Received: 21 October 2024 | Accepted: 24 April 2025

DOI: 10.1111/bjet.13604

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy

Deliang Wang | **Gaowei Chen**

Faculty of Education, The University of Hong Kong, Hong Kong, China

Correspondence
Gaowei Chen, Faculty of Education, The University of Hong Kong, Room 513, Meng Wah Complex, Pokfulam Road, Hong Kong, China.
Email: gwchen@hku.hk

Funding information
Hong Kong Research Grants Council, University Grants Committee, Grant/Award Number: 17605221

Abstract

Classroom dialogue is crucial for effective teaching and learning, prompting many professional development (PD) programs to focus on dialogic pedagogy. Traditionally, these programs rely on manual analysis of classroom practices, which limits timely feedback to teachers. To address this, artificial intelligence (AI) has been employed for rapid dialogue analysis. However, practical applications of AI models remain limited, often prioritising state-of-the-art performance over educational impact. This study explores whether higher accuracy in AI models correlates with better educational outcomes. We evaluated the performance of two language models—BERT and Llama3—in dialogic analysis and assessed the impact of their performance differences on teachers' learning within a PD program. By fine-tuning BERT and engineering prompts for Llama3, we found that BERT exhibited substantially higher accuracy in analysing dialogic moves. Sixty preservice teachers were randomly assigned to either the BERT or Llama3 group, both participating in a workshop on the academically productive talk (APT) framework. The BERT group utilized the fine-tuned BERT model to facilitate their learning, while the Llama3 group employed the Llama3 model. Statistical analysis showed significant improvements in both groups' knowledge and motivation to learn the APT framework, with high levels of satisfaction reported. Notably, no significant differences were found between the two groups in posttest knowledge, motivation, and satisfaction. Interviews further elucidated how both models facilitated teachers' learning of the APT framework. This study validates the use of AI in teacher training and is among the first to investigate the relationship between AI accuracy and educational outcomes.

KEYWORDS

APT, artificial intelligence, classroom dialogue, large language model, teacher learning

Practitioner notes

What is already known about this topic

- Given the significance of classroom dialogue, many teacher professional development programmes have been implemented focusing on dialogic pedagogy.
- To provide timely feedback to teachers, artificial intelligence (AI) techniques are increasingly utilised to investigate classroom dialogue. However, a small proportion of studies have investigated the impacts of AI models in practice, with a predominant focus on pursuing state-of-the-art performance.
- It is unclear whether more accurate AI models necessarily lead to more positive educational outcomes.

What this paper adds

- This study evaluated the performance of two AI-powered language models, BERT and Llama3, in dialogic move analysis through fine-tuning and prompt engineering. BERT exhibited significantly higher accuracy than Llama3.
- Through an experimental study, this paper revealed that teachers using either the more accurate BERT model or the less accurate Llama3 model showed substantial improvements in their knowledge and motivation to learn the APT framework and reported high levels of satisfaction.
- The performance difference between BERT and Llama3 did not cause significant differences in teachers' knowledge, learning motivation, and satisfaction during the learning of the APT framework.

Implications for Practice and/or Policy

- Deep learning models and large language models can be integrated into professional development programs to effectively facilitate teachers' learning of dialogic pedagogy.
- AI models with moderate performance can also produce impressive outcomes and provide a satisfactory experience. In some scenarios, the manner in which teachers collaborate with AI may be more pivotal than the AI's accuracy.

INTRODUCTION

Sociocultural theory indicates that learning occurs through social interaction (Vygotsky, 1978). In classroom settings, this interaction predominantly manifests as dialogue between teachers and students, as well as among students themselves, where one party addresses another and receives a reply (Howe & Abedin, 2013; Wang, Tao, & Chen, 2024). Consequently, learning is intrinsically linked to classroom dialogue, as substantiated by numerous theoretical and empirical studies (Howe et al., 2019; Resnick et al., 2015; Tao & Chen, 2024). Researchers and educators have thus explored instructional methods to orchestrate classroom dialogue to enhance students' knowledge acquisition, knowledge construction, and cognitive development (Kim & Wilkinson, 2019). Various pedagogies emphasizing classroom dialogue have been designed and developed, such as academically productive talk (APT; Michaels et al., 2010) and exploratory talk (Mercer & Littleton, 2007). However, it has been observed that teachers encounter difficulties in integrating these dialogic skills into practical teaching, with classroom dialogue often dominated by their monologic or authoritative talk (O'Connor & Snow, 2017; Thompson et al., 2024; Yang & Wang, 2022). Researchers suggest that this phenomenon may be attributed to teachers' insufficient knowledge and skills in dialogic pedagogies, as well as a lack of timely feedback to facilitate reflection (Herbel-Eisenmann et al., 2013; Jacobs et al., 2022; Resnick, Asterhan, Clarke, & Schantz, 2018).

To address the two issues, researchers and educators have developed a variety of professional development (PD) programs focused on dialogic pedagogies (e.g., Chen, Zhang, et al., 2020; Hennessy et al., 2018; Vasalampi et al., 2021). Research staff typically collect authentic teaching recordings, extract and transcribe key segments or entire classes, code dialogic practices, interpret the findings, design content and tasks, and integrate these elements into a comprehensive set of activities (Jacobs et al., 2022; Wang & Chen, 2024). Then, research facilitators organise workshops, engage teachers in active learning, analysis, and reflection, and provide guidance as needed (Borko et al., 2010; van der Linden et al., 2022). Several challenges have been identified during PD programs. First, coding teachers' dialogic practices is both time-consuming and labour-intensive, often requiring researchers to employ qualitative methods for manual annotation, as well as addressing inconsistencies (Hennessy et al., 2020). Second, providing timely feedback becomes challenging when PD programs involve a large number of teachers. Third, busy teachers often find it overwhelming to analyse their dialogic practices and reflect daily without external assistance (Resnick, Asterhan, Clarke, & Schantz, 2018).

With the advancement of technology, researchers increasingly turn to artificial intelligence (AI) for rapid and precise dialogue analysis. A systematic review by Wang, Tao, and Chen (2024) indicates that traditional machine learning and deep learning algorithms have been frequently used to model student-related interaction, teacher-related instruction, and whole-class discussion. In the past two years, there has been growing interest in employing large language models (LLMs) to detect features in classroom dialogue (e.g. Moreau-Pernet et al., 2024; Tran et al., 2024). Researchers in this domain tend to focus on comparing the performance of different AI models in analysing classroom discussions and aim to develop models with state-of-the-art performance. This tendency raises two critical issues for discussion. Firstly, despite the improved performance, the practical application of these AI models remains underexplored. Only a small proportion of studies have investigated the impacts of AI models and systems on classroom interaction (Wang, Tao, & Chen, 2024). Secondly, it is unclear whether more accurate AI models necessarily lead to more positive educational outcomes. Specifically, does the performance difference between two AI models result in a significant difference in their educational effectiveness?

To address these issues within the context of PD in dialogic pedagogies and AI in classroom dialogue, this study compares the performance of two well-known AI-based language models, Bidirectional Encoder Representations from Transformers (BERT) and Large Language Model Meta AI (Llama, specifically Llama3), in analysing classroom dialogue. The objective is to determine whether there is a performance difference in this task between these models. Following this comparison, the study then integrates these models into an exploratory PD program and evaluates whether the two models with different performances will have different effects on teachers' learning of a dialogic pedagogy, addressing both cognitive and non-cognitive aspects. The research questions (RQs) guiding this study are as follows:

RQ1. How do BERT and Llama3 differ in terms of accuracy and reliability in identifying dialogic moves in classroom dialogue

RQ2. Is there any difference in teachers' knowledge of dialogic pedagogy between the group using BERT and the group using Llama3 after the PD workshop? If so, how?

RQ3. Is there any difference in teachers' motivation to learn dialogic pedagogy between the group using BERT and the group using Llama3 after the PD workshop? If so, how?

RQ4. Is there any difference in teachers' satisfaction with the PD workshop between the group using BERT and the group using Llama3? If so, how?

RQ5. How do teachers perceive the utilization of BERT and Llama3 during their learning of dialogic pedagogy?

Specifically, RQ1 serves as the foundation of our study. RQ2 examines the cognitive aspects by investigating the impact of the two AI models on teachers' knowledge. RQs 3, 4 and 5 address the non-cognitive aspects, focusing on teachers' learning motivation, satisfaction and perception. This dual focus allows for a comprehensive evaluation of the use of the two AI models in our context.

LITERATURE REVIEW

Classroom dialogue and dialogic pedagogy

In dialogue, each utterance not only responds to previous statements but also anticipates future interactions, taking into account both the speaker's and the listener's positions (Bakhtin, 1981). Thus, the meaning of dialogue is co-constructed and negotiated between individuals (Tao & Chen, 2023). In classroom settings, a series of conversations occur between teachers and students or among students themselves, facilitating meaning-making, knowledge acquisition and cognitive practice. However, Alexander's (2001) investigation of classroom discourse across five educational systems highlights the way in which talk is utilized greatly affects students' responses and, consequently, the quality of teaching and learning. For instance, a teacher's question might elicit simple, brief answers in some classrooms, while in others, it might provoke more elaborate responses and extended student–student interactions (Kim & Wilkinson, 2019). Recognizing the critical role of classroom dialogue, researchers and educators have undertaken extensive investigations.

Based on a series of reviews by scholars (e.g., Bae et al., 2021; Howe & Abedin, 2013; Major et al., 2018; Mercer & Dawes, 2014; Rapanta & Felton, 2022; Song et al., 2019; Tao & Chen, 2023, 2024), existing research on classroom dialogue can be broadly categorized into five main areas: features of classroom dialogue, methods for evaluating classroom dialogue, factors influencing classroom dialogue, dialogic pedagogies for enhancing classroom dialogue, and the relationship between classroom dialogue and learning outcomes. In terms of features, classroom dialogue typically adheres to an initiation-response-evaluation/follow-up (IRE/F) structure, where teachers predominantly control the discourse through authoritative recitations or monologic lectures and restrict students' opportunities to elaborate on their ideas (O'Connor & Snow, 2017; Thompson et al., 2024; Yang & Wang, 2022). Regarding evaluation methods, classroom dialogue is assessed through both semi-automatic and manual efforts, utilizing coding schemes, ethnography and sociolinguistic conventions for quantitative and qualitative analyses (Wang, Tao, & Chen, 2024). Concerning influencing factors, variations in classroom dialogue have been attributed to teachers' attributes (e.g., experience), students' attributes (e.g., gender), interpersonal factors (e.g., group dynamics), and environmental factors (e.g., class size) (Bae et al., 2021; Howe & Abedin, 2013).

A variety of dialogic instructional approaches have also been developed to enhance the quality of classroom dialogue, including dialogic teaching (Alexander, 2017), APT (Michaels et al., 2010), and dialogic space (Wegerif, 2007). Despite differing terminologies, Kim and Wilkinson (2019) and Cui and Teo (2021) have summarised that most of these approaches adhere to five key principles: collective, reciprocal, cumulative, supportive and purposeful. Moreover, these dialogic pedagogies advocate for a set of dialogic skills (e.g., dialogic teaching moves) that enable teachers to facilitate student discussions and foster meaningful dialogue (Tao & Chen, 2024). Empirical evidence suggests that the implementation of these dialogic pedagogies significantly enhances students' reasoning skills and learning outcomes, with benefits that are long-lasting and transferable to other subjects (Resnick, Asterhan, Clarke, & Schantz, 2018). Thus, mastering a dialogic pedagogy is crucial for teachers aiming to improve the quality of classroom dialogue.

Teacher professional development in dialogic pedagogy

Teacher professional development (TPD) programs encompass a broad range of training and developmental activities, such as workshops and courses, designed to enhance teachers' knowledge, skills, and expertise (Huang et al., 2022). The ultimate goal of these programs is to improve teaching practices and educational outcomes (Sheridan et al., 2009). In light of the critical importance of classroom dialogue and the advantages of dialogic pedagogies, various TPD programs have been developed to enhance teachers' dialogic knowledge and skills (e.g., Chen, Zhang, et al., 2020; Hennessy et al., 2018; O'Connor & Michaels, 2019; Vasalampi et al., 2021).

