

Evaluating the performance of ChatGPT and GPT-4o in coding classroom discourse data: A study of synchronous online mathematics instruction[☆]

Simin Xu^a, Xiaowei Huang^a, Chung Kwan Lo^{a,*}, Gaowei Chen^b, Morris Siu-yung Jong^c

^a Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR

^b Faculty of Education, The University of Hong Kong, Hong Kong SAR

^c Department of Curriculum and Instruction & Centre for Learning Sciences and Technologies, The Chinese University of Hong Kong, Hong Kong SAR

ARTICLE INFO

Keywords:

ChatGPT
GPT-4o
Classroom discourse analysis
Professional development
Mathematics instruction

ABSTRACT

High-quality instruction is essential to facilitating student learning, prompting many professional development (PD) programmes for teachers to focus on improving classroom dialogue. However, during PD programmes, analysing discourse data is time-consuming, delaying feedback on teachers' performance and potentially impairing the programmes' effectiveness. We therefore explored the use of ChatGPT (a fine-tuned GPT-3.5 series model) and GPT-4o to automate the coding of classroom discourse data. We equipped these AI tools with a codebook designed for mathematics discourse and academically productive talk. Our dataset consisted of over 400 authentic talk turns in Chinese from synchronous online mathematics lessons. The coding outcomes of ChatGPT and GPT-4o were quantitatively compared against a human standard. Qualitative analysis was conducted to understand their coding decisions. The overall agreement between the human standard, ChatGPT output, and GPT-4o output was moderate (Fleiss's Kappa = 0.46) when classifying talk turns into major categories. Pairwise comparisons indicated that GPT-4o (Cohen's Kappa = 0.69) had better performance than ChatGPT (Cohen's Kappa = 0.33). However, at the code level, the performance of both AI tools was unsatisfactory. Based on the identified competences and weaknesses, we propose a two-stage approach to classroom discourse analysis. Specifically, GPT-4o can be employed for the initial category-level analysis, following which teacher educators can conduct a more detailed code-level analysis and refine the coding outcomes. This approach can facilitate timely provision of analytical resources for teachers to reflect on their teaching practices.

1. Introduction

Classroom discourse analysis is a critical aspect of teacher education (Chen et al., 2020; Ng et al., 2021). It examines how discourse is used to convey knowledge and engage students in class discussions (Adler & Ronda, 2015; Mercer et al., 2019; Resnick et al., 2010). Such analysis is particularly important in synchronous online mathematics instruction, where effective communication is crucial for students' understanding of the knowledge imparted and their engagement in class (Lo & Liu, 2022; Roberts & Olarte, 2023). In virtual classrooms, the dynamics of teacher–student interactions differ significantly from those in traditional settings, making it essential to understand how discourse practices impact virtual classroom dynamics (Gutentag et al., 2022; Jaekel et al., 2023). This understanding can enable teacher educators to identify effective instructional practices and areas for improvement.

In the context of teacher professional development (PD), particularly in mathematics, the analyses of teacher–student interactions through classroom discourse are crucial for informing instructional improvement (Chen et al., 2020). From the perspective of teacher educators, however, “this laborious and time-consuming process poses challenges” (Wang & Chen, 2024, p. 2). Therefore, this study evaluated the performance of ChatGPT (a fine-tuned GPT-3.5 series model) and GPT-4o, which are advanced Artificial Intelligence (AI) chatbots developed by OpenAI (2022, 2024), in coding classroom discourse data from synchronous online mathematics lessons. Comparisons of human and AI performance on the same task can provide important insights into how humans can offload specific tasks to AI. A parallel example is the use of AI for automated scoring of student writing, as explored by Mizumoto and Eguchi (2023). Ideally, “This feedback method can effectively reduce the workload of teachers in correcting student writing” (p. 9). In the

[☆] This manuscript has not been published in any language before; and is not under consideration concurrently for publication elsewhere.

