# Can ChatGPT Detect Student Talk Moves in Classroom Discourse? A Preliminary Comparison with Bert

Deliang Wang
Faculty of Education
The University of Hong Kong
Hong Kong
wdeliang@connect.hku.hk

Dapeng Shan
Faculty of Engineering
The University of Hong Kong
Hong Kong
dpshan@cs.hku.hk

Yaqian Zheng
Faculty of Education
Beijing Normal University
China
zhengyq@mail.bnu.edu.cn

Kai Guo
Faculty of Education
The University of Hong Kong
Hong Kong
kaiguo@connect.hku.hk

Gaowei Chen
Faculty of Education
The University of Hong Kong
Hong Kong
gwchen@hku.hk

Yu Lu
Advanced Innovation Center for Future Education, Faculty of Education
Beijing Normal University
China
luyu@bnu.edu.cn

## ABSTRACT

Student utterances in classrooms contain valuable information related to learning. Researchers have employed artificial intelligence techniques, particularly supervised machine learning, to analyze student classroom discourse and provide teachers and students with meaningful feedback. However, supervised models necessitate manual annotation of data, which is both laborious and time-consuming. Recently, OpenAI has released the pre-trained large language model, ChatGPT, which can engage in conversations and provide human-like responses to prompts. Therefore, this study examines the use of ChatGPT in automatically analyzing student utterances and evaluates its capability in addressing the challenge of manual data annotation. Specifically, we compare the performance of ChatGPT with a Bert-based model in identifying student talk moves in mathematics lessons. The preliminary results indicate that while ChatGPT may not perform as strongly as the Bert-based model, it demonstrates potential in detecting specific talk moves, such as *relating to another student*. Additionally, ChatGPT offers clear explanations for its predictions, resulting in higher interpretability compared to the Bert-based model, which operates as a black box.

## Keywords

Classroom discourse, talk move, ChatGPT, Bert.

## 1. INTRODUCTION

Student utterances in class contain rich information about their communicative goals or actions [4], ideas [5], knowledge states, and abilities [8], which are correlated to learning. To assist teachers in understanding student utterances and providing adaptive teaching, studies have adopted artificial intelligence (AI) techniques to model student utterances. For example, researchers have used Long Short-Term Memory (LSTM) networks to estimate whether students have mastered example questions based on their utterances [2]. However, most studies rely on supervised models, which have a significant limitation. Supervised models typically require researchers to manually label a large amount of data in advance, which is laborious and time-consuming. In addition, the trained models may not be easily generalized to other educational contexts.

With the advancement of natural language processing (NLP) techniques, pre-trained large language models such as BERT [3] and GPT-3 [1] have emerged and have demonstrated strong performance on various downstream tasks. Recently, ChatGPT, the latest large language model from OpenAI, has also gained popularity quickly across the whole world [1]. Based on GPT-3 [1] and InstructGPT [12], ChatGPT can engage in conversations with users and generate human-like text responses based on their prompts, such as debugging code and writing essays, which shows exceptional ability in understanding language and indicates great potential in various tasks.

Thus, this paper investigates the ability of ChatGPT to automatically analyze student utterances in classroom discourse and explores whether it can address the challenge of manually annotating data. Specifically, this paper compares ChatGPT and a BERT-based model in automatically detecting student talk moves (i.e., specific dialogic acts) in mathematics lessons. The experiment results show that the BERT-based model outperforms ChatGPT, but ChatGPT demonstrates potential in detecting specific talk moves. In addition, ChatGPT provides clear explanations for its predictions on student utterances, while the BERT-based model operates as a black box and lacks interpretability.