Typically, teachers' participation in these programs involves three main components: theoretical learning, analysis of dialogic recordings, and the implementation of acquired knowledge. For instance, Sedova et al. (2016) recruited eight experienced teachers from five secondary schools in the Czech Republic for a TPD program focused on dialogic teaching. These teachers were introduced to theories and methods for orchestrating classroom dialogue and were given opportunities to analyse their classroom teaching. Postprogram evaluations showed a significant increase in students' reasoning talk, attributed to changes in teachers' dialogic behaviours (Sedova et al., 2016). Similarly, Gröschner et al. (2018) engaged six ninth-grade teachers from Germany in a program on purposeful classroom discourse. These teachers designed lesson plans, implemented them, recorded their classroom sessions and reflected on their dialogic practices through video analysis. The program resulted in positive changes in both teachers' self-efficacy and dialogic practices (Gröschner et al., 2018). This video-based reflective approach is also utilised in other TPD programs (e.g., Gröschner et al., 2014, 2015).

Facilitating teachers' learning of dialogic pedagogy requires researchers to undertake several tasks: selecting and transcribing dialogic clips, coding specific dialogic features, summarizing findings, organizing content into a series of activities, and providing guidance to teachers (Borko et al., 2010; Jacobs et al., 2022; Major & Watson, 2018; van der Linden et al., 2022). The substantial manual effort involved has led researchers to develop (semi-) automated tools to facilitate TPD. For example, Chen (2020) designed the Classroom Discourse Analyzer (CDA) to visualize specific dialogic features in classroom videos, such as speaker names and the number of words. This tool helps teachers understand students' dialogic participation. The CDA-supported TPD program has been shown to improve teachers' self-efficacy, beliefs in classroom dialogue, and productive talk (Chen, 2020; Chen, Chan, et al., 2020). However, high-level dialogic features, such as dialogic teaching moves, still require manual labeling. To address this, Jacobs et al. (2022, 2024) developed the TalkMoves application, incorporating automatic analysis of talk moves in dialogue. This tool positively influenced teachers' perceptions and increased their use of talk moves in classrooms (Jacobs et al., 2024). These developments suggest promising directions for future programs to use automated tools to facilitate teacher learning of dialogic pedagogy.

In fact, researchers have increasingly explored incorporating automated tools, such as AI, to facilitate teachers' learning in other areas, as evidenced by several reviews (e.g., Dogan et al., 2025; Mintii & Semerikov, 2024). By contrast, AI-facilitated TPD programs specifically focused on dialogic pedagogy remain limited.

AI in classroom dialogue

Considering the high costs associated with manually analysing classroom dialogue, researchers have investigated the use of AI for automatic analysis to provide teachers with actionable insights into their dialogic practices. According to a systematic review by Wang, Tao, and Chen (2024), from 2012 to 2022, various AI models have been developed to examine multiple dimensions of classroom dialogue using traditional machine learning and deep learning algorithms. Traditional machine learning algorithms, such as Bayesian networks, hidden Markov models, and support vector machines, have been employed to analyse teachers' discourse features, students' dialogic behaviour, and classroom organization. For instance, given the significance of questions in prompting heated discussions, researchers have used these algorithms to detect the presence of questions in teachers' utterances, as well as their types and proportions (Donnelly et al., 2016; Kelly et al., 2018; Stone et al., 2019). Additionally, to assist teachers in orchestrating classroom activities, researchers have also utilized corresponding algorithms to automatically detect lecture segments, group work, and whole-class conversations (Donnelly et al., 2016).

Given the limited performance of traditional machine learning algorithms, researchers have increasingly turned to deep learning algorithms for more accurate analysis. Since 2015, various deep neural networks, such as recurrent neural networks and Transformer-based networks, have been employed to investigate teachers' uptake and questions (Demszky et al., 2021; Huang et al., 2020), students' emotions and knowledge graphs (Zhen et al., 2021; Zheng et al., 2022), and the speaking roles and semantic content of discussions (de Araujo et al., 2023; Li et al., 2020; Song et al., 2021). Despite their superior performance, deep learning models are often criticized for their opaque decision-making processes. Consequently, researchers have explored the use of explainable AI to demystify classroom dialogue analysis, thereby providing both dialogue analysis and explanations to improve user trust (Wang, Bian, & Chen, 2024).

In the past two years, large language models (LLMs) have been utilized to analyse classroom dialogue due to their exceptional ability to comprehend natural language. For instance, Kupor et al. (2023), Moreau-Pernet et al. (2024), and Wang et al. (2023) employed GPT-3 and GPT-3.5-turbo to identify dialogic moves in teachers' and students' utterances during classroom interactions. Whitehill and LoCasale-Crouch (2023) investigated the use of Meta's Llama2 to classify the presence of instructional support within teachers' utterances, while Tran et al. (2024) utilized the open-source models Mistral and Vicuna to assess instructional quality and overall discussion quality by analysing specific features in classroom transcripts. Additionally, Wang and Demszky (2023) leveraged GPT-3.5-turbo to explore its potential in coaching teachers on their instructional practices. These studies typically employed prompt engineering techniques as well as fine-tuning methods to optimize the performance of large language models. Prompt engineering involves strategically designing and optimizing task-specific prompts to enable LLMs to generate outputs without altering their parameters (Sahoo et al., 2024). While prompt engineering methods may not achieve the same level of performance as fine-tuned models in classroom dialogue analysis, they require only suitable prompts, whereas fine-tuning models demand substantial training costs.

Despite significant efforts to enhance the accuracy of automated classroom dialogue analysis, the low success rate of these models in practice raises a critical question: Will more accurate models yield more beneficial educational outcomes? As previously mentioned, it remains uncertain whether the performance differences between two AI models translate into significant variations in their educational effectiveness.

METHOD

This study aims to investigate the computational performance of two language models, Google's BERT and Meta's Llama3, in analysing dialogic moves within classroom dialogue. These models were selected as representative examples of deep neural networks and LLMs, respectively. Traditional machine learning models were excluded because they require manual selection of linguistic features, which does not align with our goal of fully automated analysis. We apply these models in an exploratory PD workshop to evaluate whether their performance differences will impact teachers' knowledge, learning motivation, satisfaction and perceptions during their learning of the APT framework.

BERT and Llama3 for dialogic move analysis

As an effective dialogic pedagogical approach for guiding and analysing classroom interaction, the APT framework emphasizes that teachers should encourage students to speak, think, share, and co-construct knowledge, both independently and through peer interactions (O'Connor & Michaels, 2019). To achieve this, teachers' talk should recognize all students as valuable contributors to collective understanding, make students' ideas and thinking public to identify errors and refine reasoning abilities, and require students to ground their claims in disciplinary knowledge (Resnick, Asterhan, & Clarke, 2018). Dialogic moves are essential skills that help teachers facilitate discussions, promoting equitable student participation in an environment where their thoughts are explicit and accessible to all.

Specifically, dialogic moves are defined as “utterance-sized units of talk, intended (as a ‘move’ in a game) to get the other player(s) to respond in some way, to bring something particular to the table” (O'Connor & Michaels, 2019, p. 168). In classroom teaching, these moves are dialogic acts designed to elicit replies from students or teachers. Empirical evidence has shown that these dialogic moves significantly enhance students' learning outcomes (Howe et al., 2019; Tao & Chen, 2024). Therefore, we chose dialogic moves as the primary focus for teacher training and the modelling unit for BERT and Llama3 to facilitate automatic analysis.

Dailogic move dataset

The public dialogic move corpus, *TalkMoves*, was selected for the fine-tuning of BERT and the prompt engineering of Llama3. This corpus comprises 567 authentic transcripts from K-12 mathematics lessons, containing 174,186 teacher utterances and 59,874 student utterances (Suresh et al., 2022). Each utterance is manually annotated with a dialogic move based on the APT framework, encompassing seven dialogic move types for teachers and five for students, as detailed in Table A1 in the appendix. Following procedures in previous studies (Suresh et al., 2022; Wang & Chen, 2024), we randomly divided the full dataset into 90% for training and 10% for testing.

BERT: Fine-tuning

BERT is a pretrained language representation model that can be fine-tuned with a single additional output layer to perform a wide range of tasks (Devlin et al., 2019). Before the advent of LLMs, BERT was regarded as one of the foundational and advanced models. In the context of classroom dialogue analysis, BERT is the most widely employed deep learning model with exceptional performance over the past decade (Wang, Tao, & Chen, 2024). Researchers have explored applying it in practical settings to facilitate learning and reported positive effects (Zheng et al., 2023). Additionally, in comparison to more recent deep learning models prior to LLMs, BERT has had a larger impact on academia, as evidenced by its significant citation count. Consequently, we selected and fine-tuned BERT to automatically identify dialogic moves among teachers' and students' utterances.

For analysing teachers' dialogic moves, we conducted a seven-way classification task (i.e., six dialogic moves and “none”). Given the importance of dialogue context, the input consisted of a teacher's utterance concatenated with the preceding student's utterance. The output was the predicted probability of each dialogic move. For analysing students' dialogic moves, we conducted a five-way classification task (i.e., four dialogic moves and “none”). The input was a student's utterance concatenated with the preceding sentence, and the output was the predicted probability of each dialogic move.

During the fine-tuning process, we utilized the AdamW optimizer and set the number of epochs to 14. The initial batch size and learning rate were 32 and 1e-5, respectively. Additionally, we set warmup steps to 0, 100, and 1000. To prevent overfitting, we adopted early stopping strategies. The code was implemented in Python 3.11 using the PyTorch and HuggingFace libraries, and the training was conducted on an RTX 4090 GPU.

Llama3: Prompt engineering

Llama3, an open-source LLM developed by Meta, was considered the most powerful among open-source LLMs when we conducted this study. Given its exceptional performance and low cost compared to commercial LLMs such as GPT-4, we selected Llama3-8B-Instruct to analyse dialogic moves. Fully training Llama3 requires an estimated 1.3 million hours of computation on H100-80GB hardware with a thermal design power of 700W (Meta-llama, 2024), which is prohibitively expensive for small laboratories like ours. Thus, we opted for prompt engineering techniques with Llama3.

Specifically, this study employed three prompt engineering techniques for dialogic move analysis: zero-shot, few-shot and few-shot chain-of-thought (CoT) prompting. Zero-shot prompting involves providing a direct task description without examples, while few-shot prompting combines the task description with a few input–output examples (Brown et al., 2020). CoT prompting facilitates LLMs in conducting a coherent, step-by-step reasoning process when addressing a question, typically involving "augmenting each exemplar in few-shot prompting with a chain of thought for an associated answer" (Wei et al., 2022, p. 24826).

Building on prompts from previous related studies (e.g., Eager & Brunton, 2023; Kojima et al., 2022; Wang et al., 2023; Wang & Demszky, 2023), we designed zero-shot prompts to include the instruction, a description of dialogic moves, and the utterances to be classified. Few-shot prompts were created by adding well-annotated utterances as examples to the zero-shot prompts. Few-shot CoT prompts incorporated the reasoning process explaining why the example utterances were classified under specific dialogic moves. Examples of these three types of prompts used in our study are illustrated in Figure 1. When using Llama3, we set the temperature parameter to 0 to ensure more deterministic answers. Notably, given the impracticality of enumerating all potential prompts, we selected three representative prompts for inclusion in the study.

Participants

To evaluate the impact of the performance difference between BERT and Llama3 on teachers' learning, we employed a convenience sampling method to recruit preservice teachers from a teacher education-focused university in Beijing. The participants would be invited to attend a PD workshop on a dialogic pedagogy applicable across various subjects (i.e., the APT framework). Participation was thus open to all preservice teachers interested in leveraging language models and learning dialogic pedagogy. The focus on preservice teachers was driven by the exploratory nature of this study, aiming to validate the effectiveness of language model-supported PD before its implementation among in-service teachers. A total of 60 preservice teachers registered and were randomly assigned to either the Llama3 group ($n=30$) or the BERT group ($n=30$). The Llama3 group consisted of 11 undergraduate and 19 graduate preservice teachers, while the BERT group included 16 undergraduate and 14 graduate preservice teachers. Each group had two males and 28 females. Their academic majors spanned both language and STEM disciplines, such as Chinese, English, mathematics, chemistry, biology, physics and computer science. Notably, there were no statistically significant differences between the two groups in terms of demographic variables including age, gender and education level ($p>0.05$; see Table 1).

We confirm that ethical approval for this study was granted by the Human Research Ethics Committee of our university. All participants provided informed consent and were informed of their right to withdraw from the study at any time without any negative consequences. Participants were also assured that their data would be anonymized and used solely for research purposes.

Experiment procedure

Prior to the experiment, participants were instructed to bring their own laptops and to download the Tencent Meeting software. Participants were advised to join the Tencent Meeting room for better visibility of the slides during the experiment. The experiment comprised a 3-hour PD workshop on the APT framework, structured into five phases as illustrated in Figure 2. In Phase 1, participants completed a pretest assessing their prior knowledge of the APT framework and a questionnaire evaluating their motivation to learn this dialogic pedagogy either by clicking the links provided in the chat box of the meeting room or by scanning a QR code using their smartphones. This phase lasted approximately 10–15 minutes.