* Corresponding author.

E-mail address: chungkwanlo@eduhk.hk (C.K. Lo).

current study, we adopted a similar approach, as we evaluated the performance of the two AI tools and explored their potential to assist teacher educators in PD contexts. The following research questions (RQ1 and RQ2) guided the study.

- RQ1: What is the level of agreement between the human standard, ChatGPT, and GPT-4o in coding discourse data from synchronous online mathematics lessons?
- RQ2: How do the coding outcomes of the AI tools align with or differ from the human standard, and what are the underlying reasons for differences, if any?

2. Background and related work

In this section, we first provide an introduction to video-supported teacher PD and its challenges, which highlights the need to explore the use of AI tools in teacher education. Subsequently, we provide a brief overview of the AI tools investigated in this study. Thereafter, we examine empirical studies related to the application of these AI tools in the field of qualitative analysis.

2.1. Video-supported teacher professional development

Darling-Hammond et al. (2017) identified that the key elements of effective teacher PD programmes include coaching, support, feedback, and reflection. Teacher educators can incorporate video-based tasks into their PD activities to coach participants and provide feedback on instructional improvement (Gaudin & Chaliès, 2015). Importantly, videos capture the richness and complexity of classroom interactions (Major & Watson, 2018), which can facilitate reflection (Larison et al., 2024). However, such video-captured information can also overwhelm teachers (Blomberg et al., 2013). To enhance coaching and support, some teacher educators (e.g., Chen et al., 2020; Ng et al., 2021) have first analysed classroom discourse data and then presented the discourse information (e.g., the proportions of teacher instructions and student responses) to aid teacher reflection.

However, analysing classroom discourse is time-consuming and labour-intensive and thus imposes a significant burden on teacher educators (Wang & Chen, 2024). In our PD programmes for teachers, for example, this process involves transcribing lesson recordings, coding each talk turn, and then importing the coding outcomes into our classroom discourse visualiser for teachers to review their teaching practices (see Chen, 2020 for a review). In particular, coding talk turns (which usually number over 100 in a 40-min lesson) requires independent efforts by two trained coders, comparing their coding outcomes, and resolving any discrepancies to ensure accuracy (Gamoran Sherin & van Es, 2009). This process delays feedback on teachers' performance and hinders their timely reflection, potentially impairing the efficacy of PD (Darling-Hammond et al., 2017; Wang & Chen, 2024). Accordingly, the current study sought to determine whether AI could offer a solution to the aforementioned problem. ChatGPT and GPT-4o have demonstrated considerable capabilities in processing and understanding natural language (Lo et al., 2024; Rodrigues et al., 2024; Roumeliotis & Tselikas, 2023). Thus, these AI tools may enable teacher educators to streamline the coding process and thereby provide more timely feedback than is possible with current methods.

2.2. ChatGPT and its subsequent releases

ChatGPT, fine-tuned from a model in the GPT-3.5 series (OpenAI, 2022), was used in our pilot and main studies. It generates human-like responses to user queries and improves itself by learning from user feedback (OpenAI, 2022; Ouyang et al., 2022; Roumeliotis & Tselikas, 2023). It excels in zero-shot and few-shot learning scenarios and in various natural language processing tasks, such as article generation and code writing (Lo et al., 2024; Wu et al., 2023). However, ChatGPT has

limitations, such as its tendency to provide incorrect information and lack knowledge of recent events (Lo et al., 2024; Wu et al., 2023). Nevertheless, it is free to use and thereby accessible to a broad audience, including teacher educators.

