Classifying Tutor Discursive Moves at Scale in Mathematics Classrooms with Large Language Models

Baptiste Moreau-Pernet
Digital Harbor Foundation
Baltimore, United States
baptiste@levi.digitalharbor.org

Yu Tian
Digital Harbor Foundation
Baltimore, United States
terry@levi.digitalharbor.org

Sandra Sawaya
University of Colorado Boulder
Boulder, United States
sandra.sawaya@colorado.edu

Peter Foltz
University of Colorado Boulder
Boulder, United States
Peter.Foltz@colorado.edu

Jie Cao
University of Colorado Boulder
Boulder, United States
jie.cao@colorado.edu

Brent Milne
Saga Education
Boulder, United States
bmilne@saga.org

Thomas Christie
Digital Harbor Foundation
Baltimore, United States
thomas.christie@levi.digitalharbor.org

ABSTRACT

In mathematics tutoring, using appropriate instructional discursive strategies, called "talk moves", is critical to support student learning. Training tutors in the appropriate use of talk moves is a key component of tutor development programs. However, tutor development at scale is a challenge. Recent research has shown that automatic talk moves classification of tutorial discourse can facilitate large-scale delivery of personalized talk moves feedback. In this paper, we build on this work and share our current progress using large language models to classify talk moves in transcripts of tutoring sessions. We report classification results from fine-tuned models, prompt optimization, and supervised embedding vectors classification. The fine-tuned strategy performed best, yielding better performance (.87 macro and .93 weighted f1 score in predicting expert labels) than the current state-of-the-art RoBERTa model. We discuss trade-offs across methods and models.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

LLM classification, math tutor training, discourse analysis

ACM Reference Format:

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), July 18–20, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3657604.3664664

L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-0633-2/24/07
https://doi.org/10.1145/3657604.3664664

1 INTRODUCTION

Students' mathematics learning is facilitated by teachers' invitations to engage in classroom discussions, interact with other students' ideas, and persist in developing their thoughts [12]. Accountable talk theory provides examples of such discussion moves with a series of well-defined, knowledge-grounded instructional strategies for eliciting student thinking [20]. These strategies, known as 'talk moves', serve as guides for teachers to facilitate productive, intentional conversation [6, 19].

Sharing personalized feedback with teachers and tutors about their use of talk moves supports reflection and instructional training [14]. Analysis of classroom conversations can be used to provide this feedback, but has historically required costly manual labeling of talk moves by instructional experts [31]. Performing conversational analysis at scale requires high-quality automated analysis of tutoring sessions to identify the talk moves employed by tutors. The need to provide such feedback at scale increased with the introduction of tutoring platforms such as those created by the National Student Support Accelerator [1] and Saga Education[1].

One recent approach to automated discourse analysis of tutoring sessions leveraged a RoBERTa-based supervised classifier trained on the TalkMoves dataset [27] to classify individual tutor utterances. This classifier achieves talk moves classification with a macro (resp. weighted) f1 score of .78 (resp. 0.91) on a validation set when compared with expert labels. While this performance was of sufficient quality to deploy within the TalkMoves application [26], [29] also reported promising classification results using ChatGPT on student talk moves, suggesting that Large Language Models (LLMs) might be successfully employed to analyze talk moves.

In this paper, we build on this work by investigating the use of LLMs to identify talk moves in video transcripts to support

[1] https://saga.org/

tutor training at scale. We compare three LLM-based classification approaches: (1) using multi-step prompting with OpenAI's GPT-3.5-turbo, (2) embedding utterances and then performing supervised learning on the embedding vectors, and (3) fine-tuning GPT-3.5-turbo to re-write the transcripts with tags added. The fine-tuning method performed best, outperforming the state-of-the-art RoBERTa-based classifier in predicting expert labels. The three approaches exhibit very different performance, highlighting that LLM-powered techniques must be applied judiciously to discourse analysis. Despite this, the success of at least one method suggests that LLMs can improve existing talk moves classification and personalized tutor feedback for training at scale.