Classify the following teacher utterance into keeping everyone together, getting students to relate to another's idea, restating, pressing for accuracy, revoicing, pressing for reasoning, or none.

1. Keeping everyone together refers to prompting students to be active listeners and orienting students to each other; 2. Getting students to relate to another's ideas refers to prompting students to react to what a classmate said; 3. Restating refers to repeating all or part of what a student said word for word; 4. Pressing for accuracy refers to prompting students to make a mathematical contribution or use mathematical language;5. Revoicing refers to repeating what a student said but adding on or changing the wording; 6. Pressing for reasoning refers to Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations. 7. None refers to that the utterance does not belong to any of the previous six talk moves.

Prior student utterance: "Well the area is what's inside and the perimeter is what's outside."
Teacher utterance: "Okay so the area is what's inside and the perimeter is what's outside"
Teacher talk move:

a) Zero-shot prompt

Classify the following teacher utterance into keeping everyone together, getting students to relate to another's idea, restating, pressing for accuracy, revoicing, pressing for reasoning, or none.

1. Keeping everyone together refers to prompting students to be active listeners and orienting students to each other; 2. Getting students to relate to another's ideas refers to prompting students to react to what a classmate said; 3. Restating refers to repeating all or part of what a student said word for word; 4. Pressing for accuracy refers to prompting students to make a mathematical contribution or use mathematical language;5. Revoicing refers to repeating what a student said but adding on or changing the wording; 6. Pressing for reasoning refers to Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations. 7. None refers to that the utterance does not belong to any of the previous six talk moves.

Prior student utterance: "X minus Y equals 10."
Teacher utterance: "Boys and girls, tell me what did Eliza just say her equation was?"
Teacher talk move: Keeping everyone together

Prior student utterance: "7/10."
Teacher utterance: "Joe, do you agree with Juan that the answer is 7/10?"
Teacher talk move: Getting students to relate to another's ideas

Prior student utterance: "I add two in this place."
Teacher utterance: "Add two here."
Teacher talk move: Restating

Teacher utterance: "Can you give an example of an ordered pair?"
Teacher talk move: Pressing for accuracy

Prior student utterance: "I add two in this place."
Teacher utterance: "Julia told us she would add two here."
Teacher talk move: Revoicing

Teacher utterance: "Why could I argue that the slope should be increasing?"
Teacher talk move: Pressing for reasoning

Teacher utterance: "Okay."
Teacher talk move: None

Prior student utterance: "Well the area is what's inside and the perimeter is what's outside."
Teacher utterance: "Okay so the area is what's inside and the perimeter is what's outside"
Teacher talk move:

b) Few-shot prompt

Classify the following teacher utterance into keeping everyone together, getting students to relate to another's idea, restating, pressing for accuracy, revoicing, pressing for reasoning, or none.

1. Keeping everyone together refers to prompting students to be active listeners and orienting students to each other; 2. Getting students to relate to another's ideas refers to prompting students to react to what a classmate said; 3. Restating refers to repeating all or part of what a student said word for word; 4. Pressing for accuracy refers to prompting students to make a mathematical contribution or use mathematical language;5. Revoicing refers to repeating what a student said but adding on or changing the wording; 6. Pressing for reasoning refers to Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations. 7. None refers to that the utterance does not belong to any of the previous six talk moves.

Prior student utterance: "X minus Y equals 10."
Teacher utterance: "Boys and girls, tell me what did Eliza just say her equation was?"
Chain of thoughts: Let's think step by step. The teacher is trying to ask other students to repeat Eliza's utterance, which aligns with the definition of keeping everyone together that tries to orient students to each other.
Teacher talk move: Keeping everyone together

Prior student utterance: "7/10."
Teacher utterance: "Joe, do you agree with Juan that the answer is 7/10?"
Chain of thoughts: Let's think step by step. The teacher is trying to ask whether Joe agrees with Juan's answer, which aligns with the definition of getting students to relate to another's idea that tries to prompt students to react to what a classmate said.
Teacher talk move: Getting students to relate to another's ideas

Prior student utterance: "I add two in this place."
Teacher utterance: "Add two here."
Chain of thoughts: Let's think step by step. The student said she/he added two in a place. The teacher repeated a part of the student's utterance, which aligns with the definition of restating that tries to repeat all or part of what a student said.
Teacher talk move: Restating

Teacher utterance: "Can you give an example of an ordered pair?"
Chain of thoughts: Let's think step by step. The teacher wants to let students or a student provide an example of an ordered pair. The ordered pair is a mathematical concept. Thus, the teacher's utterance aligns with the definition of pressing for accuracy that tries to prompt students to make a mathematical contribution or use mathematical language.
Teacher talk move: Pressing for accuracy

Prior student utterance: "I add two in this place."
Teacher utterance: "Julia told us she would add two here."
Chain of thoughts: Let's think step by step. The student said she/he added two in a place. The teacher repeated a part of the student's utterance but also added some new words, which aligns with the definition of revoicing that tries to repeat what a student said by adding on or changing the wording.
Teacher talk move: Revoicing

Teacher utterance: "Why could I argue that the slope should be increasing?"
Chain of thoughts: Let's think step by step. The teacher asked students 'why' aiming to let students explain their reasoning, which aligns with the definition of pressing for reasoning that tries to prompt students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations.
Teacher talk move: Pressing for reasoning

Teacher utterance: "Okay."
Chain of thoughts: Let's think step by step. The teacher just said yes, which does not align with any of six talk moves.
Teacher talk move: None

Prior student utterance: "Well the area is what's inside and the perimeter is what's outside."
Teacher utterance: "Okay so the area is what's inside and the perimeter is what's outside."
Chain of thoughts: Let's think step by step.
Teacher talk move:

c) Few-shot CoT prompt

FIGURE 1 Examples of zero-shot (a), few-shot (b) and few-shot CoT (c) prompts for Llama3 to analyse teacher dialogic moves.

TABLE 1 Demographic description of participants.

	Llama3 group ($n = 30$)				BERT group ($n = 30$)						
	Mean	***SD***	**Min**	**Max**	***Mean***	***SD***	**Min**	**Max**	t/χ^2	***df***	***p***
Age	22.367	2.341	18	32	22.200	2.058	18	27	$t = 0.293$	58	0.771
Gender[a]	1.933	0.254	1	2	1.933	0.254	1	2	$\chi^2 = 0.000$	1	1
Education level[b]	1.633	0.49	1	2	1.467	0.507	1	2	$\chi^2 = 1.684$	1	0.194

[a]Gender: 1, Male; 2, Female.

[b]Education level: 1, Undergraduate preservice teachers; 2, Graduate preservice teachers.

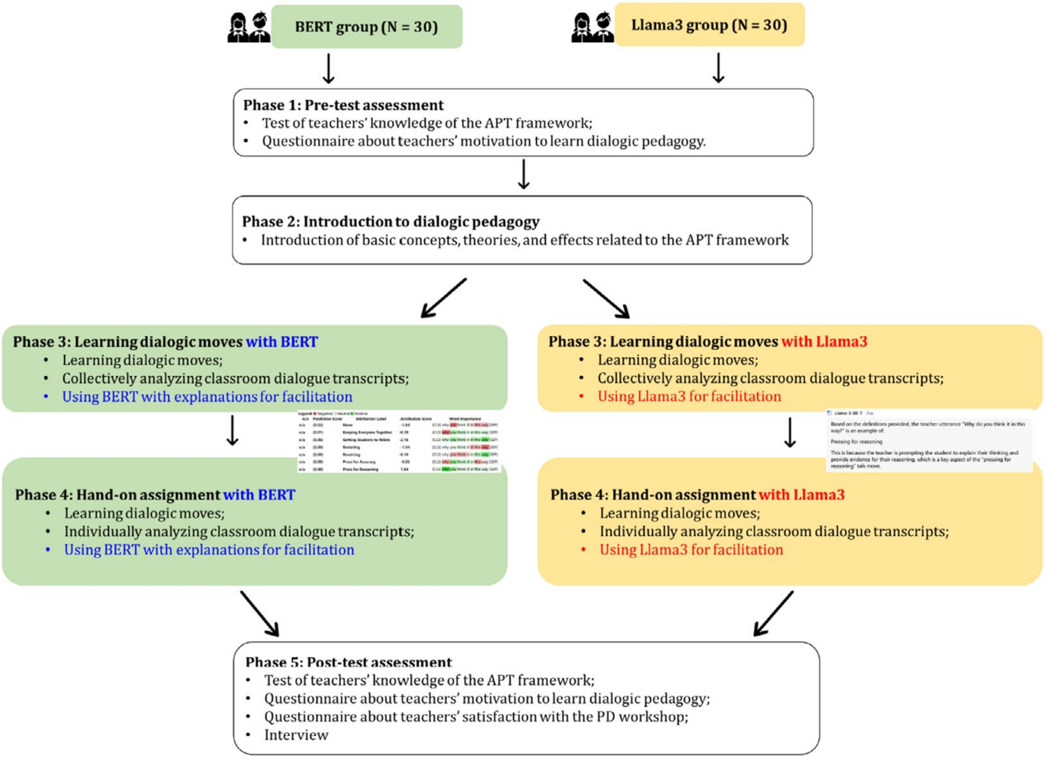

FIGURE 2 The procedures of the experiment.

Phase 2, following PD procedures in previous studies (e.g., Borko et al., 2021; Chen, Chan, et al., 2020; Osborne et al., 2019), involved introducing participants to dialogic pedagogy. This introduction covered the core principles of sociocultural theory, the definition and importance of classroom dialogue, prevailing classroom dialogue patterns, desired productive talk, key elements of the APT framework, and the effects of productive talk. For instance, we presented several authentic clips demonstrating the IRE/F pattern and prompted participants to critically evaluate its shortcomings and envision productive talk. This phase lasted about 45 minutes.

Phase 3 focused on teaching participants dialogic moves commonly used in classroom settings. We began by defining and providing examples of dialogic moves, followed by presenting authentic discourse segments with coded dialogic moves as illustrative samples. Considering that PD workshops are more effective when teachers engage as a group, incorporate subject knowledge, and practice learned concepts (Cordingley et al., 2015; Sims & Fletcher-Wood, 2021), participants were thus divided into subgroups of three to four individuals and tasked with collectively analysing dialogic moves within selected classroom dialogue clips. These clips were chosen to reflect real classroom contexts and the participants' subject majors, fostering an authentic learning environment. To assist in their analysis, we provided the Llama3 model for the Llama3 group and the BERT model for the BERT group. Before the group task, participants in the Llama3 group received instructions on writing prompts for Llama3 to analyse dialogic moves, with a recommendation to use few-shot CoT prompts. Figure A1 in the appendix presents a screenshot of Llama3's analysis of a teacher's utterance. Similarly, participants in the BERT group were given access to our fine-tuned BERT model, with explanations aligned to those provided by Llama3, following methods in Wang, Bian, and Chen (2024). Figure A2 displays a screenshot of BERT's analysis of the teacher's utterance. Utilizing Llama3 and BERT aimed to mitigate the issue of delayed facilitator feedback and to support learning in the face of inconsistencies and confusion. Phase 3 lasted approximately 70–80 minutes.

In Phase 4, participants continued learning dialogic moves through an individual hands-on assignment, which involved coding additional classroom transcript clips. They could seek assistance from Llama3 and BERT respectively when encountering difficulties. At the end of this phase, participants reflected on their learning, and a summary was provided. This phase lasted about 20–30 minutes.

Phase 5 required participants to complete a posttest on their knowledge of the APT framework and questionnaires assessing their motivation to learn dialogic pedagogy and their satisfaction with the PD workshop. This phase lasted approximately 20–30 minutes. To gain deeper insights into participants' perceptions of utilizing Llama3 and BERT for dialogic pedagogy learning, 10 participants from each group were selected for a 20-minute interview. The interview protocol focused on four primary questions: (1) How do you think of the performance of BERT/Llama3 in analysing dialogic moves? (2) Did BERT/Llama3 facilitate your understanding of the APT framework? Please elaborate. (3) Did BERT/Llama3 enhance your motivation to learn dialogic pedagogy? Please explain. (4) What aspects were you satisfied or dissatisfied with when you used BERT/Llama3 during this workshop? Please elaborate.

