our model search framework:

(1) **Trained from-scratch on SAGA22, $X^{\{\phi,1\}}_{\{,\}}$.** We retrain the models from scratch with only SAGA22 by following the training strategies described in the corresponding paper. Without any other pretraining, It aims to investigate how the existing modeling strategies perform if training only the 121 tutoring sessions.

(2) **Trained on TALKMOVES, no finetuning on SAGA22, $X^{\{t,0\}}_{\{,\}}$.** It aims to show that, without finetuning, how the models initially built for math teaching perform on tutoring data.

(3) **Pretrained on TALKMOVES then finetuning on SAGA22, $X^{\{t,1\}}_{\{,\}}$.** It aims to examine that, with further finetuning on SAGA22, how the models initially built for teaching dataset could be adapted to new tutoring data.

All the best-effort models are selected by the best macro F_1 score on validation set, and the table here only shows the performance on final evaluation on the held-out test sets. We show the macro F_1 score, the accuracy and detailed F_1 score for each label. Our zero-shot ICL models mimic the similar prompts (see Appendix C) used in (Wang et al., 2023), whiling testing on Mistral-0.2-instruct-7B models instead of ChatGPT due to consent issues on our SAGA22 datasets. However, they fail to predict the most easiest(frequent) NONE talkmoves in the supervised learning, because it requires complex reasoning to opt out all other labels. This failure causes extremely low macro F_1 on both tutor and student talkmoves.

Setting	Model	Context	Speaker	Sup-Pretrain	Finetune	F_1	Acc	NONE	KPTG	GSTUR	RESTAT	REVOIC	PRSACC	PRSREA
Majority	T_{Majority}	N/A	N/A	N/A	N/A	12.1	74	85	0	0	0	0	0	0
Suresh et al. (2022a)	$T^{\{\varnothing,1\}}_{\{-1,nospk\}}$	-1	✗	N/A	✓	68.6	89.9	94.5	68.7	5.6	71.8	65.9	86.6	87.0
	$T^{\{t,0\}}_{\{-1,nospk\}}$	-1	✗	TALKMOVES	✗	76.4	89.1	93.7	69.0	56.7	73.7	68.0	84.7	88.9
	$T^{\{t,1\}}_{\{-1,nospk\}}$	-1	✗	TALKMOVES	✓	77.6	91.0	95.0	72.9	51.0	72.2	70.9	87.8	93.3
Suresh et al. (2022b)	$T^{\{\varnothing,1\}}_{\{\pm7,nospk\}}$	±7	✗	N/A	✓	58.6	68.2	93.8	62.5	0.0	47.1	40.2	82.2	86.2
	$T^{\{t,0\}}_{\{\pm7,nospk\}}$	±7	✗	TALKMOVES	✗	75	89.7	94.1	71.3	50.9	73.7	58.2	87.8	88.9
	$T^{\{t,1\}}_{\{\pm7,nospk\}}$	±7	✗	TALKMOVES	✓	73.7	90.7	95.0	**73.4**	43.5	70.3	55.4	87.7	90.9
Wang et al. (2023)	ICL-zero-shot	±7	✓	N/A	✗	24.5	18.5	3.1	30.4	14	36.5	33.3	20.0	34
Best From-Scratch	$T^{\{\varnothing,1\}}_{\{-1,spk\}}$	-1	✓	N/A	✓	70.6	89.8	94.5	69.2	20.5	75.7	64.4	85.9	83.6
Best of All	$T^{\{t+n,1\}}_{\{\pm7,spk\}}$	±7	✓	TALKMOVES +NCTE-119	✓	**82.4**	**91.4**	**95.1**	71.6	**75.4**	**81.1**	**70.3**	**89.6**	**93.3**

Table 2: Best tutor models for each setting on SAGA22 test set.