[1]https://openai.com/blog/chatgpt/

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. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519, Bengaluru, India, July 2023. International Educational Data Mining Society.


https://doi.org/10.5281/zenodo.8115772

## 2. RELATED WORK

### 2.1 Automated Models on Student Discourse

Recently, many studies have employed AI techniques to analyze student discourse and provide feedback for learners and teachers. This can be further divided into offline and online learning based on their educational contexts. In offline learning, researchers have not only explored the use of AI chatbots to support students' learning in multiple subjects such as English [10] and engineering [23] but also leveraged LSTM to detect breakdowns in students' conversations with the chatbot in classrooms [11]. Additionally, they have investigated building a convolutional neural network (CNN) based model to automatically identify the semantic content of student dialogue (e.g., prior knowledge, uptake, and querying) in math, science, and physics lessons [17]. In online learning, researchers have used decision trees and naive bayes to classify learners' speech acts (e.g., statement and request)[15], and utilized a Bert-based model to predict learners' dialogue acts (e.g., question, answer, and statement) in science lessons[9]. Student dialogue in collaborative learning is often analyzed to facilitate their learning. For example, researchers have leveraged transformers to automatically classify the dialogue into cumulative, disputational, and exploratory talk [21], and built learners' knowledge graphs to estimate their knowledge [24]. Additionally, students' emotions (e.g., positive and negative) and their behaviour (e.g., knowledge building and off-topic activities) has also been modeled by Bert-based models [26].

### 2.2 ChatGPT

ChatGPT is one of the latest pre-trained large language models developed by OpenAI, which has attracted over 1 million users within 5 days of its release in 2022. Compared to previous language models (i.e., GPT-1 [13], GPT-2 [14], GPT-3 [1]) that may generate harmful and untruthful content, ChatGPT employs the reinforcement learning from human feedback (RLHF) method [18, 12] that changes the training objective from predicting the next token to following human instructions safely, which enables it to generate human-like answers to users' questions. This makes it a powerful tool for various applications, such as composing poetry, commenting on news, and editing language. In the context of education, ChatGPT demonstrates great potential in facilitating learning. For example, users have explored using ChatGPT in language learning (e.g., translating language and providing feedback on writing) [22, 7] and programming learning (e.g., interpreting and debugging code) [20]. As ChatGPT is a relatively new model, there are limited studies examining its use in education. In this paper, we investigate ChatGPT's capability in identifying student talk moves in classroom discourse, to evaluate its potential for providing teachers with effective feedback.

## 3. METHOD

This section describes how this paper compares the performance of a Bert-based model (i.e., BertForSequenceClassification) and ChatGPT in detecting student talk moves in a dataset.

### 3.1 Data

In this paper, we selected *TalkMoves* [19], a classroom discourse dataset on K-12 mathematics lessons as our data source. Due to the unavailability of an API interface from OpenAI[2], we were only able to repeatedly utilize ChatGPT for predicting the talk move of a student utterance, which was a time-consuming and challenging task. To address this limitation, we selected a subset from from the *TalkMoves* dataset. Specifically, we chose all primary school lessons in 2021, consisting of 34 transcripts with a total of 4041 student utterances, each of which was annotated with a talk move label. Talk moves refer to specific dialogic acts reflecting the intention of an utterance and speakers' communicative goals [16], and accurately identifying student talk moves is important for teachers to make appropriate response to students. Student talk moves in the *TalkMoves* dataset include *relating to another student*, *asking for more information*, *making a claim*, *providing evidence*, and *None*[19]. The data were not evenly distributed, which can be seen in Table 1. For each type of talk move, we randomly selected 90% of the data as the training set and used the remaining 10% as the testing set. We compared the performance of a Bert-based model and ChatGPT on the testing set.

**Table 1: Distribution of student talk moves**

| Talk Move | Number |
|---|---|
| Relating to Another Student | 353 |
| Asking for more Information | 108 |
| Making a Claim | 1135 |
| Providing Evidence | 664 |
| None | 1781 |