Notably, all the tests and questionnaires were made available to participants in both Chinese and English. They were collected on a platform similar to Google Forms (https://www.wjx.cn/). The interview data were collected by us using a digital voice recorder.

Measures

Two experts in the APT framework developed both the pretest and posttest to evaluate participants' knowledge of this dialogic pedagogy. The pretest included four questions focused on the definitions and theories of classroom dialogue and the APT framework, aimed at assessing participants' prior knowledge. The posttest consisted of four questions on the concept and core ideas of the APT framework, along with 12 questions requiring participants to analyse dialogic moves in carefully selected classroom dialogue transcript clips, designed to measure their theoretical mastery of the APT framework following the workshop. Both the pretest and posttest were scored out of a maximum of 100 points.

Participants' motivation to learn dialogic pedagogy was assessed using a widely recognized learning motivation questionnaire, initially developed to evaluate students' motivation in a natural science course (Hwang et al., 2013). This instrument has been thoroughly reviewed, adopted, and adapted by various researchers (e.g., Huang et al., 2023; Hwang et al., 2023; Woo et al., 2024) across diverse contexts to assess motivation for learning different subjects. In accordance with these adaptation procedures, we modified the questionnaire by replacing the course name with “dialogic pedagogy” to maintain content validity. This instrument comprised seven items and employed a six-point Likert scale for responses. It demonstrated high internal consistency, with Cronbach's alpha values of 0.840 in the pretest and 0.872 in the posttest.

Participants' satisfaction with the PD workshop was evaluated using a tailored questionnaire designed to measure teachers' satisfaction with PD programs (Fisher et al., 2010). This questionnaire included 14 items aimed at capturing participants' feelings during the workshop (e.g., engagement, enjoyment, and perceived usefulness). Responses were rated on a seven-point Likert scale. The instrument achieved a Cronbach's alpha value of 0.936, indicating a high level of reliability.

Notably, all tests and questionnaires utilized in this study are included in the appendix, as seen in Tables A2–A5.

Data analysis

To investigate the effects of BERT- or Llama3-supported PD workshops on teachers' pedagogical knowledge and learning motivation (RQ2 and RQ3), we employed independent samples *t*-tests, paired samples *t*-tests, and analysis of covariance (ANCOVA). First, independent samples *t*-tests were utilized to determine whether there were significant differences in teachers' initial knowledge and learning motivation between the two groups. Second, paired samples *t*-tests aimed to assess whether significant differences existed in teachers' knowledge and motivation before and after the workshops, thereby evaluating the effectiveness of BERT- and Llama3-supported PD workshops. Third, ANCOVA was conducted to examine whether significant differences in posttest scores for knowledge and motivation existed between the two groups, indicating which PD workshop was more effective. To examine teachers' satisfaction with the PD workshops (RQ4), an independent samples *t*-test was employed to compare any difference in satisfaction scores between the two groups. The gathered data notably met the underlying assumptions required for conducting *t*-tests and ANCOVA.

To explore teachers' perceptions of utilizing BERT and Llama3 for learning dialogic pedagogy (RQ5), a thematic analysis of the interview data was conducted. Following the steps recommended by Braun and Clarke (2012), we familiarised ourselves with the interview data by transcribing the audio recordings and reviewing the transcripts. The first author generated initial codes and iteratively developed these codes into themes. To enhance the reliability and validity of the analysis, a well-trained research assistant independently reviewed the coding scheme, and discrepancies were resolved through iterative discussions until a consensus was reached. The refined coding scheme was then applied to the interview data by both researchers, with all differences addressed through discussions.

Our data collection and analysis encompass both computational and educational aspects. The computational aspect involves comparing the accuracy of BERT and Llama3 in analysing dialogic moves. Their difference in accuracy forms the foundation of our study to examine whether higher accuracy in AI models correlates with a better educational impact. The educational aspect evaluates the effectiveness of the BERT-supported and Llama3-supported PD workshops from both quantitative and qualitative perspectives. The quantitative data include teachers' self-reported knowledge, learning motivation, and satisfaction, gathered through tests and questionnaires. These data aim to assess whether the BERT-supported and Llama3-supported PD workshops can enhance their knowledge of the APT framework, increase their motivation to learn, and provide satisfaction and to assess whether there is a significant difference in these variables between the two groups. The qualitative data, obtained from interviews, allow us to examine teachers' perceptions and experiences in the workshops, further triangulating the findings from the quantitative tests and questionnaires. Ultimately, the integration of both quantitative and qualitative findings enables us to assess the overall effectiveness of the BERT-supported and Llama3-supported PD workshops and determine whether higher accuracy in AI models correlates with a better educational impact on preservice teachers' learning of dialogic pedagogy.

RESULTS

Model accuracy

Table 2 presents the accuracy of BERT and Llama3 in identifying dialogic moves. Specifically, the fine-tuned BERT model achieved an accuracy of 0.869 for teachers' dialogic moves and 0.777 for students' dialogic moves. In contrast, Llama3's accuracy for teachers' dialogic moves was as follows: 0.393 in zero-shot prompts, 0.510 in few-shot prompts, and 0.573

in few-shot CoT prompts. For students' dialogic moves, Llama3 achieved an accuracy of 0.214 in zero-shot prompts, 0.494 in few-shot prompts, and 0.528 in few-shot CoT prompts. These results indicate that the fine-tuned BERT model outperforms Llama3 employing three prompt engineering techniques in dialogic move analysis in our study.

Knowledge

The independent samples *t*-test was conducted to compare teachers' knowledge of the APT framework in the pretest between the BERT and Llama3 groups. As shown in Table 3, the BERT group exhibited a mean score of 12.400 with a standard deviation (*SD*) of 12.926, while the Llama3 group showed a mean score of 10.200 with an *SD* of 10.193. The *t*-test result ($t=0.732$, $p=0.467$) indicates no significant difference in teachers' pretest knowledge between the two groups, suggesting comparable levels of prior knowledge.

The paired samples *t*-test was then conducted to compare teachers' knowledge of the APT framework before and after the workshop, as illustrated in Table 4. For the BERT group, teachers achieved a mean of 78.783 ($SD=15.254$). The *t*-test result ($t=19.502$, $p<0.001$) indicates a significant increase in teachers' knowledge in the posttest compared to the pretest, suggesting the effectiveness of the BERT-supported PD workshop in enhancing teachers' dialogic pedagogy knowledge. Similarly, for the Llama3 group, the results also indicate the effectiveness of the Llama3-supported PD workshop in improving teachers' dialogic pedagogy knowledge.

Subsequently, ANCOVA was utilized to compare teachers' posttest knowledge of the APT framework between the BERT and Llama3 groups, with pretest scores as the covariate and posttest scores as the dependent variable. As shown in Table 5, the BERT group had an adjusted mean score of 78.593 with a standard error (*SE*) of 2.533, while the Llama3 group had an adjusted mean score of 75.373 ($SE=2.533$). The ANCOVA result ($F=0.804$, $p=0.374$) does not reveal a significant difference in the posttest knowledge between the two groups. To further identify which phases contributed to specific aspects of teachers'

TABLE 2 The accuracy of BERT and Llama3 in analysing dialogic moves in the TalkMoves dataset.

Language models	Teachers' dialogic moves	Students' dialogic moves
BERT	0.869	0.777
Llama3 – zero-shot	0.393	0.214
Llama3 – few-shot	0.510	0.494
Llama3 – few-shot CoT	0.573	0.528

TABLE 3 The independent samples *t*-test comparing teachers' knowledge in the pretest.

Group	*n*	*Mean*	*SD*	*t*	*p*
BERT	30	12.400	12.926	0.732	0.467
Llama3	30	10.200	10.193		

TABLE 4 The paired samples *t*-test comparing teachers' knowledge before and after the workshop.

		Pretest	Posttest		
Group	*n*	*Mean* (*SD*)	*Mean* (*SD*)	*t*	*p*
BERT	30	12.400 (12.926)	78.783 (15.254)	19.502	<0.001
Llama3	30	10.200 (10.193)	75.183 (12.334)	24.298	<0.001

improved knowledge of the APT framework, we conducted additional analyses, which are detailed in the appendix. The results suggest that the increased knowledge of the APT framework is attributed to both the introduction in Phase 2 and the learning of dialogic moves with BERT and Llama3 support in Phases 3 and 4.

Motivation

The independent samples *t*-test was conducted to compare teachers' motivation to learn the APT framework in the pretest between the BERT and Llama3 groups. As shown in Table 6, the BERT group exhibited a mean score of 4.733 ($SD=0.589$), while the Llama3 group showed a mean score of 4.743 ($SD=0.557$). The *t*-test result ($t=0.064$, $p=0.949$) indicates no significant difference in teachers' pretest motivation between the two groups, suggesting similar learning motivation.

The paired samples *t*-test was then conducted to assess changes in teachers' learning motivation before and after the workshop, as detailed in Table 7. For the BERT group, teachers achieved a mean of 5.400 ($SD=0.482$). The *t*-test result ($t=5.793$, $p<0.001$) indicates a significant increase in teachers' learning motivation in the posttest compared to the pretest, suggesting the effectiveness of the BERT-supported PD workshop in enhancing teachers' motivation to learn dialogic pedagogy. Likewise, findings for the Llama3 group also indicate the effectiveness of the Llama3-supported PD workshop in improving teachers' motivation to learn dialogic pedagogy.

Subsequently, ANCOVA was utilized to compare teachers' posttest motivation to learn the APT framework between the BERT and Llama3 groups, with pretest scores as the covariate and posttest scores as the dependent variable. As delineated in Table 8, the BERT

TABLE 5 The ANCOVA comparing teachers' knowledge in the posttest.

Group	*n*	*Mean*	*SD*	Adjusted mean	Std.error	*F*	*p*
BERT	30	78.783	15.254	78.593	2.533	0.804	0.374
Llama3	30	75.183	12.334	75.373	2.533		

TABLE 6 The independent samples *t*-test comparing teachers' learning motivation in the pretest.

Group	*n*	*Mean*	*SD*	*t*	*p*
BERT	30	4.733	0.589	0.064	0.949
Llama3	30	4.743	0.557		

TABLE 7 The paired samples *t*-test comparing teachers' learning motivation before and after the workshop.

		Pretest	Posttest		
Group	*n*	*Mean* (*SD*)	*Mean* (*SD*)	*t*	*p*
BERT	30	4.733 (0.589)	5.400 (0.482)	5.793	<0.001
Llama3	30	4.743 (0.557)	5.276 (0.514)	4.542	<0.001

TABLE 8 The ANCOVA comparing teachers' learning motivation in the posttest.

Group	*n*	*Mean*	*SD*	Adjusted mean	Std.error	*F*	*p*
BERT	30	5.400	0.482	5.401	0.088	1.040	0.312
Llama3	30	5.276	0.514	5.275	0.088		

group had an adjusted mean score of 5.401 (SE = 0.088), while the Llama3 group had an adjusted mean score of 5.275 (SE = 0.088). The ANCOVA result (F = 1.040, p = 0.312) does not reveal a significant difference in learning motivation in the posttest between the two groups.

Satisfaction

Table 9 presents the results of the independent samples t-test comparing teachers' satisfaction with the workshops between the BERT and Llama3 groups. Notably, teachers in both groups reported a high level of satisfaction. Specifically, the BERT group reported a mean score of 5.910 (SD = 0.649) out of seven points, while the Llama3 group reported a mean of 5.895 (SD = 0.662). Although the BERT group achieved a slightly higher satisfaction score than the Llama3 group, the t-test result (t = 0.084, p = 0.933) indicates no significant difference in teachers' satisfaction levels between the two groups.

To further describe the satisfaction levels of the two groups, Figure 3 presents the participants' average scores on each item. Both groups reported high scores in the dimensions of engagement, enjoyment, and perceived usefulness. For instance, participants in both groups reported an average score higher than six out of seven points on enjoyment-related items (e.g., items 8, 12 and 14). Additionally, they gave high ratings on engagement-related items (e.g., items 2, 6 and 10).

Perceptions

We conducted a thematic analysis of interview data to further examine teachers' perceptions of using BERT and Llama3 for learning dialogic pedagogy. As shown in Table 10, two primary themes emerged: model features and the relationship between these features and teacher learning.

TABLE 9 The independent samples t-test comparing teachers' satisfaction with the workshops.