Setting	Model	Context	Speaker	Sup-Pretrain	Finetune	F_1	Acc	NONE	RELTO	ASKMI	MCLAIM	PRSEVI
Majority	S_{Majority}	N/A	N/A	N/A	N/A	15	59.7	74.8	0	0	0	0
Suresh et al. (2022a)	$S^{\{\varnothing,1\}}_{\{-1,nospk\}}$	-1	✗	N/A	✓	61.5	82.5	89.2	0.0	67.1	77.5	71.5
	$S^{\{t,0\}}_{\{-1,nospk\}}$	-1	✗	TALKMOVES	✗	66.8	82.1	90	25.9	69.1	77.5	71.4
	$S^{\{t,1\}}_{\{-1,nospk\}}$	-1	✗	TALKMOVES	✓	68.2	84.9	91.0	27.1	71.4	82.5	69.0
Suresh et al. (2022b)	$S^{\{\varnothing,1\}}_{\{\pm7,nospk\}}$	±7	✗	N/A	✓	45.0	74.1	85.3	0.0	30.8	66.0	43.1
	$S^{\{t,0\}}_{\{\pm7,nospk\}}$	±7	✗	TALKMOVES	✗	69.5	84.4	91.2	30.8	68.7	81.9	74.7
	$S^{\{t,1\}}_{\{\pm7,nospk\}}$	±7	✗	TALKMOVES	✓	69.6	85.6	91.9	29.1	70.9	82.7	73.5
Wang et al. (2023)	ICL-zero-shot	±7	✓	N/A	✗	25.9	27.3	4.2	40.9	24.4	34.2	25.6
Best From-Scratch	$S^{\{\varnothing,1\}}_{\{-1,spk\}}$	-1	✓	N/A	✓	63.6	84.9	90.8	0.0	73.5	82.5	71.7
Best of All	$S^{\{t+n+s,1\}}_{\{\pm7,spk\}}$	±7	✓	TALKMOVES +NCTE-119 +SAGA22	✓	**76.5**	**87.4**	**92.8**	**48.1**	**79.0**	**84.4**	**78.1**

Table 3: Best student models for each setting on SAGA22 test set.

Best Tutor Model Our best tutor model $T^{\{t+n,1\}}_{\{\pm7,spk\}}$ achieves **82.4 macro** F_1, reaching the same level of performance of existing models for the classroom domain (Suresh et al., 2022b). It is firstly pretrained on the combination of teaching-only datasets TALKMOVES and NCTE-119 using previous 7 and subsequent 7 utterances as context, with speaker information, then further fine-tuned on SAGA22. It achieves the best performance over all talk move categories except for KPTG. With our SAGA22 datasets, our best-effort model trained from-scratch can only get 70.6 macros F_1, particularly failing at predicting GSTUR (20.5 macro F_1), which is using a single utterance as context but adding speaker information on that[10]. Because when we trained models with ±7 context with only the SAGA22 dataset, our best-effort model can get a 61.3 macros F_1 score, and 0 F_1 score on GSTUR, indicating **the limited SAGA22 dataset is not enough to support learning from a longer speaker-aware context**. Figure 3 shows the confusion matrix of best tutor model. The darkness of the diagonal indicates that our model could robustly predict all 7 labels except for KPTG,GSTUR,REVOIC. This pattern is highly coorelated with the frequency of each label in our datasets (see 3.4). **GSTUR are often predicted as KPTG**, because they are relatively similar in the semantics of connecting to students, while the GSTUR is towards a specific student (on a idea) not all general students. However, our current tutoring model didn't distinguish different student speaker, all students utterances are noted as the same speaker prefix "S:", which calls for better multiparty dialogue modeling.

Best Student Model Our best student model $S^{\{t+n+s,1\}}_{\{\pm7,spk\}}$ is firstly pretrained on all three datasets using previous 7 and subsequent 7 utterances as context, with speaker information, then further finetuned on SAGA22. It achieves 76.5 macro F_1, which significantly outperforms all the existing talk move models and best-effort training-from-scratch models, in all student talkmoves, particularly on RELTO and ASKMI. Figure 4 shows the confusion matrix of our best student model. It highlights the failure of predicting RELTO, which is widely mis-predicted with NONE, PRSEVI and MCLAIM. However, without the help of supplementary training on teaching dataset, our best-effort from-scratch model only achieves 0.0 F_1 on RELTO. This poor performance is highly due to its **rare portion of 2.3%** as shown in the subsec-

[10]As shown in 4.1, adding speaker info to tutor model T_{-1} should be similar with no speaker setting, because the previous sentence must be from the student. Although the prefix "SpeakerName:" is likely to be sparse and not a dominant feature in the training data of RoBERTa, such as OpenWebText (Gokaslan et al., 2019), but the prefix seems still help.

tion 3.4), and the unified student speaker prefix "S:" may also **lose important discussion thread information between different students**, and causes the model confused with various student behaviors.

6 Discussion

We first conduct ablation studies on longer context (§6.1) and speaker info (§6.2), respectively. Then we fixed the best fundamental settings using speaker info with ± 7 context, so that we could focus on the impact of supplementary pretraining (§6.3) and finetuning (§6.4) on the model family of $X^{\{,\}}_{\{\pm 7, spk\}}$. For five combinations of pretraining datasets (including ϕ, without pretraining also can be denoted as $X^{\{s,0\}}_{\{,\}}$), we order them in increasing size (s, t, t+n, t+s, t+n+s).