We used GPT-4 and GPT-4o in our pilot study (see Section 3.1) and main study, respectively. It was reported that GPT-4 excels in following complex instructions and achieves human-level performance on various benchmarks (OpenAI, 2023). Korkmaz Guler et al. (2024) and Plevris et al. (2023) found that GPT-4 surpasses ChatGPT in performance on certain professional examinations and in terms of accuracy in mathematical tasks. However, Bubeck et al. (2023) uncovered limitations in GPT-4's planning, working memory, and reasoning. GPT-4o, an advanced version of GPT-4, understands and generates text, audio, video, and images faster and better than its predecessors (Islam & Moushi, 2024; OpenAI, 2024; Shahriar et al., 2024). GPT-4o also offers multimodal support, exhibits superior mathematical reasoning capabilities, and demonstrates accuracy and efficiency in managing complex tasks (Shahriar et al., 2024; Zhu et al., 2024). However, the utility of GPT-4o in various application areas has remained under-evaluated, necessitating further assessment to fully understand its capabilities and limitations (Shahriar et al., 2024). Therefore, we used GPT-4o in our main study to explore its potential utility for assisting teacher educators in classroom discourse analysis.

2.3. The use of ChatGPT and its subsequent releases in qualitative analysis

ChatGPT and its subsequent releases have the potential to support human researchers in qualitative analysis processes (Hamilton et al., 2023; Zambrano et al., 2023). These AI tools excel in conducting efficient analysis, facilitating initial exploration, and providing qualitative insights (Yan et al., 2024). In coding analysis, ChatGPT has demonstrated capability in specifying codes, categories, and descriptive themes and analysing the meaning of datasets (Morgan, 2023; Sen et al., 2023). It can capture various language structures and explain its coding decisions (Zambrano et al., 2023; Zhang et al., 2024). Furthermore, Drápal et al. (2023) found that GPT-4 performed well in identifying themes from raw data in empirical legal studies. However, although AI tools have shown promise, concerns have persisted regarding their validation, accuracy, and reliability in qualitative research (Xiao et al., 2023). For example, Wang et al. (2023) reported that ChatGPT exhibited only a fair agreement with human coders in analysing online student feedback on mathematics lessons in terms of pedagogy. Similarly, Gandolfi (2024) highlighted a moderate level of inaccuracies in GPT-4's analysis of students' mathematics work. Furthermore, concerns have been expressed regarding AI tools' contextual comprehension in qualitative analysis (Hamilton et al., 2023; Hitch, 2024; Yan et al., 2024; Zhang et al., 2024).

In the context of teacher PD, some educators have developed codebooks to analyse classroom discourse data based on teacher behaviours that are targeted for reflection or improvement (see Chen et al., 2020 for a review). However, only a few studies (e.g., Misiejuk et al., 2024; Xiao et al., 2023; Zambrano et al., 2023) have provided such codebooks for AI tools to utilise in data analysis, despite the important implications from these studies for future research. Xiao et al. (2023) found that while their codebook provided transparency and explicit control, it might have constrained AI performance. Therefore, they suggested the need for more effective codebook design, which led us to adopt established coding frameworks from the literature in the current study (see Section 3.2). Misiejuk et al. (2024) explored AI coding of online discourse written by students. Their results indicated that GPT-4 agreed substantially with human coding on major categories. However, they observed that GPT-4 struggled with coding tasks requiring contextual understanding. This observation led us to adapt the prompt design proposed by Zhang et al. (2024) in the current study, with a focus on providing sufficient contextual information (see Section 3.4). Zambrano et al.

(2023) reported that based on their codebook for analysing governmental press releases and public addresses, the agreement between GPT-4 coding and human coding was not substantial (Cohen's Kappa < 0.6 ; Nili et al., 2020). Nevertheless, they demonstrated GPT-4's ability to explain its coding decisions. Likewise, in the current study, we asked ChatGPT and GPT-4o to explain their coding decisions to allow us to clearly understand the reasoning behind these decisions (see Section 3.4). Given the scarcity of comparative studies on ChatGPT and GPT-4o in qualitative analysis, particularly in coding classroom discourse data, we assessed the accuracy of these two AI tools in this domain.