Group	*n*	*Mean*	*SD*	*t*	*p*
BERT	30	5.910	0.649	0.084	0.933
Llama3	30	5.895	0.662		

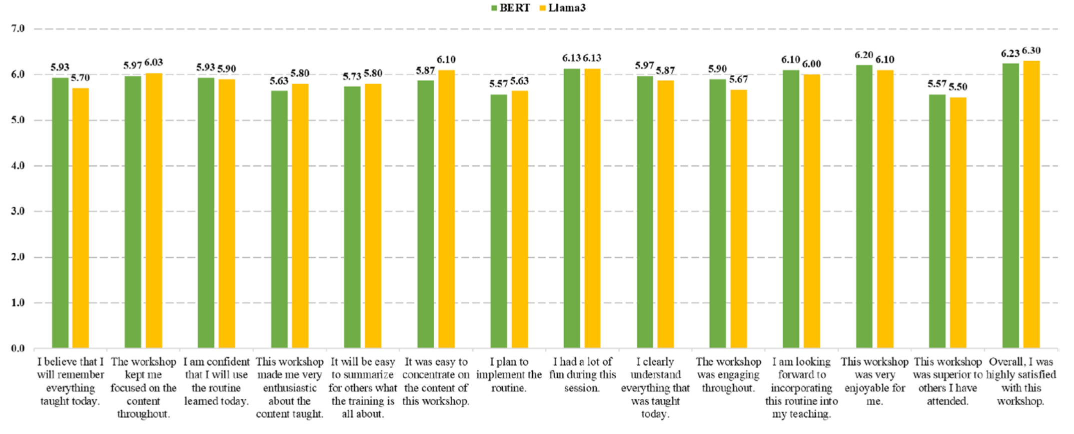

FIGURE 3 Teachers' average satisfaction scores on each item.

TABLE 10 Thematic analysis of the interview.

Theme	Sub-theme	Definition	Example
Model features	Efficiency	Teachers think BERT/Llama3 is fast to generate an analysis.	[*BERT group*] Because BERT is a machine, it's quicker at understanding human language than our brains. To put it in another words, it is more efficient and convenient.
			[*Llama3 group*] This large language model can give me answers in a snap, way faster than I can. It can even analyze a bunch of conversations all at once.
	Accuracy	Teachers think BERT/Llama3 is accurate/inaccurate in its answer.	[*BERT group*] After going through a series of exercises just now, I noticed that BERT was pretty precise most of the time. Although there were a couple of small inconsistencies sometimes, BERT always ranked our answer in its top 2 candidate answers!
			[*Llama3 group*] I think the Llama model mainly struggles with accuracy. There were many times when I felt its answers were wrong. For example, during a group discussion, we all agreed the answer should be "getting students to relate to another's idea", but the Llama model gave a completely different answer, and its explanation wasn't convincing.
	Explanation	How teachers perceive the explanations provided by BERT and Llama3	[*BERT group*] This model pointed out key words or phrases highlighted in different colours to explain its dialogic move analysis. The explanation was clear and quite different from my usual approach, where I tend to look at the context instead of diving into some word details. However, I think it could enhance the explanation by condensing some descriptive sentences into the figure.
			[*Llama3 group*] Llama provided a very textbook response. It presented all potential options, adhered closely to the definition in the prompt, and assessed which dialogic move best matched the intention behind the statement. It mirrored human reasoning to some extent. However, its explanation seemed crafted from a lengthy and wordy template, lacking a more concise feel.
The relationship between model features and teacher learning	Understanding	Teachers think BERT/Llama3 facilitated their understanding of dialogic pedagogy.	[*BERT group*] At first, I struggled to recall the precise definitions of various dialogic moves. So, I turned to BERT for assistance. By noting the emphasized keywords like "what" and "why", I grasped how to employ keywords to interpret the intentions behind teachers' and students' statements. By blending this approach with my own insights, I swiftly mastered the art of analysing dialogic moves. Moreover, I can utilize these keywords in my future teaching to prompt students' responses effectively.

TABLE 10 (Continued)

Theme	Sub-theme	Definition	Example
			[*Llama3 group*] When my response differs from Llama's, I instinctively trust my own answer, either by defending my perspective or challenging its viewpoints. This back-and-forth resembles a reflective exercise, where I contrast two distinct answers. Through this process, my comprehension of dialogic moves has significantly deepened.
	Motivation	Teachers think BERT/Llama3 made them more interested in learning dialogic pedagogy.	[*BERT group*] BERT, being an AI machine, has its unique way of thinking that intrigued me. I was curious to see how its thought process compared to mine when I completed an analysis. Whether our answers matched or not, I eagerly awaited BERT's response after each analysis. This anticipation enhanced my interest and overall engagement in the learning experience.
			[*Llama3 group*] The introduction of Llama into the workshop added a very interesting element. On one hand, it acts almost like a virtual person, given its current worldwide popularity. Its usage attracts our interest. On the other hand, Llama is not a person but a machine, strictly speaking. Personally, I found it fascinating to observe the distinctions between human logic and the reasoning of a sophisticated language model like Llama.
	Confidence	Teachers think BERT/Llama3 made them more confident in analysing classroom dialogue.	[*BERT group*] Thanks to BERT's assistance, I now feel well-equipped with a solid grasp of this pedagogy. I am confident that I will apply it in my future classes, enhancing the quality of my teaching. Knowing that BERT is always available to support me gives me reassurance. I can rely on it to analyse my classroom discussions and subsequently refine my teaching methods.
			[*Llama3 group*] As a preservice teacher, I initially felt apprehensive about engaging with students due to my inexperience in teaching. However, following this workshop, I discovered that having meaningful conversations with students was not as daunting as I had imagined, particularly after seeking guidance from these advanced language models. I utilized them not only for analysing dialogic moves but also for insights on productive communication. Their versatility instilled in me a sense of confidence.

TABLE 10 (Continued)

Theme	Sub-theme	Definition	Example
	Partnership	Teachers think BERT/Llama3 served as a learning partner.	[*BERT group*] As I was learning, it felt like I had a virtual learning buddy. I studied its thinking process, trying to figure out how I could match or even surpass its performance. While it excelled in dialogic move analysis, I saw BERT as a fellow runner on the same path, each of us pushing ourselves forward.
			[*Llama3 group*] I viewed Llama as a study companion with whom I could freely interact and seek clarification whenever questions arose. While I could also approach the authoritative facilitator or my group members for assistance, engaging with Llama eliminated any social or emotional barriers—I could inquire about anything without hesitation. In contrast, when communicating with humans, I had to be mindful of my language choices, which at times led to added stress.

Under model features, we identified three subthemes: efficiency, accuracy and explanation. Efficiency refers to the rapid analysis by both BERT and Llama3. Teachers from both groups reported that the models could perform analyses almost instantaneously, far quicker than human capability. Accuracy pertains to the correctness or imprecision of the models' responses. Several teachers in the BERT group praised BERT's precision, while a few teachers in the Llama3 group noted that Llama3's accuracy had room for improvement. Explanations denote how teachers evaluated the models' explanations. For example, a teacher from the BERT group appreciated the highlighted keywords and phrases in BERT's explanations, which offered a different mode of thinking compared to humans. Conversely, a teacher from the Llama3 group observed that Llama3's method of matching dialogue with definitions was akin to human cognitive processes. Teachers also suggested improvements for the explanations, such as incorporating more descriptive sentences in BERT's responses and shortening Llama3's lengthy sentences.

Regarding the relationship between model features and teacher learning, we identified four subthemes: understanding, motivation, confidence, and partnership. Understanding refers to how the models facilitated teachers' comprehension of the APT framework. For instance, one teacher in the BERT group struggled with analysing dialogic moves initially, and BERT provided timely assistance. Another teacher in the Llama3 group noted that discrepancies between her answers and Llama3's prompted reflective thinking, thereby deepening her understanding. Motivation indicates that the use of BERT and Llama3 heightened teachers' interest in learning. Teachers in both groups expressed curiosity about AI responses, eager to compare machine intelligence with human intelligence. Confidence reflects the models' role in enhancing teachers' self-efficacy in analysing classroom dialogue. A teacher from the BERT group mentioned applying her learnings to teaching and relying on BERT for professional feedback. Similarly, a teacher from the Llama3 group reported her confidence in conducting productive dialogue by learning through the versatile Llama model. Finally, partnership denotes BERT and Llama3 serving as learning partners. A teacher in the Llama3 group emphasized the absence of socio-emotional conflicts with Llama3, allowing her to ask questions freely. Likewise, a teacher in the BERT group described the model as a fellow traveller on the same educational journey.

DISCUSSION

Discussion of findings

Given the significance of classroom dialogue and the benefits of teachers' mastery of dialogic skills, many TPD programs on dialogic pedagogy have been implemented. Traditional programs rely on manual analysis of classroom practices and cannot provide timely feedback to teachers. Consequently, researchers have increasingly turned to AI for automatic analysis. Despite advancements in using AI, practical applications of these models remain limited, with a predominant focus on pursuing the accuracy of performance. To address these issues, we investigated two language models with varying performance levels, incorporated them into an exploratory TPD program on dialogic pedagogy, and evaluated the impact of their performance differences on teachers' learning.

Specifically, we utilized a pretrained language model named BERT and an open-source large language model called Llama3 to automatically analyse dialogic moves within classroom dialogue. By fine-tuning BERT and engineering zero-shot, few-shot, and few-shot CoT prompts for Llama3, we found that BERT's accuracy was notably higher than that of Llama3 in this task, consistent with previous studies (e.g., Moreau-Pernet et al., 2024; Wang & Demszky, 2023). To assess the effects of the performance differences between BERT and Llama3, 60 preservice teachers were randomly assigned to either the BERT group or the Llama3 group, both of which participated in a PD workshop on the APT framework. The BERT group utilized the fine-tuned BERT model to facilitate their learning, while the Llama3 group employed the Llama3 model. Both questionnaire and interview data were collected, yielding insightful findings.

First, teachers in both the BERT and Llama3 groups demonstrated significant improvement in their knowledge of the APT framework after the PD workshop. There was no significant difference in their posttest knowledge scores between the BERT and Llama3 groups. These results suggest that both BERT and Llama3 were effective in facilitating teachers' learning of dialogic pedagogy. However, their performance difference did not result in a significant difference in learning the APT framework. The interview findings further elucidate these results. Although teachers noticed the accuracy of BERT's answers and the imprecision of Llama3's responses, they reported that both models facilitated their understanding. Teachers indicated that BERT's accurate answers and explanations provided them with help when they were confused. Conversely, the discrepancies between Llama3's answers and their own thoughts prompted comparison and reflection, deepening their understanding. These findings align with He et al. (2023) that noted that humans determine their reliance level on AI based on their perceived accuracy. As indicated by Glikson and Woolley (2020) and Nazaretsky et al. (2022), teachers are more inclined to trust and accept decisions from more accurate AI models with reliable explanations, while they reject those with evident errors. During the workshop, BERT typically provided correct answers and intuitive explanations, thereby earning teachers' trust and acceptance. This phenomenon is also observed in other studies (Bansal et al., 2021). In contrast, Llama3 often offered incorrect analyses and lengthy explanations, leading to teachers' distrust and reflection, a behaviour known as diagnostic reasoning (Lambe et al., 2016), which subsequently deepened their understanding. In our study, it appears that the critical factor for learning is not the AI model's accuracy but how teachers collaborate with the AI—an argument also proposed in Tammets and Ley (2023). However, in other tasks, the relationship between AI accuracy, human-AI collaboration and educational outcomes still needs further exploration.

Second, teachers in both the BERT and Llama3 groups exhibited a significantly heightened motivation to learn dialogic pedagogy after the PD workshop. Notably, there was no significant difference in posttest motivation levels between the BERT and Llama3 groups. These findings suggest that incorporating either BERT or Llama3 into teacher training effectively enhanced learning motivation. This is consistent with prior research (e.g., Huang et al., 2023; Woo et al., 2024) demonstrating that the use of AI can promote participants' motivation to learn. Furthermore, the lack of a significant difference in motivational outcomes between the two AI models implies that the specific AI models used did not differentially impact teachers' learning motivation. Interviews with teachers provided further explanations for these findings. Teachers from both groups reported that their curiosity and interest in understanding the distinctions between human intelligence and AI reasoning motivated them to engage more actively in the learning process. This aligns with intrinsic and extrinsic motivation theories, which posit that curiosity and fantasy enhance intrinsic motivation (Malone & Lepper, 2021; Ryan & Deci, 2000). The innovative nature of BERT and Llama3, as leading AI models, naturally evokes curiosity and anticipation, thereby increasing teachers' motivation to learn. Another potential explanation is that teachers recognized the importance of dialogic pedagogy during the learning process. According to self-determination theory (Deci & Ryan, 2012), individuals are motivated by feelings of effectiveness and the desire to acquire knowledge. This recognition likely contributed to their increased motivation to learn dialogic pedagogy.