6.1 Ablation Study on Longer Context

In Figure 5a, each bar value shows the performance change when improving the context from -1 to ± 7 while keeping the other modeling options consistent by comparing $X^{\{\cdot,\cdot\}}_{\{-1,\cdot\}}$ and $X^{\{\cdot,\cdot\}}_{\{\pm 7,\cdot\}}$. Most performance change are positive(ranging from 0.1-7.2), except the SAGA22 from-scratch training setting (the first cluster). Because SAGA22 data only is insufficient to support longer context training. Further more, from the left to right, adding the large **TALKMOVES dataset significantly helps to** release the power of long context. However, further adding more NCTE-119 or SAGA22 only have **diminished returns**. This indicates the longer context in talk move analysis requires sufficient training data to help, but adding more data may not help further. Finally, the performance gains from the model with "Speaker" info (red and orange) almost always outperform their corresponding "Non-Speaker" variant (blue and yellow). This indicates that **simple speaker prefix generally make the longer context more efficient**.

6.2 Ablation Study on Speaker Information

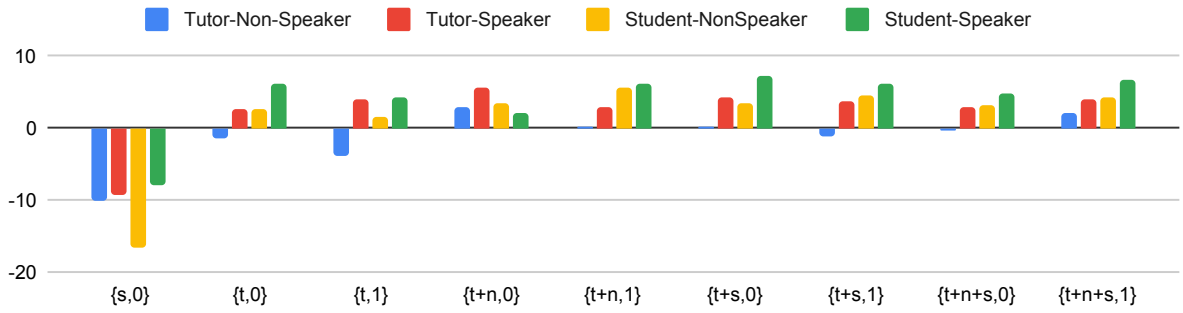
We apply a similar method to illustrate the performance changes when adding speaker information to our models. All positive bar values in Figure 5b shows that **adding speaker information generally helps all models**, and more significant on ± 7 context (red and orange) than on -1 context. More results on teaching-only datasets TALKMOVES in Appendix A shows that retraining with longer context and speaker information also outperformed the previous models on TALKMOVES in large margin, and RoBERTa-large could further help. However, together with the previous findings in best model analysis, the findings in the longer context ablation study, and the known **deficit of bi-party speaker prefixes "T:" and "S:"**, we believe fine-grained speaker modeling could support the personalized learning in the tutoring settings better.

6.3 Ablation Study on Pretraining

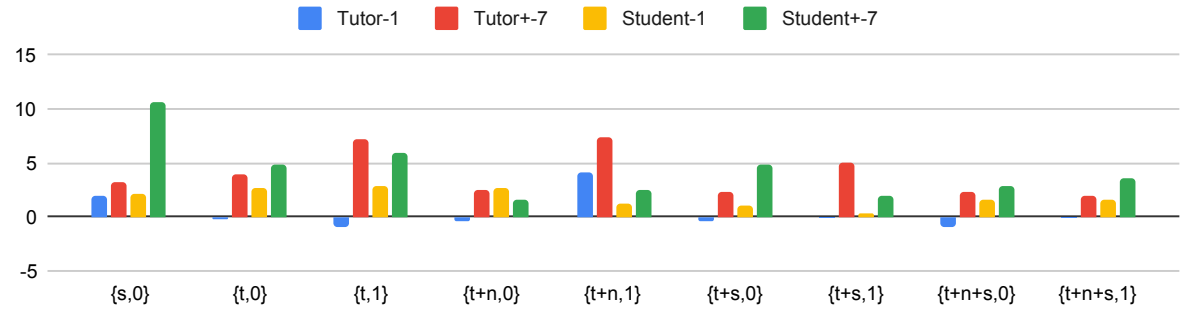
Comparing row 2 and row 1 in Table 4, adding TALKMOVES into pertaining significantly boosts the performance for every talk move label. Even without using any SAGA22 tutoring dataset, the best models in the zero-shot settings (pretraining without SAGA22 and with finetuning tag $\mathbf{F} = 0$) perform 81.8 and 74.4 on tutor and student talk moves. It indicates that, with large teaching talk move datasets, **previous models built for teaching could just work fine on tutoring without finetuning on any the target tutoring data**. Furthermore, adding NCTE-119 helps on tutor models, generally not on student models (comparing row 2/3 vs. 4/5, and 6/7 vs. 8/9). Finally, we noticed that jointly training with SAGA22 adds more information about tutoring domain, leading better performance on tutor models, while not on student models. The best performance for each talk move label (bold numbers) are achieved by different combinations of pretraining datasets, which highlight a future research direction of **finding an optimal mixture or data augmentation for lab distributions that could help all labels**.