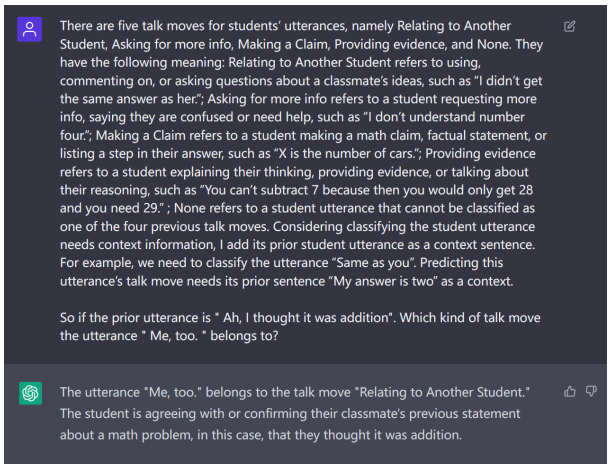


**Figure 1: An example of the prompt for ChatGPT and its answer.**

### 3.2 Bert-based Model

In this paper, we selected a Bert-based model (i.e., BertForSequenceClassification) as a baseline because its training process (e.g., next sentence prediction) considered the context information [3] and it showed strong performance in text classification tasks [25]. For this specific task of student talk move detection in the TalkMoves dataset, we treated it as a 5-way sequence classification problem. To account for

[2]This work was conducted in December 2022, and the API interface of ChatGPT was made publicly available by OpenAI in March 2023.

**Table 2: Overall performance of the Bert-based model and ChatGPT**

| | Bert-based | ChatGPT |
|---|---|---|
| **accuracy** | 0.746667 | 0.582222 |
| **precision** | 0.651488 | 0.503348 |
| **recall** | 0.561072 | 0.519613 |
| **f1 score** | 0.599339 | 0.483108 |

the importance of dialogue context in identifying talk moves, we set the input of the model as a student utterance concatenated with its preceding utterance. The representation of the input, obtained from the BERT architecture, was fed into a linear layer, and the softmax function was used to predict the talk move. When training the model, we set the learning rate, optimizer, batch size, and number of epochs as 1e-5, AdamW, 32, and 6 respectively.

### 3.3 ChatGPT

The key to using ChatGPT to detect student talk moves is to provide suitable prompts. We explored several different prompts and selected a suitable one. Specifically, inspired by the idea of few-shot learning from GPT-3 [1], we first provided ChatGPT with the definition and an example of each talk move based on their original description [19]. For example, *Relating to Another Student refers to using, commenting on, or asking questions about a classmate's ideas, such as "I didn't get the same answer as her."* Then, we also clarified the importance of context information, similar to what we did in the Bert-based model. Finally, we asked ChatGPT to predict the talk move of a student utterance. We attempted to provide a batch of student utterances for ChatGPT, but it outputted multiple predictions that did not match the number of inputs in the batch. Thus, we asked ChatGPT to identify the student talk move one utterance by one utterance. An example of the prompt we gave to ChatGPT and its answer can be seen in Figure 1. Considering that ChatGPT may generate inconsistent answers to the same question, this preliminary study adopted the first output as the prediction.

## 4. RESULT

### 4.1 Performance

The Bert-based model achieved 0.7523 in F1 score and 0.8164 in accuracy on the testing set. Considering the role of talk moves in improving learning [6], we eliminated student utterances tagged with *None* and only compared the performance of the Bert-based model and ChatGPT in identifying the other four meaningful talk moves, as seen in Table 2. It is evident that the Bert-based model outperforms ChatGPT in accuracy, precision, recall, and F1 score. For instance, the accuracy of the Bert-based model in detecting the four talk moves is around 0.747 while that of ChatGPT is only around 0.58.

Table 3 illustrates the performance of the Bert-based model and ChatGPT in each type of the four talk moves. The Bert-based model performs better in *asking for more information, making a claim, and providing evidence* while ChatGPT achieves stronger performance in *relating to another student.* Additionally, ChatGPT also shows potential in *asking for more information* with 1.0 in the recall metric.