Third, teachers in both the BERT and Llama3 groups reported a high level of satisfaction with the PD workshops, with no significant difference between the two groups. These findings suggest that the performance differences between BERT and Llama3 did not significantly affect user satisfaction. Given that satisfaction arises from specific benefits obtained (Perse & Ferguson, 2000), we can infer from the satisfaction items that PD workshops incorporating either BERT or Llama3 provided teachers with high levels of enjoyment, engagement, and perceived usefulness, as evidenced by the high scores teachers gave on each satisfaction item. These conclusions are corroborated by subsequent interviews, where teachers' curiosity about AI reasoning and their perception of AI as a tool aiding their understanding contributed to their sense of engagement and perceived usefulness. Furthermore, teachers in both groups indicated that BERT and Llama3 served as learning partners that did not induce stress, thereby enhancing their enjoyment. These findings align with previous research (Cheng & Jiang, 2020; Xie et al., 2024), which posits that user satisfaction with AI is influenced by utilitarian, technological, hedonic, and social gratifications. This corresponds to AI's facilitative role in learning, teachers' curiosity about AI, and AI's function as a collaborative learning partner.

Implications and future directions

This study holds significant implications regarding the use of AI in education. First, teachers can integrate AI models (e.g., LLMs) into professional development programmes to enhance pedagogical knowledge acquisition and foster positive learning experiences. However, our results emphasize the importance of critical engagement: teachers should systematically evaluate AI-generated outputs and integrate reflective practices rather than blindly trusting or distrusting the technology. For practical classroom applications, teachers can strategically use reliable AI systems like BERT to analyse domain-specific dimensions (e.g., cognitive engagement, socio-emotional dynamics) in alignment with teachers' instructional goals. For example, it is crucial for teachers to learn how to foster and assess students' reasoning abilities (Talman et al., 2021). AI models like BERT and Llama3 can be used to identify reasoning-related components in classrooms, help teachers reflect on their teaching practices, and further facilitate effective guidance and assessment of students' thinking. Second, technologists may shift from a performance-first focus to application-priority optimization when developing AI-powered educational models, as enhanced computational performance does not inherently correlate with improved educational outcomes. This shift is not intended to diminish the importance of pursuing more accurate AI models; rather, it emphasizes the need for contextual relevance and modular functionality to meet diverse pedagogical needs. For example, technologists can implement human-in-the-loop frameworks to embed pedagogical theories and develop teacher-centric interfaces, ensuring that AI-powered educational applications are both effective and adaptable to real-world classroom environments. Third, educational administrators should adopt a multidimensional evaluation framework when purchasing AI-powered educational applications. This framework should encompass factors such as teacher training requirements, potential impacts, and deployment costs, rather than solely prioritizing technical benchmarks. For instance, pilot validations can be implemented to assess the real-world educational value of these applications before large-scale adoption. Furthermore, administrators should provide contextualized training for teachers on how to effectively utilize AI tools in their teaching practices. This can be achieved by organizing training workshops and inviting experienced educators to share their strategies for AI application in education.

Despite these findings and implications, several issues warrant further exploration in future research. First, the performance of LLMs in a given task is heavily influenced by the design of the prompts. Factors such as task formulation, context length, and the inclusion of few-shot examples can significantly impact the performance of LLMs, as indicated by Tran et al. (2024). Although we adhered to methodologies from previous studies and presented three representative prompts in this research, the potential for more effective prompt designs remains unexplored. For instance, prompt effectiveness could be enhanced by investigating new advanced prompt engineering techniques. Second, due to cost constraints, we selected Llama3, which was claimed to be the most powerful at the time of our study. Currently, other powerful open-source LLMs are available, such as DeepSeek-R1, which could be considered in future research. Third, our study is limited to the context of learning the APT framework. It would be more convincing to explore the outcomes when teachers use BERT and Llama3 to analyse their classroom teaching and subsequently adjust their dialogic practices based on the analysis. Fourth, only preservice teachers were recruited in this exploratory study. It remains uncertain whether different results would be observed with in-service teachers. Fifth, the multi-phased workshop presents challenges for us to accurately identify which phases contributed to specific aspects of teachers' improved learning motivation and high level of satisfaction, although the interview analysis partially supports that the use of BERT and Llama3 increased participants' motivation and provided them with satisfaction. Future research should focus on recruiting in-service teachers to participate in PD programs on dialogic pedagogy with incorporated AI models and comprehensively evaluate the effects of these programs on both theoretical and practical aspects.

ACKNOWLEDGEMENTS

This work was supported by the Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

FUNDING INFORMATION

This work was supported by the Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the first author [DW] upon reasonable request.

ETHICS STATEMENT

This work has obtained ethical approval from the Human Research Ethics Committee at the University of Hong Kong and all participants provided informed consent.

ORCID

Deliang Wang https://orcid.org/0009-0008-6488-0234
Gaowei Chen https://orcid.org/0000-0002-6847-4013

REFERENCES

Alexander, R. J. (2001). *Culture and pedagogy: International comparisons in primary education*. Blackwell Publishers. https://api.semanticscholar.org/CorpusID:153884653

Alexander, R. J. (2017). *Towards dialogic teaching: Rethinking classroom talk* (5th ed.). Dialogos.

Bae, C. L., Mills, D. C., Zhang, F., Sealy, M., Cabrera, L., & Sea, M. (2021). A systematic review of science discourse in K–12 urban classrooms in the United States: Accounting for individual, collective, and contextual factors. *Review of Educational Research*, *91*(6), 831–877. https://doi.org/10.3102/0034654321104

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. University of Texas Press.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. M. Drucker (Eds.), *CHI '21: CHI conference on human factors in computing systems* (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

Borko, H., Jacobs, J., & Koellner, K. (2010). Contemporary approaches to teacher professional development. *International Encyclopedia of Education*, *7*(2), 548–556. https://doi.org/10.1016/B978-0-08-044894-7.00654-0

Borko, H., Zaccarelli, F. G., Reigh, E., & Osborne, J. (2021). Teacher facilitation of elementary science discourse after a professional development initiative. *The Elementary School Journal*, *121*(4), 561–585. https://doi.org/10.1086/714082

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chen, G. (2020). A visual learning analytics (VLA) approach to video-based teacher professional development: Impact on teachers' beliefs, self-efficacy, and classroom talk practice. *Computers & Education*, *144*, 103670. https://doi.org/10.1016/j.compedu.2019.103670

Chen, G., Chan, C. K., Chan, K. K., Clarke, S. N., & Resnick, L. B. (2020). Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. *Journal of the Learning Sciences*, *29*(4–5), 642–680. https://doi.org/10.1080/10508406.2020.1783269

Chen, G., Zhang, J., Chan, C. K., Michaels, S., Resnick, L. B., & Huang, X. (2020). The link between student-perceived teacher talk and student enjoyment, anxiety and discursive engagement in the classroom. *British Educational Research Journal*, *46*(3), 631–652. https://doi.org/10.1002/berj.3600

Cheng, Y., & Jiang, H. (2020). How do AI-driven chatbots impact user experience? Examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use. *Journal of Broadcasting & Electronic Media*, *64*(4), 592–614. https://doi.org/10.1080/08838151.2020.1834296

Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B., Saunders, L., & Coe, R. (2015). *Developing great teaching: Lessons from the international reviews into effective professional development*. Teacher Development Trust. https://durham-repository.worktribe.com/output/1631202

Cui, R., & Teo, P. (2021). Dialogic education for classroom teaching: A critical review. *Language and Education*, *35*(3), 187–203. https://doi.org/10.1080/09500782.2020.1837859

de Araujo, A., Papadopoulos, P. M., McKenney, S., & de Jong, T. (2023). Automated coding of student chats, a trans-topic and language approach. *Computers and Education: Artificial Intelligence*, *4*, 100123. https://doi.org/10.1016/j.caeai.2023.100123

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 416–436). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n21

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1638–1653). Association for Computational Linguistics. https://doi.org/10.18653/V1/2021.ACL-LONG.130

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/V1/N19-1423

Dogan, S., Nalbantoglu, U. Y., Celik, I., & Agacli Dogan, N. (2025). Artificial intelligence professional development: A systematic review of TPACK, designs, and effects for teacher learning. *Professional Development in Education*, 1–28. https://doi.org/10.1080/19415257.2025.2454457

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016). Automatic teacher modeling from live classroom audio. In J. Vassileva, J. Blustein, L. Aroyo, & S. K. D'Mello (Eds.), *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 45–53). ACM. https://doi.org/10.1145/2930238.2930250

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice*, *20*(5), 1–19. https://doi.org/10.53761/1.20.5.02

Fisher, J. B., Schumaker, J. B., Culbertson, J., & Deshler, D. D. (2010). Effects of a computerized professional development program on teacher and student outcomes. *Journal of Teacher Education*, *61*(4), 302–312. https://doi.org/10.1177/0022487110369556

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gröschner, A., Schindler, A. K., Holzberger, D., Alles, M., & Seidel, T. (2018). How systematic video reflection in teacher professional development regarding classroom discourse contributes to teacher and student self-efficacy. *International Journal of Educational Research*, *90*, 223–233. https://doi.org/10.1016/j.ijer.2018.02.003

Gröschner, A., Seidel, T., Kiemer, K., & Pehmer, A. K. (2015). Through the lens of teacher professional development components: The 'dialogic video Cycle' as an innovative program to foster classroom dialogue. *Professional Development in Education*, *41*(4), 729–756. https://doi.org/10.1080/19415257.2014.939692

Gröschner, A., Seidel, T., Pehmer, A. K., & Kiemer, K. (2014). Facilitating collaborative teacher learning: The role of "mindfulness" in video-based teacher professional development programs. *Gruppendynamik und Organisationsberatung*, *45*, 273–290. https://doi.org/10.1007/s11612-014-0248-0

He, G., Buijsman, S., & Gadiraju, U. (2023). How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–29. https://doi.org/10.1145/3610067

Hennessy, S., Dragovic, T., & Warwick, P. (2018). A research-informed, school-based professional development workshop programme to promote dialogic teaching with interactive technologies. *Professional Development in Education*, *44*(2), 145–168. https://doi.org/10.1080/19415257.2016.1258653

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction*, *25*, 100404. https://doi.org/10.1016/j.lcsi.2020.100404

Herbel-Eisenmann, B. A., Steele, M. D., & Cirillo, M. (2013). (Developing) teacher discourse moves: A framework for professional development. *Mathematics Teacher Educator*, *1*(2), 181–196. https://doi.org/10.5951/mathteaceduc.1.2.0181

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education*, *43*(3), 325–356. https://doi.org/10.1080/0305764X.2013.786024

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Huang, B., Jong, M. S. Y., Tu, Y. F., Hwang, G. J., Chai, C. S., & Jiang, M. Y. C. (2022). Trends and exemplary practices of STEM teacher professional development programs in K-12 contexts: A systematic review of empirical studies. *Computers & Education*, *189*, 104577. https://doi.org/10.1016/j.compedu.2022.104577

Huang, G. Y., Chen, J., Liu, H., Fu, W., Ding, W., Tang, J., Yang, S., Li, G., & Liu, Z. (2020). Neural multi-task learning for teacher question detection in online classrooms. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Proceedings of the 21st International conference on artificial intelligence in education* (pp. 269–281). Springer. https://doi.org/10.1007/978-3-030-52237-7_22

Huang, A. Y., Lu, O. H., & Yang, S. J. (2023). Effects of artificial intelligence–enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, *194*, 104684. https://doi.org/10.1016/j.compedu.2022.104684

Hwang, G. J., Tu, Y. F., & Chu, H. C. (2023). Conceptions of the metaverse in higher education: A draw-a-picture analysis and surveys to investigate the perceptions of students with different motivation levels. *Computers & Education*, *203*, 104868. https://doi.org/10.1016/j.compedu.2023.104868

Hwang, G. J., Yang, L. H., & Wang, S. Y. (2013). A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, *69*, 121–130. https://doi.org/10.1016/j.compedu.2013.07.008

Jacobs, J., Scornavacco, K., Clevenger, C., Suresh, A., & Sumner, T. (2024). Automated feedback on discourse moves: Teachers' perceived utility of a professional learning tool. *Educational Technology Research and Development*, *72*, 1307–1329. https://doi.org/10.1007/s11423-023-10338-6

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, 103631. https://doi.org/10.1016/j.tate.2022.103631

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464. https://doi.org/10.3102/0013189X18785613

Kim, M. Y., & Wilkinson, I. A. (2019). What is dialogic teaching? Constructing, deconstructing, and reconstructing a pedagogy of classroom talk. *Learning, Culture and Social Interaction*, *21*, 70–86. https://doi.org/10.1016/j.lcsi.2019.02.003

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 22199–22213). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf

Kupor, A., Morgan, C., & Demszky, D. (2023). Measuring five accountable talk moves to improve instruction at scale. CoRR, abs/2311.10749. https://doi.org/10.48550/ARXIV.2311.10749

Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality and Safety*, *25*(10), 808–820. https://doi.org/10.1136/bmjqs-2015-004417

Li, H., Wang, Z., Tang, J., Ding, W., & Liu, Z. (2020). Siamese neural networks for class activity detection. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Proceedings of the 21st international conference on artificial intelligence in education* (pp. 162–167). Springer. https://doi.org/10.1007/978-3-030-52240-7_30

Major, L., Warwick, P., Rasmussen, I., Ludvigsen, S., & Cook, V. (2018). Classroom dialogue and digital technologies: A scoping review. *Education and Information Technologies*, *23*, 1995–2028. https://doi.org/10.1007/s10639-018-9701-y

Major, L., & Watson, S. (2018). Using video to support in-service teacher professional development: The state of the field, limitations and possibilities. *Technology, Pedagogy and Education*, *27*(1), 49–68. https://doi.org/10.1080/1475939X.2017.1361469

Malone, T. W., & Lepper, M. R. (2021). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction* (pp. 223–254). Routledge.