6.4 Ablation Study on Finetuning

Comparing the adjacent rows with the same pretraining datasets (e.g., $\{t+s, 0\}$ vs. $\{t+s, 1\}$) in Table 4, we noticed that the performance gains for tutor models are **all positive** (ranged from 0.6-1.9 F_1), while not for student models (finetuning the model pretrained with TALKMOVES and SAGA22 on SAGA22 again hurt the student model performance). It indicates that **further finetuning may not always help especially when the finetuning dataset is small**. Appendix B shows similar finetuning results with RoBERTa-large , which could further improve tutor models from 82.4 to 85.3, but not on student models. More needs to be done to overcome the **catastrophic forgetting** (Kirkpatrick et al., 2017) of continuing finetuning as shown in Table 7 in Appendix B. Furthermore, (Moreau-Pernet et al., 2024) trained GPT-3.5-turbo to re-write transcripts by append-



(a) Performance gains when improving the context from -1 to ± 7.



(b) Performance gains when adding speaker information.

Figure 5: Ablation studies on longer context and speaker information.

$X^{\{,\}}_{\{\pm 7, spk\}}$	Tutor Models									Student Models						
	F_1	Acc	NONE	KPTG	GSTUR	RESTAT	REVOIC	PRSACC	PRSREA	F_1	Acc	NONE	RELTO	ASKMI	MCLAIM	PRSEVI
$\{s, 0\}$	61.3	88.8	94.2	64.5	0.0	47.4	51.4	85.3	86.0	55.6	80.3	89.0	0.0	53.0	76.4	79.0
$\{t, 0\}$	78.9	90.1	94.4	70.8	57.1	82.1	71.3	85.9	90.7	74.4	85.5	91.7	51.4	66.7	82.2	80.0
$\{t, 1\}$	80.8	90.7	94.9	70.4	60.4	**88.9**	70.6	87.2	93.5	75.6	87.2	92.6	44.4	75.6	84.7	80.7
$\{t+n, 0\}$	81.8	90	94.2	69.5	73.8	81.1	71.7	87.6	**94.4**	72.7	86.5	92.7	35.7	71.1	83.9	80.0
$\{t+n, 1\}$	**82.4**	91.4	95.1	71.6	**75.4**	81.1	70.3	**89.6**	93.3	74.4	87.4	82.9	39.5	77.8	**84.9**	76.7
$\{t+s, 0\}$	80.4	91.3	95.2	73.1	63.3	80.0	73.9	88.0	89.5	**77.1**	**87.8**	**93.5**	**49.5**	76.3	84.8	**81.6**
$\{t+s, 1\}$	81.2	91.7	**95.6**	**73.7**	65.6	83.3	71.8	87.3	91.3	74.5	86.6	92.6	49.4	70.9	83.2	76.2
$\{t+n+s, 0\}$	80.2	91.2	95.1	72.0	66.7	75.7	73.7	88.6	89.3	76.2	87.6	93.1	46.3	**79.7**	84.8	76.9
$\{t+n+s, 1\}$	81.5	**91.7**	**95.6**	73.2	66.7	81.1	**74.4**	87.7	91.6	76.5*	87.4	92.8	48.1	79.0	84.4	78.1

Table 4: Model performance for our best fundamental model settings $X^{\{,\}}_{\{\pm 7, spk\}}$ on different pretrain-finetuning settings on SAGA22 test set. Since the model checkpoints are selected from the validation set, we notice some highlighted numbers are better than our selected best student models in the main results.

ing a label to the end of each tutor utterance, the promising results highlight exciting future directions on LLM-based instruction finetuning.

7 Conclusion

In this paper, we investigate how to apply the rich resources on talk move analysis in math teaching to the tutoring domain. We collect a small math tutoring dataset with talk move annotations on 121 tutoring sessions. Then we conduct a thorough examination on existing talk move models on our new tutoring dataset. Based on a unified pretraining-finetuning framework, we systematically search over 4 modeling choices on dialogue context, speaker information, pretraining datasets, and further finetuning to reuse and improve the previous models. We show that without the help on existing models and datasets in the teaching domain, our small amount tutoring data fails to get acceptable performance, and fails to modeling longer context. Our discovered best models with RoBERTa-base achieve 82.4 macro F_1 on tutor talk moves, and 76.5 on student talk moves. Using RoBERTa-large could further improve the tutor models to 85.3, while not on student models. Lastly, extensive ablations studies show that longer context modeling requires sufficient training data and speaker information support; The current bi-party speaker information always helps; However, better tutoring discourse analysis still calls for future support on modeling multi-party speaker information, optimizing the mixture of pretraining data, and better model finetuning strategies.