### 4.2 Interpretability

Despite the superior performance in detecting student talk moves, the Bert-based model is limited in interpretability, as it cannot provide the reason why the prediction is obtained. By contrast, ChatGPT offers clear explanations for each prediction. For example, given a student utterance, "*I did the same thing as Josh did*", the prior student utterance of which is "*I did partial products, and I got it correct*", ChatGPT annotates the utterance with *Relating to Another Student* and gives the following explanations:

> *The student is using and commenting on a classmate's idea (Josh's method, as mentioned in the prior utterance). The student is indicating that they followed the same process as Josh in solving the problem. This talk move involves using, commenting on, or asking questions about a classmate's ideas.*

In this case, ChatGPT points out how the prediction follows the definition of *Relating to Another Student* in the prompt. Another example is that, given a student utterance, "*I'm struggling all the steps in the process.*", the prior utterance of which is "*Okay, we solve it separately and then we all talk about how we each did it?*", ChatGPT labels the utterance with *asking for more information* and gives the following explanations:

> *This talk move is characterized by a student requesting more information or asking for help. In this case, the student is indicating that they are struggling with understanding the steps in the process, which is a request for more information and assistance.*

Similarly, we can see that ChatGPT explains how the prediction is made by analyzing which part of the utterance follows the definition. This feature demonstrates great potential in addressing the interpretability issue of deep learning-based classroom discourse models (e.g., the Bert-based model in this study).

## 5. DISCUSSION AND CONCLUSION

To automatically analyze classroom discourse without laborious and time-consuming manual annotation of data, the work investigates the capability of the latest large language model, ChatGPT, in identifying student talk moves in mathematics lessons. To achieve this goal, we compare ChatGPT and a Bert-based model in the subset of a classroom discourse dataset. The preliminary results show that although the BERT-based model achieves superior performance, ChatGPT demonstrates the potential in detecting specific talk moves (e.g., *relating to another student*). Specifically, ChatGPT can effectively analyze student utterances that include obvious indicators of talk moves as they align with the definition of the prompt. However, ChatGPT struggles to detect talk moves that are hidden in complex classroom discourse.

In addition, ChatGPT has a significant advantage over the Bert-based model, as it is able to provide detailed and clear explanations for its predictions on student utterances. This feature makes it more interpretable and can increase user

Table 3: Performance of the Bert-based model and ChatGPT in each type of talk move

| | Model | Relating to Another Student | Asking for more Information | Making a Claim | Providing Evidence |
|---|---|---|---|---|---|
| **precision** | Bert-based | 0.695652 | **0.888889** | **0.864078** | **0.808824** |
| | ChatGPT | **0.727273** | 0.458333 | 0.64486 | 0.686275 |
| **recall** | Bert-based | 0.457143 | 0.727273 | **0.787611** | **0.833333** |
| | ChatGPT | 0.457143 | **1.000000** | 0.610619 | 0.530303 |
| **f1 score** | Bert-based | 0.551724 | **0.800000** | **0.824074** | **0.820896** |
| | ChatGPT | **0.561404** | 0.628571 | 0.627273 | 0.598291 |

trust. In contrast, the Bert-based model directly gives predictions without explanations, operating as a black box for users.

As a preliminary study, this exploratory work has several limitations. Firstly, because when the study was conducted, OpenAI did not make the API interface public, the sample size was limited to a relatively small scale, which may cause a bias in the findings. Secondly, as ChatGPT is sensitive to the prompts, changing the prompt may result in different answers. Thus, the choice of prompts may also introduce a bias in the findings. Additionally, it is difficult to determine the optimal prompt for generating the most accurate responses. Thirdly, even if ChatGPT is given the same prompt, it may still generate different answers at different times, which may lead to inconsistency in the results. Fourthly, the study only examines the use of ChatGPT in identifying student talk moves while classroom discourse also carries other valuable information, not limited to talk moves. Besides, teachers' dialogic approach in class can significantly affect teaching and learning. Thus, promising research directions for ChatGPT in classroom discourse include evaluating its ability to identify multiple meaningful characteristics of dialogues between teachers and students in a more extensive dataset with well-crafted prompts and addressing its consistency issue in gnerating answers.