Mercer, N., & Dawes, L. (2014). The study of talk between teachers and students, from the 1970s until the 2010s. *Oxford Review of Education*, *40*(4), 430–445. https://doi.org/10.1080/03054985.2014.934087

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach*. Routledge. https://doi.org/10.4324/9780203946657

Meta-llama. (2024). *Meta-llama/Meta-Llama-3-8B*. Hugging Face. https://huggingface.co/meta-llama/Meta-Llama-3-8B

Michaels, S., O'Connor, C., Hall, M. W., & Resnick, L. B. (2010). *Accountable talk sourcebook: For classroom conversation that works*. University of Pittsburgh.

Mintii, I., & Semerikov, S. (2024). Optimizing teacher training and retraining for the age of AI-powered personalized learning: A bibliometric analysis. In *Information technology for education, science, and technics. ITEST 2024* (pp. 339–357). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71804-5_23

Moreau-Pernet, B., Tian, Y., Sawaya, S., Foltz, P., Cao, J., Milne, B., & Christie, T. (2024). Classifying tutor discursive moves at scale in mathematics classrooms with large language models. In D. Joyner, M. K. Kim, X. Wang, & M. Xia (Eds.), *Proceedings of the eleventh ACM conference on learning @ scale* (pp. 361–365). ACM. https://doi.org/10.1145/3657604.3664664

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring teachers' trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 56–66). Association for Computing Machinery. https://doi.org/10.1145/3506860.3506866

O'Connor, C., & Michaels, S. (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. https://doi.org/10.1016/j.ijer.2017.11.003

O'Connor, C., & Snow, C. (2017). Classroom discourse: What do we need to know for research and for practice? In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *The Routledge handbook of discourse processes* (2nd ed., pp. 315–342). Routledge.

Osborne, J. F., Borko, H., Fishman, E., Gomez Zaccarelli, F., Berson, E., Busch, K. C., Reigh, E., & Tseng, A. (2019). Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational Research Journal*, *56*(4), 1067–1112. https://doi.org/10.3102/0002831218812059

Perse, E. M., & Ferguson, D. A. (2000). The benefits and costs of web surfing. *Communication Quarterly*, *48*(4), 343–359. https://doi.org/10.1080/01463370009385603

Rapanta, C., & Felton, M. K. (2022). Learning to argue through dialogue: A review of instructional approaches. *Educational Psychology Review*, *34*, 477–509. https://doi.org/10.1007/s10648-021-09637-2

Resnick, L. B., Asterhan, C. S., & Clarke, S. N. (Eds.). (2015). *Socializing intelligence through academic talk and dialogue*. American Educational Research Association.

Resnick, L. B., Asterhan, C. S., & Clarke, S. N. (2018). *Accountable talk: Instructional dialogue that builds the mind*. The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the United Nations Educational, Scientific and Cultural Organization (UNESCO).

Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, L. F. Quinn, & D. M. Gollnick (Eds.), *Wiley handbook of teaching and learning* (pp. 323–338). Wiley-Blackwell. https://doi.org/10.1002/9781118955901.ch13

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. CoRR, abs/2402.07927. https://doi.org/10.48550/ARXIV.2402.07927

Sedova, K., Sedlacek, M., & Svaricek, R. (2016). Teacher professional development as a means of transforming student classroom talk. *Teaching and Teacher Education*, *57*, 14–25. https://doi.org/10.1016/j.tate.2016.03.005

Sheridan, S. M., Edwards, C. P., Marvin, C. A., & Knoche, L. L. (2009). Professional development in early childhood programs: Process issues and research needs. *Early Education and Development*, *20*(3), 377–401. https://doi.org/10.1080/10409280802582795

Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, *32*(1), 47–63. https://doi.org/10.1080/09243453.2020.1772841

Song, Y., Chen, X., Hao, T., Liu, Z., & Lan, Z. (2019). Exploring two decades of research on classroom dialogue by using bibliometric analysis. *Computers & Education*, *137*, 12–31. https://doi.org/10.1016/j.compedu.2019.04.002

Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2021). Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, *59*(3), 496–521. https://doi.org/10.1177/0735633120968554

Stone, C., Donnelly, P. J., Dale, M., Capello, S., Kelly, S., Godley, A., & D'Mello, S. K. (2019). Utterance-level modeling of indicators of engaging classroom discourse. In M. C. Desmarais, C. F. Lynch, A. Merceron, & R. Nkambou (Eds.), *Proceedings of the 12th international conference on educational data mining* (pp. 420–425). International Educational Data Mining Society.

Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4654–4662). European Language Resources Association. https://aclanthology.org/2022.lrec-1.497

Talman, K., Vierula, J., Kanerva, A. M., Virkki, O., Koivisto, J. M., & Haavisto, E. (2021). Instruments for assessing reasoning skills in higher education: A scoping review. *Assessment & Evaluation in Higher Education*, *46*(3), 376–392. https://doi.org/10.1080/02602938.2020.1776212

Tammets, K., & Ley, T. (2023). Integrating AI tools in teacher professional learning: A conceptual model and illustrative case. *Frontiers in Artificial Intelligence*, *6*, 1255089. https://doi.org/10.3389/frai.2023.1255089

Tao, Y., & Chen, G. (2023). Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature. *Learning, Culture and Social Interaction*, *39*, 100702. https://doi.org/10.1016/j.lcsi.2023.100702

Tao, Y., & Chen, G. (2024). The relationship between teacher talk and students' academic achievement: A meta-analysis. *Educational Research Review*, *45*, 100638. https://doi.org/10.1016/j.edurev.2024.100638

Thompson, P., Paatsch, L., & Nolan, A. (2024). Understanding the talk patterns in secondary school classrooms. *Language and Education*, *39*(2), 489–511. https://doi.org/10.1080/09500782.2024.2343292

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L. C. (2024). Analyzing large language models for classroom discussion assessment. In B. Paaßen & C. D. Epp (Eds.), *Proceedings of the 17th international conference on educational data mining* (pp. 500–510). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729870

van der Linden, S., van der Meij, J., & McKenney, S. (2022). Teacher video coaching, from design features to student impacts: A systematic literature review. *Review of Educational Research*, *92*(1), 114–165. https://doi.org/10.3102/00346543211046984

Vasalampi, K., Metsäpelto, R. L., Salminen, J., Lerkkanen, M. K., Mäensivu, M., & Poikkeus, A. M. (2021). Promotion of school engagement through dialogic teaching practices in the context of a teacher professional

development programme. *Learning, Culture and Social Interaction*, *30*, 100538. https://doi.org/10.1016/j.lcsi.2021.100538
Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Harvard University Press.
Wang, D., Bian, C., & Chen, G. (2024). Using explainable AI to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology*, *55*, 2530–2556. https://doi.org/10.1111/bjet.13466
Wang, D., & Chen, G. (2024). Are perfect transcripts necessary when we analyze classroom dialogue using AIoT? *Internet of Things*, *25*, 101105. https://doi.org/10.1016/j.iot.2024.101105
Wang, D., Shan, D., Zheng, Y., Guo, K., Chen, G., & Lu, Y. (2023). Can ChatGPT detect student talk moves in classroom discourse? A preliminary comparison with Bert. In M. Feng, T. Kaser, & P. Talukdar (Eds.), *Proceedings of the 16th international conference on educational data mining* (pp. 515–519). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115772
Wang, D., Tao, Y., & Chen, G. (2024). Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, *123*, 102275. https://doi.org/10.1016/j.ijer.2023.102275
Wang, R. E., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications* (pp. 626–667). Association for Computational Linguistics. https://doi.org/10.18653/V1/2023.BEA-1.53
Wegerif, R. (2007). *Dialogic education and technology*. Springer. https://doi.org/10.1007/978-0-387-71142-3
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
Whitehill, J., & LoCasale-Crouch, J. (2023). Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. CoRR, abs/2310.01132. https://doi.org/10.48550/ARXIV.2310.01132
Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*, *29*, 24963–24990. https://doi.org/10.1007/s10639-024-12819-4
Xie, C., Wang, Y., & Cheng, Y. (2024). Does artificial intelligence satisfy you? A meta-analysis of user gratification and user satisfaction with AI-powered chatbots. *International Journal of Human Computer Interaction*, *40*(3), 613–623. https://doi.org/10.1080/10447318.2022.2121458
Yang, Z., & Wang, S. (2022). Dialogic teaching in secondary classrooms in China: Features, commonalities, and distinctiveness. *Learning, Culture and Social Interaction*, *34*, 100619. https://doi.org/10.1016/j.lcsi.2022.100619
Zhen, Y., Zheng, L., & Chen, P. (2021). Constructing knowledge graphs for online collaborative programming. *IEEE Access*, *9*, 117969–117980. https://doi.org/10.1109/ACCESS.2021.3106324
Zheng, L., Zhong, L., & Fan, Y. (2023). An automated analysis of topic distributions and features approach to promoting group performance, collaborative knowledge building and socially shared regulation in online collaborative learning. *Australasian Journal of Educational Technology*, *39*(3), 114–132. https://doi.org/10.14742/ajet.7995
Zheng, L., Zhong, L., & Niu, J. (2022). 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. https://doi.org/10.1080/02602938.2021.1883549

How to cite this article: Wang, D., & Chen, G. (2025). Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy. *British Journal of Educational Technology*, *00*, 1–34. https://doi.org/10.1111/bjet.13604

APPENDIX A

TABLE A1 The description of dialogic moves in the corpus (Jacobs et al., 2022; Suresh et al., 2022).

Role	Dialogic move	Description
Teacher	Keeping everyone together	Encouraging students to engage in active listening and get to know each other.
	Getting students to relate to another's idea	Encouraging students to actively respond to their peers' contributions.
	Restating	Echoing a student's words exactly or partially.
	Revoicing	Rephrasing a student's statement with modifications or additions.
	Pressing for accuracy	Encouraging students to use mathematical terminology and contribute mathematically.
	Pressing for reasoning	Asking students to provide evidence or explain their reasoning.
	None	The teacher's utterance lacks any of the dialogic moves mentioned above.
Student	Relating to another student	Engaging in commentary or questioning a classmate's ideas.
	Asking for more information	The student asks for more information, expresses confusion or seeks help.
	Making a claim	The student offers a mathematical assertion, states a fact or lists a step in their response.
	Providing evidence or reasoning	The student explains their thought process, provides supporting evidence or elaborates on the reasoning.
	None	The student's utterance lacks any of the dialogic moves mentioned above.