8 Limitations

Our primary focus is on the pretrain-finetuning framework to transfer the model learning from the classroom teaching to math tutoring. We keep the same foundational architectures and model components unchanged, such as the speaker information, dialogue context, etc. This is suboptimal for two reasons: (1) to be compatible with existing datasets TALKMOVES and NCTE-119, where different students are all deidentified as the same student "S". However, our analysis shows that tutoring dialogue have more personalized behaviors and closer interactions, where multiple party dialogue is required. (2) the optimal dialogue context may be also different in the tutoring sessions, and we only demonstrate the premilinary result for ICL methods where "None" label could not be well-classified by 7B public models. We plan to conduct more comprehensive experiments in the future, incorporating longer contexts and LLMs.

Beyond the above modeling limitations, this work is also limited by the datasets themselves. Specifically: (1) the datasets are all from U.S. classrooms with English-only discourse, (2) the domain is limited to mathematics instruction, and (3) the transcripts alone do not provide sufficient context to adequately ground the participants' discourse behavior. Finally, the sensitive nature of the data, including readily available personally identifiable information about teachers, tutors, and students, poses challenges in evaluating potential biases within the models.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback. This research was supported by the National Science Foundation grant #2222647 and the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805. All opinions are those of the authors and do not reflect those of the funding agencies.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Jon Cai, Brendan D. King, Margaret Perkoff, Shiran Dudy, Jie Cao, Marie Grace, Natalia Wojarnik, Ganesh Ananya, James Martin, Martha Palmer, Marilyn Walker, and Jeffrey Flanigan. 2022. Dependency dialogue acts — annotation scheme and case study. *The 13th International Workshop on Spoken Dialogue Systems Technology*.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Swedish Language Technology Conference and NLP4CALL*, pages 23–35.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E Margaret Perkoff, Michael Regan, Katharina Kann, James H Martin, Martha Palmer, and Sidney D'Mello. 2023. 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*, pages 250–262.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Richard Correnti, Mary Kay Stein, Margaret S Smith, James Scherrer, Margaret McKeown, James Greeno, and Kevin Ashley. 2015. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. *Socializing Intelligence Through Academic Talk and Dialogue. AERA*, 284.

Dorottya Demszky and Heather Hill. 2022. The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Per Engzell, Arun Frey, and Mark D Verhagen. 2021. Learning loss due to school closures during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 118(17):e2022376118.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.

Emily Jensen, Meghan Dale, Patrick J Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Sean Lewis, Victoria Locke, and Chalie Patarapichayatham. 2021. Research brief: student engagement in online learning during covid school closures predicts lower learning loss in fall 2020.

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*.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Sarah Michaels and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue*, pages 347–362.

Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.

Baptiste Moreau-Pernet, Yu Tian, Sandra Sawaya, Peter Foltz, Jie Cao, Brent Milne, and Thomas Christie. 2024. Classifying tutor discursive moves at scale in mathematics classrooms with large language models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 361–365, New York, NY, USA. Association for Computing Machinery.

Heli Muhonen, Eija Pakarinen, Helena Rasku-Puttonen, Anna-Maija Poikkeus, Martti Siekkinen, and Marja-Kristiina Lerkkanen. 2024. Investigating educational dialogue: Variations of dialogue amount and quality among different subjects between early primary and secondary school classrooms. *Learning, Culture and Social Interaction*, 45:100799.

Catherine O'Connor and Sarah Michaels. 2019. Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, 97:166–175.

Catherine O'Connor, Sarah Michaels, and Suzanne Chapin. 2015. Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. *Socializing intelligence through academic talk and dialogue*, pages 111–126.

Chalie Patarapichayatham, Victoria N Locke, and Sean Lewis. 2021. Covid-19 learning loss in texas. *Istation: Dallas, TX, USA*.

Lauren B Resnick, Christa SC Asterhan, and Sherice N Clarke. 2018. Accountable talk: Instructional dialogue that builds the mind. *Geneva, Switzerland: The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the United Nations Educational, Scientific and Cultural Organization (UNESCO)*.

Lauren B Resnick, Sarah Michaels, and Catherine O'Connor. 2010. How (well structured) talk builds the mind. *Innovations in educational psychology: Perspectives on learning, teaching and human development*, pages 163–194.