3. Materials and methods

In this section, we first discuss the major findings and lessons learnt from a pilot study that we conducted to inform the main study. Subsequently, we describe the development of the codebook tailored to classroom discourse analysis, outlining its components and theoretical underpinnings. Thereafter, we elaborate on the preparation of the dataset and the establishment of the human standard, which are crucial for benchmarking the performance of AI tools. Subsequently, we present the coding process adopted in the use of ChatGPT and GPT-4o, followed by a description of the data extraction and analysis.

3.1. Major findings and lessons learnt from the pilot study

In view of the lack of studies on AI-supported classroom discourse analysis, we conducted a pilot study to test our methodology and identify its limitations to facilitate the main study. As presented in Lo et al. (in press), our pilot study analysed a 40-min synchronous online mathematics lesson on plane geometry. The lesson involved one mathematics teacher and 19 Grade 12 students. The language of instruction was Cantonese, the predominant Chinese language spoken in Hong Kong. We developed a codebook based on the Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith et al., 2014) and the framework of Academically Productive Talk (APT; Michaels et al., 2008; Resnick et al., 2010). We used the ChatGPT and GPT-4 hosted on our campus and instructed these AI tools to code each talk turn using the codebook. Their coding outcomes were then compared with a human standard. The results indicated substantial discrepancies between the human standard and both the ChatGPT (Cohen's Kappa = 0.18) and GPT-4 (Cohen's Kappa = 0.37) outputs. Although GPT-4 performed better than ChatGPT, the level of agreement in both comparisons was not satisfactory.

The lessons learnt from the pilot study led to the incorporation of five major enhancements into the main study. First, the COPUS codes were too general to inform teacher instructional improvement in the context of mathematics teaching. Therefore, we incorporated the Mathematics Discourse in Instruction (MDI) framework proposed by Adler and Ronda (2015) to better capture variations in teaching practices. Second, we found that one AI-generated code was useful for enhancing the comprehensiveness of our analysis. Specifically, the code "Affirmation" (see Section 3.2), generated by ChatGPT in the pilot study, was added to the codebook in the main study. Third, both ChatGPT and GPT-4 had difficulty recognising and classifying dialogue related to technical problems in online teaching. This limitation might have stemmed from their training data (Roumeliotis & Tselikas, 2023; Wu et al., 2023), which lacked sufficient information about synchronous online instruction because it was not widely adopted before the COVID-19 pandemic. To address this limitation, we enriched our description of the code "Fixing technical problems" (see Section 3.2) in the main study. Fourth, we asked the AI tools to provide explanations for their coding decisions to enhance the transparency and traceability of their coding outcomes (Zhang et al., 2024). This approach was previously shown to allow researchers to understand the underlying reasons for AI tools' mistakes (Sen et al., 2023). Fifth, the main study included a large dataset comprising multiple lessons across different stages of a two-month fully

online mathematics enrichment programme. We also utilised the most updated version of the GPT series, GPT-4o, to evaluate its capabilities in classroom discourse analysis. These enhancements were adopted to facilitate the main study and fully exploit the potential of the AI tools in classroom discourse analysis.

3.2. Codebook for classroom discourse analysis

As shown in Table 1, we developed our codebook to cover all teacher–student dialogues in synchronous online mathematics lessons. Based on the literature and the lessons learnt from the pilot study, the

Table 1

The codebook for classroom discourse analysis.