2 RELATED WORK

Text classification with LLMs can be performed through supervised and unsupervised learning, serving a wide range of classification objectives: AI-generated language detection [7], conversation features [17], sentiment analysis [11], data annotation [13], etc. Common supervised approaches include using labeled examples to fine-tune a model on a classification task [10] and training machine learning classifiers on text embeddings [22, 28]. Unsupervised approaches use zero or few-shot learning [33] and prompt engineering to induce general-purpose LLMs to classify user inputs. Other approaches use a blend of fine-tuned models and reasoning prompting to overcome LLM context windows limits [24], or use ensembles of LLMs [2].

Until recently, BERT-based models were the state-of-the-art approach for text classification [23, 23, 27], including the optimized transformer model 'RoBERTa' [18, 26]. Lately, these approaches have been outperformed by LLMs. As few-shot learners [5], LLMs show better generalization abilities, thereby improving performance on unbalanced datasets [29]. Their ability to provide decision rationales via chain-of-thought is also helpful for user interpretation of classification decisions. LLMs can handle larger context windows than BERT models, providing valuable context for classification tasks where an utterance's context is critical [27].

3 DATA

We analyzed a dataset of 101 transcripts of 20-60 minute in-class high-school math tutoring sessions. Each session is a math tutoring lesson record with 1 to 4 students. The data was provided by Saga Education, which has implemented a high-dosage tutoring model with small-group, in-class instruction to provide mentorship and personalization in high-poverty urban schools. Tutor conversations were encoded using OpenAI's Whisper Medium speech-to-text algorithm and anonymized. Utterances were broken down at the sentence level and labeled by a research group at University of Colorado Boulder, according to the original Talk Moves classification [25]. Two expert raters labeled each transcript and reported an inter-rater reliability Krippendorff's alpha of .87 [4] on 14 transcripts. We only considered the tutors' talk moves, setting aside student utterances for future work. The distribution of talk moves categories includes seven, very unbalanced labels (see Table 1). We used a 90-10 train/test random split (89 and 12 transcripts).

The present dataset uses similar annotation categories as the 'TalkMoves Dataset' [25], a larger open-source dataset (567 human-annotated transcripts) previously released to help scale the analysis and encoding of teachers' talk moves. The TalkMoves Dataset enabled deeper understanding of teacher and student discourse during classroom mathematics instruction, and is part of recent research on scaling teacher discourse analyses from classroom conversations [3, 8, 9, 15].

Table 1: Distribution of annotated talk moves in the training and testing sets.

Talk Move	Train (25.5k)		Test (3.8k)	
None	18,624	(73%)	2,802	(74%)
Pressing for accuracy	3,256	(13%)	412	(11%)
Keeping everyone together	2,277	(9%)	377	(10%)
Revoicing	811	(3%)	109	(3%)
Restating	204	(0.8%)	51	(1.3%)
Pressing for reasoning	158	(0.6%)	34	(0.9%)
Getting students to relate to another's ideas	92	(0.4%)	16	(0.4%)

4 METHODS

We compared three LLM-based utterance classification methods: (1) an unsupervised method using binary classifications in a multi-step prompting fashion; (2) a supervised classification method that used text embeddings to convert utterances to vector representations, then classified the vectors using an XGBoost model; and (3) a supervised classification method that prompted an LLM to reproduce the entire transcript with TalkMove annotations appended to each utterance. We describe each method in detail below.

4.1 Unsupervised multi-step prompting

To label each utterance with a TalkMove category, we decomposed the multi-category classification task into multiple binary decision steps using a technique called prompt chaining [32]. Early experimentation with LLM prompting suggested that a single-prompt approach (i.e., one rubric describing all categories) suffered due to both the complex nature of the classification task and the poor performance of GPT-3.5-turbo in handling lengthy prompts. Rubrics for each binary decision were adapted from the coding manual used by human raters to annotate the transcripts for talk moves.