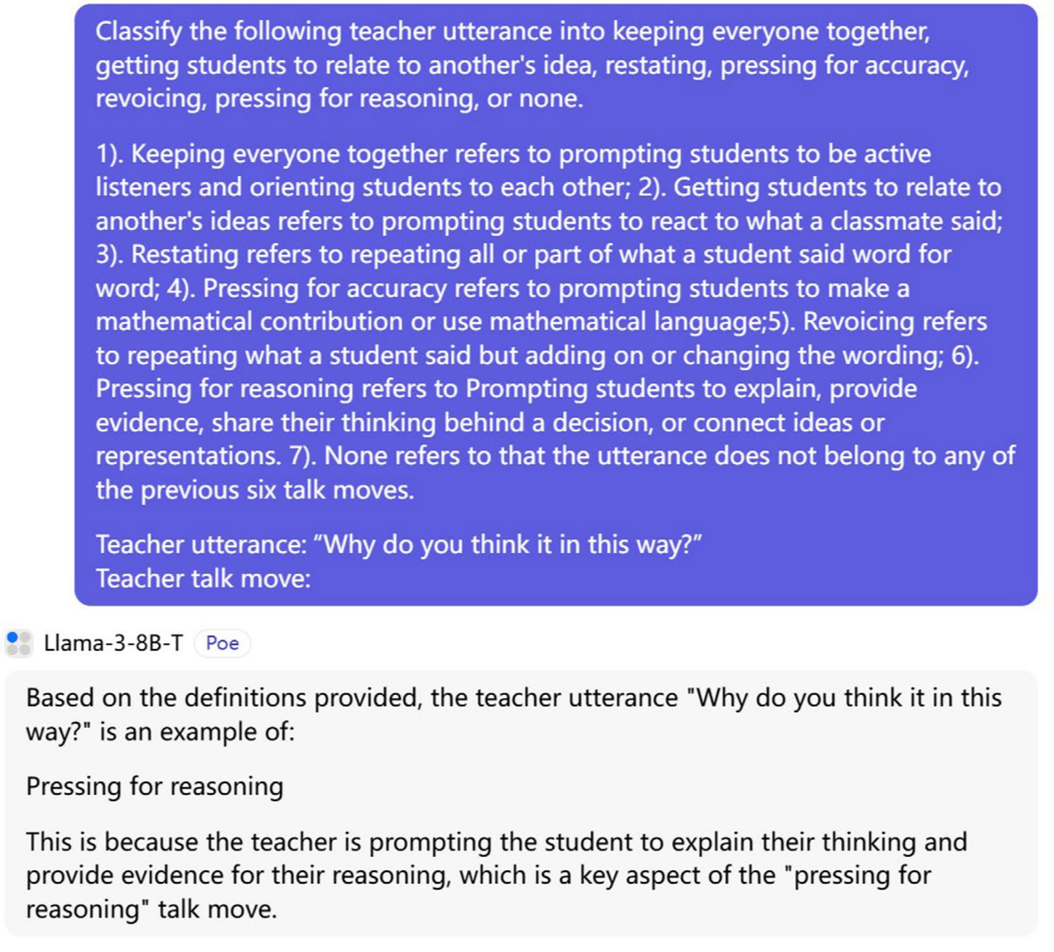

FIGURE A1 An example of Llama3's analysis regarding a teacher's utterance.

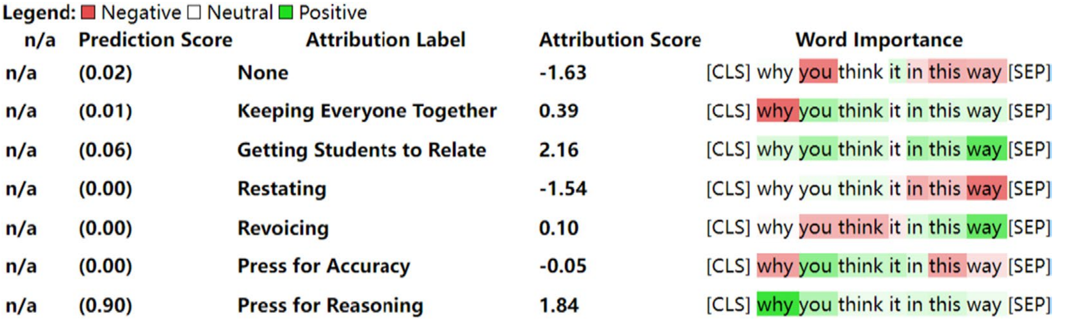

FIGURE A2 An example of BERT's analysis with explanations regarding a teacher's utterance.

TABLE A2 The test for evaluating participants' prior knowledge of the APT framework.

No.	Questions
1	Are you familiar with the concept of academically productive talk? If so, please describe your understanding of it.
2	Do you know the common patterns within classroom dialogue? If so, please elaborate on these patterns.
3	Can you identify the theories that underpin the academically productive talk framework? If so, please specify them.
4	Do you understand what a dialogic move is? If so, please explain your understanding and write down the dialogic moves that you frequently use.

TABLE A3 The test for evaluating participants' knowledge of the APT framework after the workshop.

No.	Questions		
1	What is the definition of classroom dialogue?		
2	What is the common pattern of dialogue observed in traditional classrooms?		
3	According to the APT framework, to what three aspects should classroom talk be accountable? Please explain the meaning of each aspect.		
4	How is a dialogic move defined within the APT framework? Please talk about your understanding.		
5	Identify the dialogic move the teacher employed in the provided utterance. If possible, explain the reasoning behind your answer.		
	Dialogue 1	Teacher: “Okay, you are going to estimate benchmarks.”	
		Teacher: “Is it closer to zero, half, one, one and a half, or two?”	Dialogic move: ______________
		Teacher: “When you've got your answers to both of those, stand up, please.”	Dialogic move: ______________
	Dialogue 2	Teacher: “Close your eyes.”	
		Teacher: “On the first one, three-fifths plus one-fourth, what is it closest to?”	Dialogic move: ______________
	Dialogue 3	Student A: “I want to see why they think it's a half.”	
		Teacher: “Oh, okay.”	
		Teacher: “So why did you say that was a half, Michael?”	Dialogic move: ______________
	Dialogue 4	Teacher: “Is that what your answer was, or was it just a half?”	
		Michael: “A half.”	
		Teacher: “Just a half, okay.”	Dialogic move: ______________
		Teacher: “Who can help Michael clarify his thinking?”	Dialogic move: ______________
	Dialogue 5	Teacher: “I can tell you, you needed to connect here to here.”	
		Student: “Oh, that's okay.”	
		Teacher: “Okay. So now, let's triple these.”	
		Teacher: “So what does that mean I'm multiplying everything by?”	Dialogic move: ______________

TABLE A3 (Continued)

No.	Questions		
	Dialogue 6	Teacher: "And all I have to add four plus three, this is seven plus, that's where those numbers come in."	
		Teacher: "I had to figure out what times my denominator was going to give me my new denominator."	
		Teacher: "Does that make sense?"	Dialogic move: ______________
6	Identify the dialogic move the student employed in the provided utterance. If possible, explain the reasoning behind your answer.		
	Dialogue 1	Teacher: "Can you tell me what these numbers mean over here?"	
		Student: "Alright. Okay."	
		Student: "These are the dice numbers. Like, numbers that can be rolled on the two dice, and one's not there cause you can't roll one."	Dialogic move: ______________
	Dialogue 2	Erik: "I think A should be the right answer."	
		Teacher: "David, what do you think?"	
		Teacher: "Did you want to say something?"	
		David: "Um, I agree with Erik"	Dialogic move: ______________
	Dialogue 3	Teacher: "Don't forget once you have completed drawing your- your plot, then you must show those five points on that scale, yes?"	
		Student: "Mr. Learoyd, this- is this correct?"	
		Teacher: "Looks all right to me."	
		Student: "I don't know- I don't get how to, like, do it, how to make the box plot."	Dialogic move: ______________
	Dialogue 4	Teacher: "Why did you select the mean rather than the, uh, the mode for example?"	
		Student: "Cause that's on an average of, um, all of the … things."	Dialogic move: ______________

TABLE A4 The learning motivation questionnaire.

No.	Items
1	I think learning dialogic pedagogy is interesting and valuable.
2	I would like to learn more and observe more in the workshop on dialogic pedagogy.
3	It is worth learning those things about dialogic pedagogy.
4	It is important for me to learn dialogic pedagogy well.
5	It is important to know the dialogic pedagogy knowledge related to classroom teaching.
6	I will actively search for more information and learn about dialogic pedagogy.
7	It is important for me to take the training workshop on dialogic pedagogy.

TABLE A5 The satisfaction questionnaire.

No.	Items
1	I believe that I will remember everything taught today.
2	The workshop kept me focused on the content throughout.
3	I am confident that I will use the routine learned today.
4	This workshop made me very enthusiastic about the content taught.
5	It will be easy to summarize for others what the training is all about.
6	It was easy to concentrate on the content of this workshop.
7	I plan to implement the routine.
8	I had a lot of fun during this session.
9	I clearly understand everything that was taught today.
10	The workshop was engaging throughout.
11	I am looking forward to incorporating this routine into my teaching.
12	This workshop was very enjoyable for me.
13	This workshop was superior to others I have attended.
14	Overall, I was highly satisfied with this workshop.

Supplementary analysis of RQ2

To further identify which phases contributed to specific aspects of teachers' improved knowledge of the APT framework, we divided the posttest knowledge data into two parts: Phase 2 scores and Phases 3 & 4 scores. Table A3 in the appendix illustrates this division: the first four questions primarily focus on the content of Phase 2, while the remaining questions target the content of Phases 3 and 4. Since the content of Phase 2 was not revisited in Phases 3 and 4, we assume that the Phase 2 scores can be considered an independent variable, reflecting participants' understanding of the basic components of the APT framework introduced in Phase 2. In contrast, the Phases 3 & 4 scores may be influenced by both BERT- or Llama3-supported activities and the foundational knowledge from Phase 2. To isolate the impact of Phase 2 on Phases 3 and 4, we performed additional analyses.

First, we conducted a repeated-measures ANOVA to compare participants' pretest knowledge scores, Phase 2 scores, and Phases 3 and 4 scores, aiming to determine whether participants showed significant improvement across the different phases. Given that the four questions in Phase 2 have a total of 40 points and the remaining questions in Phases 3 and 4 have a total of 60 points, we scaled the scores to ensure both sets had a total of 100 points for a fair comparison with the pretest scores. Second, we performed ANCOVA to compare the two groups' Phase 2 scores, using the pretest scores as a covariate. Third, we conducted another ANCOVA, considering the Phase 2 scores as a covariate to isolate their effects on the Phases 3 & 4 scores.

Table A6 shows a significant difference between teachers' pretest knowledge scores, Phase 2 scores, and Phases 3 and 4 scores, with Phases 3 and 4 scores being significantly higher than Phase 2 scores, and Phase 2 scores significantly higher than pretest scores. Table A7 indicates no significant difference between the two groups' Phase 2 scores. Table A8, which treats the Phase 2 scores as a covariate, reveals no significant difference in the Phases 3 and 4 scores between the two groups.

These results, together with the analyses presented in previous sections, suggest that the increased knowledge of the APT framework (reflected in the total posttest knowledge scores) is attributed to both the introduction in Phase 2 and the learning of dialogic moves with AI support in Phases 3 and 4. Specifically, the enhanced understanding of the basic APT framework is due to Phase 2, while the improved knowledge of how to analyse dialogic moves is facilitated by the activities supported by BERT and Llama3. No significant differences were found between the two groups in posttest knowledge.

TABLE A6 Repeated-measures ANOVA results comparing teachers' pretest knowledge scores, Phase 2 scores, and Phases 3 and 4 scores.

Group	*n*	Pretest *Mean* (*SD*)	Phase 2 *Mean* (*SD*)	Phases 3 and 4 *Mean* (*SD*)	*F*	*df*	*Location of significance*
BERT	30	12.40 (12.93)	71.00 (16.32)	83.97 (18.36)	235.78**	2	Pretest < Phase 2 < Phases 3&4
Llama3	30	10.20 (10.19)	63.96 (17.32)	82.67 (15.73)	232.29**	2	Pretest < Phase 2 < Phases 3&4

Note: Adjustment for multiple comparisons: Bonferroni.

**$p < 0.01$.

TABLE A7 The ANCOVA comparing teachers' Phase 2 scores (with pretest scores as a covariate).

Group	*n*	*Mean* (*SD*)	*Adjust mean* (*Std. error*)	*F*	*p*
BERT	30	71.00 (16.32)	70.462 (2.92)	2.07	0.16
Llama3	30	63.96 (17.32)	64.496 (2.92)		

TABLE A8 The ANCOVA comparing teachers' Phases 3 and 4 scores (with Phase 2 scores as a covariate).

Group	*n*	*Mean* (*SD*)	*Adjust mean* (*Std. error*)	*F*	*p*
BERT	30	83.97 (18.36)	82.89 (3.03)	0.04	0.84
Llama3	30	82.67 (15.73)	83.75 (3.03)		