Code (sources)	Description	Representative quotations
Category 1. MDI codes (adopted from Adler & Ronda, 2015, pp. 10–14)		
1.1. Examples	Using examples by providing experiences of similarity, contrast, and fusion in mathematics	"Example 1: $x + x + 4x + y$," "Example 2: $x + y$," "Example 4: $2(x + y)$," ...
1.2. Tasks	Requiring students to carry out known operations, apply skills, select procedures, and make connections in mathematics	"... to simplify expressions by first factorising expressions and then cancelling common factors."
1.3. Naming	Naming conventions for words used within and across episodes, including colloquial language, mathematics words as names, and proper mathematical language	"So, over here it concerns the common factor. Why? Because we want to have one, one term at the top and one term below."
1.4. Legitimisations	Justifying concepts using localised and general criteria, considering non-mathematical factors like everyday knowledge, visual cues, and authority	"So, you just apply the same principle, it's just that when it looks complicated ..."
Category 2. APT codes (adopted from Resnick et al., 2010, p. 180)		
2.1. Say more	Asking the student(s) for further elaboration	"Say more about that."
2.2. Revoice	Re-voicing the expressions of the student(s) to prompt further responses	"So let me see if I've got your thinking right. You're saying XXX?"
2.3. Press for reasoning	Asking the student(s) to explain the reasoning	"Why do you think that?"
2.4. Challenge	Challenging the idea put forth by the student(s)	"Is this always true?"
2.5. Explain others	Asking the student(s) to explain someone else's reasoning	"Can you [explain] what he just said in your own words?"
2.6. Restate	Asking the student(s) to restate someone else's reasoning	"Can you repeat what he just said in your own words?"
2.7. Add on	Prompting other student(s) for further participation	"Would someone like to add on?"
2.8. Agree/disagree	Asking the student(s) to state the standpoint	"Do you agree or disagree and why?"
Category 3. Other codes (pilot study)		
3.1. Checking attention	Checking the attention of the student(s)	"Let's ask a student. Let me see. Johnny, are you here?"
3.2. Fixing technical problems	Addressing connectivity issues, audio or video problems, and software or platform glitches	"Hey, why is there no sound from the mic, no sound?"
3.3. Guiding and administration	Introduction, follow-up feedback on homework or tests, and assigning homework or specific tasks	"Okay, good. This is done by Roy. N equals 6. Okay, your equation is well formulated."
3.4. Student responses	The student(s) respond to the teacher	A student answered: "x, the one on the top."
3.5. Affirmation	Validating and acknowledging a student's response	"Very good! Ha, well done, Stephen! Very good!"

Note. APT = Academically productive talk; MDI = Mathematics discourse in instruction.

codebook encompassed the following categories: Category 1: MDI codes (i.e., teacher presentation of mathematics knowledge); Category 2: APT codes (i.e., teacher engagement of students in class discussions); and Category 3: Other codes (mainly everyday classroom interactions, not covered by the first two categories). In the following paragraphs, we present representative talk turns from the main study as illustrative examples.

To analyse teachers' presentation of mathematics knowledge, we adopted the MDI framework proposed by Adler and Ronda (2015). The theoretical foundation of the framework is Variation Theory, which has been widely applied in mathematics teaching (Essien, 2021; Thanheiser & Melhuish, 2019). The core idea of this theory is discernment – seeing or experiencing critical features of the object of learning (Essien, 2021). It explains how to systematically represent variations through a range of examples and hands-on practices to facilitate student learning (Thanheiser & Melhuish, 2019). Consequently, the MDI framework focuses on four elements of mathematics instruction: (1) examples, (2) tasks, (3) naming, and (4) legitimations (Category 1 in Table 1; Adler & Ronda, 2015). For example, in one of the classes used in this study, the teacher said, “ m_{OA} multiplied by m_{AA} equals $5/2$ multiplied by $-3/7$, which simplifies to $-15/14$. This result is important because it is not equal to -1 . Therefore, they are not perpendicular” (Lesson 3 Turn 93). This turn was coded as “Legitimations” because the teacher justified the conclusion based on mathematical calculations and criteria. The MDI framework is useful in teacher PD because it provides structured insights into effective mathematics instruction (Adler & Ronda, 2015).