Carly D Robinson and Susanna Loeb. 2021. High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*, 21(284):1–53.

Karla Scornavacco, Jennifer Jacobs, and Charis Harty. 2021. Automated feedback on discourse moves teachers' perceived utility of a big data tool. *Annual conference of the American Educational Research Association*.

Yu Song, Shunwei Lei, Tianyong Hao, Zixin Lan, and Ying Ding. 2020. Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, page 0735633120968554.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset

for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

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.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022b. Finetuning 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.

Abhijit Suresh, Tamara Sumner, Isabella Huang, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2018. Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5445–5447. IEEE.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Margaret Walshaw and Glenda Anthony. 2008. The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of educational research*, 78(3):516–551.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert.

Noreen M Webb, Megan L Franke, Marsha Ing, Angela C Turrou, Nicholas C Johnson, and Joy Zimmerman. 2019. Teacher practices that promote productive dialogue and learning in mathematics classrooms. *International Journal of Educational Research*, 97:176–186.

Mikyung Kim Wolf, Amy C Crosson, and Lauren B Resnick. 2005. Classroom talk for rigorous reading comprehension instruction. *Reading Psychology*, 26(1):27–53.

Tiffany Zhou, Tomas Molfino, and Jonathan Travers. 2021. The cost of covid: Understanding the full financial impact of covid-19 on districts and schools. *Education Resource Strategies*.

A Results on TALKMOVES

When investigating the two fundamental variables such as the dialogue context (**C**) and speaker information (**SI**)) on talk move analysis, we noticed that the test set of SAGA22 is realtively small. To further verify our findings, we also conduct extensive experiments on TALKMOVES dataset. Table 5 shows the results on the same test set used in two previous papers on teaching talk moves, where $\{-1, nospk\}$ is equivalent to Suresh et al. (2022a), while $\{\pm 7, nospk\}$ is equivalent to Suresh et al. (2022b). Combing both speaker information with ± 7 dialogue context already outperform the previous models. Furthermore, using RoBERTa-large could achieve the SOTA for both teacher and student models on every talk move label.

B Ablation Study on Finetuning with RoBERTa-large

Our model search is mainly based on RoBERTa-base. Table 6 shows the similar ablation studies on finetuning as described in subsection 6.4, but using RoBERTa-large model. Simply scale-up the model could raise the best model performance on teacher models but not on student models. Furthermore, if comparing the adjacent rows with the same pre-training datasets (e.g., t + s, 0 vs. t + s, 1) in Table 6, we noticed that further finetuning the stronger RoBERTa-large model on our small SAGA22 dataset slightly hurts the performance on tutor models, and not stable in student models. Table 7 shows the performance changes on TALKMOVES test set before and after finetuning on the tutoring setting SAGA22, which indicating the forgetting behavior of the pretrained teaching setting(TALKMOVES) after further finetuning on the small SAGA22 training set in tutoring setting. We noticed the RoBERTa-large model forget more than RoBERTa-base model, which may be related to the relative less training data for the student models, In the future work, we plan to investigate more adaptive finetuning strategies such as LoRA (Hu et al., 2021) to finetune models on a small dataset settings, and conduct more comprehensive study one this.

C In-Context Learning

In this paper, we only offer the preliminary studies on prompting methods with longer context and demonstrations. In this section, we describe the detailed prompts we used for each of our setting with running examples to show the possibility of LLMs and highlight the challenges of prompt engineering.

Prompt-based Baseline Models As the prompt shown in Listing 1, we first reuse a zero-shot

$X^{\{t,0\}}_{\{\cdot,\cdot\}}$	Teacher Models									Student Models						
	F_1	Acc	None	KpTg	GStuR	Restat	Revoic	PrsAcc	PrsRea	F_1	Acc	None	RelTo	AskMI	MClaim	PrsEvi
$\{-1, nospk\}$	77.2	87.3	92.7	71.5	61.3	84.1	67.7	82.8	80.0	67.8	78.1	84.6	41.4	56.2	78.7	77.9
$\{-1, spk\}$	78.2	87.8	92.8	72.5	67.2	82.9	67.7	84.4	79.6	71.1	80.0	85.4	52.2	57.7	80.3	79.7
$\{\pm7, nospk\}$	77.0	88.4	93.3	75.6	64.7	79.0	60.1	85.0	81.1	70.1	80.8	87.0	47.0	54.8	80.8	80.7
$\{\pm7, spk\}$	79.2	89.0	93.8	76.0	65.3	83.4	70.1	85.4	80.3	73.4	82.1	87.9	58.7	56.3	81.8	82.3
$\{\pm7, spk\}^*$	**81.3**	**90.1**	**94.5**	**78.5**	**68.9**	**84.9**	**73.0**	**86.8**	**82.4**	**75.5**	**83.9**	**89.3**	**61.6**	**59.8**	**83.6**	**83.2**