To build the chain of prompts, we incorporated several rubrics in a prompt template where we first provided a general definition of 'talk move', inserted the transcription data with the current conversational turn along with the prior and following turns as context, and then specified the classification task (see supplementary materials for the prompt template). Each rubric focused on one talk move category, including its descriptions, examples, and exception cases (see supplementary materials for the rubrics). In each step, the LLM was queried to determine whether the current turn was the target category or not by printing either 'Yes' or 'No'. If the decision was 'No', then the LLM was queried again with the next prompt in the chain. If the decision was 'Yes', the result was recorded and the same process was repeated for the next turn. Note that if 'No'

Previous turns: *'tutor': A closed circle does include that point. So that's like a greater than, an equal to or less than or equal to sign. So it does include that point. So this comes into play here, because if I were to draw my vertical line, let's say right here, it's a little off, but it's meant to pass through those two. That closed circle and that open circle. That closed circle, I am hitting my function here. I am hitting my relation, but in this open circle, that's there to tell me there's nothing there.\n*
Utterance: *'tutor': Thumbs up if we're following.*
Following turns: *'tutor': So on this vertical line that I drew, I'm only hitting once at that closed circle. Are we all following here? Thumbs up if we're following. Awesome. Thank you. Okay, so, Treziah, if you look at this graph here, using what we now know about open and closed circles, does this look like a function?\n'student': I think so, yeah.\n*

Figure 1: Example of a 'Keep Student Together' talk move utterance and 7-sentence contexts.

Tutor: *What would be the reciprocal? <Press for Accuracy>*
Tutor: *Go ahead. <Keep Together>*
Student 1: *What was the question?*
Tutor: *What would be the reciprocal of three over nine? <Press for Accuracy>*
Student 1: *Nine over three.*
Tutor: *Nine over three. <Restating>*

Figure 2: A transcript tagged with correct talk moves, used as expected 'assistant' answer during fine tuning.

was reported for all six categories, the turn would be automatically identified as the 'None' category (i.e., not a talk move). We found that results were improved by ordering rubrics such that better-performing rubrics (tested in isolation) were placed earlier in the chain. The order of the rubrics used in this study was as follows: 1. Restating, 2. Revoicing, 3. Getting students to relate to another's ideas, 4. Pressing for reasoning, 5. Pressing for accuracy, 6. Keeping everyone together. Apart from prompt chaining, we also adopted chain-of-thought (CoT) prompting [30] by requesting the LLM to produce intermediate reasoning concurrently with its final output.

We used OpenAI's GPT-3.5-turbo model for each step, both because it facilitates comparison between results from the supervised and unsupervised methods, and because it is less costly than GPT-4. To extract contextual information for each conversational turn, we used a 7-sentence bidirectional context window recommended by [27] which included 7 preceding and 7 subsequent utterances.

4.2 Supervised classification with text embeddings

Our second approach involved generating vector embeddings of each utterance and then training a machine learning classifier using the embeddings as inputs and human labels as targets. For each tutor utterance to classify, we concatenated the embedding of the conversation's context with the embedding of the utterance itself. For consistency, we used the same context window of 7 sentences (Figure 1). We used OpenAI's embedding API [21], set the vector embedding output dimension to 256, and used 'text-embedding-3-large', the current best-performing model.

We then trained an XGBoost classifier on the resulting embedding vectors. The true labels are the 7 talk moves' categories with a very unbalanced distribution (see Table 1). To account for this distribution, we performed oversampling by duplicating all samples from the 3 least frequent classes. We manually tuned XGBoost parameters to avoid over-fitting.