## 6. ACKNOWLEDGMENTS

This work was supported by Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221), and by the Innovation and Technology Commission of the Government of the HKSAR (Grant No.: ITB/FBL/7026/20/P).

## 7. REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[2] J. Chen, Z. Liu, and W. Luo. Wide & deep learning for judging student performance in online one-on-one math classes. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II*, volume 13356 of *Lecture Notes in Computer Science*, pages 213–217. Springer, 2022.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 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, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] A. Ezen-Can and K. E. Boyer. Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of the 6th International Conference on Educational Data Mining, 2013*, pages 20–27. International Educational Data Mining Society, 2013.

[5] A. Ezen-Can, J. F. Grafsgaard, J. C. Lester, and K. E. Boyer. Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, LAK '15*, pages 280–289. ACM, 2015.

[6] 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.

[7] L. Kohnke, B. L. Moorhouse, and D. Zou. Chatgpt for language teaching and learning. *RELC Journal*, page 00336882231162868, 2023.

[8] S. P. Leeman-Munk, E. N. Wiebe, and J. C. Lester. Assessing elementary students' science competency with text analytics. In *Learning Analytics and Knowledge Conference 2014*, pages 143–147. ACM, 2014.

[9] J. Lin, S. Singh, L. Sha, W. Tan, D. Lang, D. Gašević, and G. Chen. Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127:194–207, 2022.

[10] K. Mageira, D. Pittou, A. Papasalouros, K. Kotis, P. Zangogianni, and A. Daradoumis. Educational ai chatbots for content and language integrated learning. *Applied Sciences*, 12(7):3239, 2022.

[11] W. Min, K. Park, J. B. Wiggins, B. W. Mott, E. N. Wiebe, K. E. Boyer, and J. C. Lester. Predicting dialogue breakdown in conversational pedagogical agents with multimodal lstms. In *Artificial Intelligence in Education - 20th International Conference, AIED*

*2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II*, volume 11626 of *Lecture Notes in Computer Science*, pages 195–200. Springer, 2019.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[15] B. Samei, H. Li, F. Keshtkar, V. Rus, and A. C. Graesser. Context-based speech act classification in intelligent tutoring systems. In *International conference on intelligent tutoring systems*, pages 236–241. Springer, 2014.

[16] J. R. Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

[17] Y. Song, S. Lei, T. Hao, Z. Lan, and Y. Ding. Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, 59(3):496–521, 2021.

[18] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[19] 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. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4654–4662. European Language Resources Association, 2022.

[20] H. Tian, W. Lu, T. O. Li, X. Tang, S.-C. Cheung, J. Klein, and T. F. Bissyandé. Is chatgpt the ultimate programming assistant–how far is it? *arXiv preprint arXiv:2304.11938*, 2023.

[21] S. Ubani and R. Nielsen. Classifying different types of talk during collaboration. In *International Conference on Artificial Intelligence in Education*, pages 227–230. Springer, 2022.

[22] D. Yan. Impact of chatgpt on learners in a l2 writing practicum: An exploratory investigation. *Education and Information Technologies*, pages 1–25, 2023.

[23] C.-C. Yuan, C.-H. Li, and C.-C. Peng. Development of mobile interactive courses based on an artificial intelligence chatbot on the communication software line. *Interactive Learning Environments*, pages 1–15, 2021.

[24] Y. Zhen, L. Zheng, and P. Chen. Constructing knowledge graphs for online collaborative programming. *IEEE Access*, 9:117969–117980, 2021.

[25] L. Zheng, J. Niu, and L. Zhong. Effects of a learning analytics-based real-time feedback approach on knowledge elaboration, knowledge convergence, interactive relationships and group performance in cscl. *British Journal of Educational Technology*, 53(1):130–149, 2022.

[26] L. Zheng, L. Zhong, and J. Niu. Effects of personalised feedback approach on knowledge building, emotions, co-regulated behavioural patterns and cognitive load in online collaborative learning. *Assessment & Evaluation in Higher Education*, 47(1):109–125, 2022.