Table 5: Model performance on TalkMoves. * denotes the RoBERTa-large results

$X^{\{,\}}_{\{\pm7,spk\}}$	Tutor Models									Student Models						
	F_1	Acc	None	KpTg	GStuR	Restat	Revoic	PrsAcc	PrsRea	F_1	Acc	None	RelTo	AskMI	MClaim	PrsEvi
$\{s, 0\}$	76.9	91.8	96.8	71.6	22.2	92.9	56.9	90.5	93.0	66.0	84.8	93.9	9.5	77.9	81.3	60.2
$\{t, 0\}$	83.5	91.5	94.6	77.6	66.7	96.4	65.0	90.9	96.5	77.2	88.2	90.8	45.2	80.2	88.9	89.2
$\{t, 1\}$	83.3	92.6	96.7	74.9	52.8	82.1	75.6	89.3	94.7	75.6	88.3	93.3	33.3	72.1	88.4	83.9
$\{t+n, 0\}$	85.0	92.2	95.4	76.9	63.9	100.0	67.5	92.5	94.7	75.3	87.7	91.4	42.9	74.4	88.9	82.8
$\{t+n, 1\}$	80.2	92.1	95.7	75.9	33.3	89.3	73.2	92.3	93.0	77.5	89.0	93.2	35.7	76.7	89.6	84.9
$\{t+s, 0\}$	86.1	93.0	96.3	78.1	63.9	100.0	70.7	91.9	98.2	75.7	88.2	91.8	33.3	82.6	89.1	82.8
$\{t+s, 1\}$	**85.3**	92.9	96.1	76.4	66.7	96.4	74.8	91.7	96.5	**76.7**	88.8	95.1	35.7	76.7	86.3	79.6
$\{t+n+s, 0\}$	86.7	93.1	97.2	73.0	72.2	96.4	68.3	90.7	98.2	75.9	89.1	93.3	28.6	81.4	89.6	84.9
$\{t+n+s, 1\}$	85.4	92.7	96.7	72.3	63.9	96.4	78.9	90.1	93.0	75.7	88.0	93.3	35.7	82.6	85.1	82.8

Table 6: Model performance for our best fundamental model settings $X^{\{,\}}_{\{\pm7,spk\}}$ with different pretrain-finetuning settings on SAGA22 test set. All results are based on RoBERTa-large, and comparable to Table 4. Since the model checkpoints are selected with the validation set, we notice some numbers are better than the bolded best student models in the main results.

prompt template from (Wang et al., 2023) to predict the talkmoves. The prompt is made by using the label description and examples in Tables 1 and 2 in the original TalkMoves dataset paper (Suresh et al., 2022a).

Listing 1: System Prompt for Student Talk Moves

```
System:
You are a dialogue analyzer to understand the five talk
    moves for students' utterances, namely "Relating to
    another student", "Asking for more information", "
    Making a claim", "Providing evidence" and "None". They
     have the following meaning: "Relating to another
    student" refers to using commenting on, or asking
    questions about a classmates' ideas, such as "I did not
     get the same answer as her."; "Asking for more
    information" refers to a student requesting more info,
    saying they are confused or need help, such as "I don't
     understand number four."; "Making a claim" refers to a
     student making a math claim, factual statement, or
    listing a step in their answer, such as "X is the
    number of cars."; "Providing evidence" refers to a
    student explaining their thinking, providing evidence,
    or talking about their reasoning, such as "You can't
    subtract 7 because they would only get 28 and you need
    29."; "None" refers to a student utterance that cannot
    be classified as one of the four previous talk moves.
    Considering classifying the student utterance needs
    context information, I add its prior student utterances
     as a context sentence. For example, we need to
    classify the utterance "Same as you". Predicting this
    utterance'talk move need its prior sentence "May answer
     is two" as a context.

So if the prior utterance is "Ah, I thought it was addition
    ". Which kind of talk move the utterance " Me too."
    belongs to?
```

Zero-shot Prompt: Longer Context and Normalized Output When conducting prompt-based methods with previous templates, we noticed that different LLMs could have different outputs in quite different formats. To make our analysis easier, we strictly constrain the output format as Listing 2, and an example of the input and output example are shown in Listing 3. Given the relatively normalized output format, we automatically extract the predicted labels to analyze the generated talk moves and explanations. In this paper, we only use the response explanation to help debug and our prompt engineering. For example, when we use ±7 context for our input example, we found the generated explanation may mistakenly classify other text in the dialogue context instead of the target utterance in the middle.