4.3 Supervised fine-tuning

In our final approach, we used fine-tuning to train GPT-3.5-turbo to re-write transcripts and append talk moves labels to the end of each tutor utterance. This approach provided the LLM with more conversational context. The system prompt instructed the LLM to reproduce the transcript line-for-line, appending a talk move annotation in brackets to the end of each tutor utterance. The system prompt also included an abridged rubric, describing the talk moves labels and giving a few examples of each. We fine-tuned OpenAI's 'gpt-3.5-turbo-0125' using pairs of unlabeled and labeled conversation segments of approximately 200 utterances each, as this roughly corresponded to the model's maximum output length.

The tagged transcripts produced by the model were parsed to recover each utterance's talk move label. The parsing technique required sentence-level alignment for performance evaluation. We found that the fine-tuned model correctly produced output that corresponded utterance-by-utterance to the input for >95% of segments. Re-requesting a prediction typically fixed any misalignment.

5 RESULTS

We report results in terms of Macro and Weighted f1 scores, which are commonly used metrics of prediction quality for multi-class classification problems. All metrics reflect the models' ability to correctly predict expert labels for utterances in the test dataset. High-level averaged results and class-level breakdown are reported in Table 2.

Fine-tuning GPT-3.5-turbo produced the most promising results, achieving a macro (resp. weighted) f1 score of .87 (resp. .93) on the test set. This method outperformed the baseline RoBERTa model by 9 points on the macro f1 score. The most significant improvements were concentrated in a few categories. The fine-tuned GPT-3.5-turbo classified 'Restating' talk moves with a .95 f1 score compared to .65 for RoBERTa, and for 'Getting students to relate to another's ideas' it achieved f1 scores of .75 (GPT-3.5-turbo) vs .55 (RoBERTa). The fine-tuned model also achieved better performance on the 'Keeping everyone together' and 'Revoicing' moves (resp. by 7 and 4 points). However, we observed that the RoBERTa model performed equally or better on the 3 remaining talk moves.

The embeddings classification method delivered a .43 macro f1 score and .82 weighted f1, which is a lower performance than the other trained models. Weak f1 scores on some talk moves categories ('Getting students to realte to another's ideas' (.14), 'Restating' (.14), 'Revoicing'(.15)) and to a smaller extent on 'Pressing for reasoning' (.35) brought the overall macro f1 score down. It demonstrated very good performance on 'None' (.91 f1 score), the most frequent category, with average scores on the rest of the talk moves.

Finally, multi-step prompting using GPT-3.5-turbo yielded underperforming results in comparison to both the baseline RoBERTa-based classifier and our other approaches. The unsupervised method produced average f1 scores (both macro and weighted) far below those of other approaches. A detailed breakdown of how the unsupervised method performed on average and with each label as presented in Table 2 shows that slightly better results were yielded for 'None', 'Getting students to relate to another's ideas', and 'Pressing for accuracy', followed by 'Pressing for reasoning', 'Restating', and 'Revoicing'.

Table 2: Talk moves binary and average classification f1 scores on test set.

Classification setting	Binary classification							Multi-class average	
	None	KET*	Relate*	Restating	Revoicing	Accuracy*	Reasoning*	Macro f1	Weighted f1
Supervised, RoBERTa-based classifier	0.95	0.74	0.55	0.65	0.71	**0.90**	**0.97**	0.78	0.91
Unsupervised multi-step prompting	0.54	0.41	0.39	0.33	0.30	0.39	0.37	0.16	0.43
Supervised fine-tuning	**0.96**	**0.81**	**0.75**	**0.95**	**0.76**	0.88	0.94	**0.87**	**0.93**
Supervised classification with text embeddings	0.91	0.61	0.14	0.14	0.15	0.71	0.35	0.43	0.82

**KET: Keeping everyone together; *Relate: Getting students to relate to another's ideas; *Accuracy: Pressing for accuracy; *Reasoning: Pressing for reasoning*