Listing 2: Noramlizing the output format

```
Considering classifying the target student utterance needs
    context information, I add its prior and future
    utterances as dialogue context. Each utterance contains
     a prefix speaker tag "T:" or "S:" indicating the
    speaker is a teacher or a student, respectively. Please
     predict the label of the target text from one of the
    five talk move labels in the first line as "label:␣X"
    and explain the reason in a new line.
```

Listing 3: Zero-Shot ICL for Student Talk Move

```
User:
prior_text:
T: What should be our first step
target_text:
S: Subtract A two from both sides
future_text:

Assistant:
label: Making a claim
The target text "Subtract A two from both sides" is a step
    in solving a math problem, which falls under the "
    Making a claim" talk move category as the student is
    stating a step in their solution process.
```

Besides that, to make the request more efficient to LLM, we also conduct studies on supporting batch predictions on multiple input examples.

C.1 Results on In-Context Learning

Due to the consent issues, we cannot use our SAGA22 data in ChatGPT. Hence, we select a subset of dataset from TalkMoves to show the baseline performance. Overall, due the limitation

$X^{\{,\}}_{\{\pm 7, spk\}}$	Tutor Models									Student Models						
	F_1	Acc	NONE	KPTG	GSTUR	RESTAT	REVOIC	PRSACC	PRSREA	F_1	Acc	NONE	RELTO	ASKMI	MCLAIM	PRSEVI
$\{t, 1\}$, base	-2.4	-0.9	0.8	-7.6	-6.8	-1.9	1.8	-2.4	-0.4	-2.5	-1.2	3.3	-12.2	-2.8	-3.3	-5.6
$\{t, 1\}$, large	-3.5	-1.4	1.0	-9.5	-1.6	-17.7	3.9	-5.5	-2.9	-5.6	-2.5	6.0	-22.7	-33.6	-9.2	1.6
$\{t+n, 1\}$,base	-3.4	-1.8	1.7	-10.5	-14.4	-15.0	-1.3	-8.4	-1.1	-0.3	-0.3	5.6	-5.9	-7.6	-6.1	-1.9
$\{t+n, 1\}$,large	-1.5	-1.2	-0.2	-3.5	-13.4	-11.4	-1.3	-2.9	6.4	-6.9	-2.5	3.9	-34.0	-23.3	-5.7	3.9
$\{t+s, 1\}$,base	-1.9	-1.4	-0.6	-4.9	-3.9	-7.8	0.9	-0.8	-8.2	-2.5	-1.1	3.1	-13.2	-7.6	-4.4	0.4
$\{t+s, 1\}$,large	-1.7	4.7	0.3	-4.3	-0.7	-3.9	2.9	-2.3	0.3	-2.7	-1.7	3.2	-10.9	-14.4	-5.8	-1.9
$\{t+n+s, 1\}$,base	-1.1	-0.7	-0.4	-0.2	-9.8	-6.2	-0.9	-1.2	0.7	-1.7	-0.2	-1.5	-7.3	10.9	4.0	-2.2
$\{t+n+s, 1\}$,large	-0.8	-0.5	-0.9	0.7	1.7	-3.9	4.6	-0.9	-5.0	-0.4	-0.8	1.0	3.8	-1.3	-3.9	-1.2

Table 7: Model forgetting of the pretrained teaching setting(TALKMOVES) after further finetuning the pretrained models on the small SAGA22 in the tutor setting. All models are based our best fundamental model settings $X^{\{,\}}_{\{\pm 7, spk\}}$. The numbers show the performance differences of TALKMOVES test sets between the models with and without finetuning on SAGA22. "base" and "large" means using RoBERTa-base and RoBERTa-large respectively.

Table 8: In-Context Learning on TALKMOVES

		Teacher	Student
Method	Metrics	F_1	F_1
	Mojority	12.1	15.0
	0-shot(ChatGPT*)	37.5*	32.3*
Zero-ICL	0-shot(LLama2-7B)	23.1	26.2
	0-shot(Mixtral-7B)	24.0	25.9
	2-shot(ChatGPT*)	39.0*	35.6*
Few-ICL	2-shot(LLama2-7B)	25.4	28.4
	2-shot(Mixtral-7B)	26.2	27.9

of computing resource, we only use 7B version of Llama2 and Mixtral for our ICL testing. As shown in Table 8, overall, the performance is much worse than the above supervised learning methods. While, the ICL prompt-enginering indeed took more time for manual tuning, instead of the above model search over existing modeling factors. The ChatGPT performance is using the latest gpt-3.5-0125 on a subset of TALKMOVES dataset, which is not comparable with our preliminary results on SAGA22 test set.