6 DISCUSSION AND FUTURE WORK

The purpose of this study was to evaluate whether Large Language Models can improve upon state-of-the-art classification of talk moves in tutorial transcripts in support of tutor training. We evaluated several LLM approaches and found that fine-tuning GPT-3.5-turbo produces utterance labels that align better with human annotations than the baseline RoBERTa approach. Critically, fine-tuning GPT-3.5-turbo is not an arduous task: it requires a relatively small dataset of example conversations (less than 100) and a Python script to break the text into chunks, query the API, and parse the model outputs. In contrast, RoBERTa is typically trained and deployed on managed compute resources, requiring both configuration and maintenance. Moreover, the fine-tuned GPT-3.5-turbo model is accessed as any OpenAI model, increasing ease-of-use.

We showed that GPT-3.5-turbo with careful prompting performs worse. This is not surprising: supervised models learn from many examples, and the unsupervised approach was not afforded the same luxury. However, it does challenge the received wisdom that LLMs are few or zero-shot learners [5], at least in this context.

We explored the supervised classification of embedded utterances to know whether the semantic content of utterances would effectively encode each talk move category. While overall classification results were promising, a few categories performed particularly poorly. In particular, model predictions 'Relating', 'Restating', and 'Revoicing' were very low quality. As these talk move categories describe a tutor follow-up to previous utterances, we believe that sufficient context, *and the relationship between the utterance and that context*, are critical for classifying them. It is likely that the embedding approach does not sufficiently capture that relationship. This emphasizes that a talk move's meaning is very context-specific and not captured solely through vocabulary.

Several possible advantages might explain the success of a fine-tuned model over the other approaches. First, the 200-utterance segments of conversation input provided substantially more context than was available to the other models. A fairer comparison would provide larger context in the other approaches, though some models (like RoBERTa) have architectural constraints on context length. Second, in re-writing transcripts with talk move tags appended, the token production process is similar to using chain-of-thought methods. Chain-of-thought allows LLMs to produce a sequence of statements whereby the models can condition each token on all tokens already produced. Similarly, appending tags to tutor utterances allows the model to utilize both the utterance itself and all past turns to determine the appropriate talk move label. This supposition is supported by its high performance on exactly those context-specific categories: Relate, Restating, and Revoicing.

Furthermore, the fine-tuning strategy with GPT-3.5-turbo provides top performance at an average inference cost of $0.20 per tutoring session, cheaper than the prompt chaining method ($.75), but higher than embeddings classification ($.02). Costs depend on model providers and are subject to change.

We intend to explore several avenues to extend this work. First, both prompts and chain-of-prompt sequencing have potential for improving the unsupervised approach, as well as prompt optimization tools such as DSPy [16]. Second, we avoided using GPT-4 in this analysis due to its prohibitive costs in production but using more powerful models could substantially improve results. It may be a scalable option if costs decrease. We would also like to investigate the use of open source LLMs. Third, the supervised classification of text embeddings included a few parameter choices that could be modified. We used small embedding dimensions, fearing that large ones would lead to model overfitting. However, this assumption could be tested and collecting a larger dataset may alleviate this problem. Finally, all approaches except the fine-tuned model are constrained by a bi-directional context of 7 utterances around the 'target' utterance to be classified. While we expected this to provide sufficient context, a larger context window might be beneficial.

In conclusion, LLMs are useful for automated analysis of tutoring transcripts at scale, but only when fine-tuned and used in a manner that leverages their unique strengths. The fine-tuned strategy requires an upfront training cost, but produces highly efficient predictions of tags: Each utterance must be included as input exactly once, whereas the other approaches require each utterance to be included in the context of multiple utterances (14 in this study). This reduces the number of input tokens used for context, massively decreasing the labeling cost for tutoring sessions at scale.

ACKNOWLEDGMENTS

This work was supported by the Learning Engineering Virtual Institute (LEVI)[2] and is housed at the Digital Harbor Foundation.

[2]http://learning-engineering-virtual-institute.org

REFERENCES

[1] National student support accelerator. toolkit for tutoring programs., 2024. https://studentsupportaccelerator.org/tutoring, Last accessed on 2024-04-05.

[2] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*, 2023.

[3] S. Alic, D. Demszky, Z. Mancenido, J. Liu, H. Hill, and D. Jurafsky. Computationally identifying funneling and focusing questions in classroom discourse. *arXiv preprint arXiv:2208.04715*, 2022.

[4] B. Booth, J. Jacobs, J. Bush, B. Milne, T. Fischaber, and S. DMello. Human-tutor coaching technology (htct): Automated discourse analytics in a coached tutoring model. pages 725–735, 03 2024.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] A. Candela, M. Boston, and J. Dixon. Discourse actions to promote student access. *Mathematics Teacher: Learning and Teaching PK-12*, 113:266–277, 04 2020.

[7] C. Chen and K. Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

[8] D. Demszky and H. Hill. The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*, 2022.

[9] D. Demszky, J. Liu, H. C. Hill, D. Jurafsky, and C. Piech. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 0(0):01623737231169270, 0.

[10] S. Do, E. Ollion, and R. Shen. The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, page 004912412211345, 12 2022.

[11] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.

[12] M. Franke, A. Turrou, N. Webb, M. Ing, J. Wong, N. Shin, and C. Fernandez. Student engagement with others' mathematical ideas. *The Elementary School Journal*, 116:126–148, 09 2015.

[13] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, et al. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*, 2023.

[14] J. Jacobs, K. Scornavacco, C. Harty, A. Suresh, V. Lai, and T. Sumner. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631, 2022.

[15] E. Jensen, M. Dale, P. Donnelly, C. Stone, S. Kelly, A. Godley, and S. D'Mello. Toward automated feedback on teacher discourse to enhance teacher learning. pages 1–13, 04 2020.

[16] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[17] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*, 2023.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] S. Michaels, C. O'Connor, and L. Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Stud Philos Educ*, 27:283–297, 07 2008.

[20] C. O'Connor, S. Michaels, and S. Chapin. *"Scaling Down" to Explore the Role of Talk in Learning: From District Intervention to Controlled Classroom Study*, pages 111–126. 04 2015.

[21] OpenAI. Embeddings, 2024. https://platform.openai.com/docs/guides/embeddings, Last accessed on 2024-04-15.

[22] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada. Text clustering with llm embeddings. *arXiv preprint arXiv:2403.15112*, 2024.

[23] S. Pugh, S. K. Subburaj, A. R. Rao, A. E. Stewart, J. Andrews-Todd, and S. K. D'Mello. Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. *Proceedings of The 14th International Conference on Educational Data Mining*.

[24] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

[25] A. Suresh, J. Jacobs, C. Harty, M. Perkoff, J. H. Martin, and T. Sumner. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *arXiv preprint arXiv:2204.09652*, 2022.

[26] A. Suresh, J. Jacobs, V. Lai, C. Tan, W. Ward, J. H. Martin, and T. Sumner. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. *arXiv preprint arXiv:2105.07949*, 2021.

[27] A. Suresh, J. Jacobs, M. Perkoff, J. H. Martin, and T. Sumner. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington, July 2022. Association for Computational Linguistics.

[28] C. Wang, P. Nulty, and D. Lillis. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '20, page 37–46, New York, NY, USA, 2021. Association for Computing Machinery.

[29] D. Wang, D. Shan, Y. Zheng, K. Guo, G. Chen, and Y. Lu. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. 07 2023.

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[31] M. Wolf, A. Crosson, and L. Resnick. Classroom talk for rigorous reading comprehension instruction. *Reading Psychology*, 26:27–53, 03 2005.

[32] T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–22, 2022.

[33] J. Zhang, P. Lertvittayakumjorn, and Y. Guo. Integrating semantic knowledge to tackle zero-shot text classification. In J. Burstein, C. Doran, and T. Solorio, editors, *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 1031–1040, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.