To analyse how the teacher engaged students in class discussions, we adopted the APT framework proposed by Michaels et al. (2008) and Resnick et al. (2010). The theoretical foundation of this framework lies in Sociocultural Theory, which suggests that learning is a social process (Vygotsky, 1978). In the educational context, this social process involves interaction between teachers and students through dialogue. The framework consists of eight types of talk moves that increase the likelihood of students' participation in knowledge construction: (1) say more, (2) re-voice, (3) press for reasoning, (4) challenge, (5) explain others, (6) restate, (7) add on, and (8) agree/disagree (Category 2 in Table 1; Resnick et al., 2010). For example, our teacher asked, “Steve [a pseudonym of a student], can you explain why you took 160 as the denominator?” (Lesson 2 Turn 38). This turn was coded as “Press for reasoning” because the teacher asked the student to explain his reasoning for using 160 as the denominator in his calculation. The APT framework is valuable in teacher PD because it provides clear strategies to foster student participation and enables teachers to assess whether their students have acquired the knowledge the teachers communicated (Chen et al., 2020; Ng et al., 2021).

In addition to the talk turns represented by the MDI and APT codes, other types of talk turns identified in the pilot study were incorporated into the main study, as they facilitated the classroom discourse analysis. As shown in Category 3 in Table 1, these additional codes were (1) checking attention, (2) fixing technical problems, (3) guiding and administration, (4) student responses, and (5) affirmation. For example, the online teaching context necessitated the inclusion of unique codes, such as “Checking attention” and “Fixing technical problems.” In Lesson 1, the teacher said, “Oh, someone has sent me a chat message. I should have shared my screen. I am going to share it again” (Turn 57). This turn was coded as “Fixing technical problems” because the teacher addressed a technical issue that arose during the online lesson. These other codes are important to facilitate teacher reflection on classroom management during synchronous online lessons.

3.3. Preparation of the dataset and the human standard

We used three 45-min lesson recordings from a two-month synchronous online mathematics enrichment programme at a Hong Kong secondary school. These lessons (Lessons 1 to 3) represented the start, middle, and end stages of the programme and were aimed at helping

struggling students overcome deficiencies in their learning. The programme involved one mathematics teacher and 15 students. As in the pilot study, the language of instruction was Cantonese. The lesson recordings were transcribed into Chinese. Table 2 presents the summary characteristics of the lessons. The dataset comprised 24,166 words and 414 talk turns. Due to the token limit for each prompt in the interface for the ChatGPT and GPT-4o hosted on our campus, we divided the lesson recordings into episodes based on the content of the classroom discourse (see Section 3.4). The three lessons consisted of 31 episodes in total.

To establish the human standard for the coding of the lessons, we adopted the negotiated agreement approach described by Campbell et al. (2013). In this approach, “two or more researchers code a transcript, compare codings, and then discuss their disagreements in an effort to reconcile them and arrive at a final version in which as many discrepancies as possible have been resolved” (p. 305). Specifically, two trained human coders (the first two authors) independently coded the dataset using the codebook developed in Section 3.2. Their coding outcomes were then compared and were found to demonstrate almost perfect agreement (Cohen's Kappa = 0.98; Nili et al., 2020; Viera & Garrett, 2005). In keeping with the final step of the methodology of Campbell et al. (2013), the discrepancies were resolved through discussion. The finalised human coding results, which served as the ground truth, were then compared with the outcomes generated by ChatGPT and GPT-4o.

3.4. Coding process in the use of ChatGPT and GPT-4o

We utilised two chatbot-based large language models, ChatGPT and GPT-4o, provided by the first author's university through the Microsoft Azure OpenAI service (Office of the Chief Information Officer, 2024). This service provides functionality similar to that of OpenAI's ChatGPT and GPT-4o. However, because it is hosted on campus, the conversation history could be deleted at our discretion and would not be used for the training of future AI models, thereby ensuring privacy.

The prompts input to AI tools can significantly impact the quality, coherence, and applicability of their outputs. We adapted the prompt design for thematic analysis proposed by Zhang et al. (2024). The processes adopted for coding with ChatGPT and GPT-4o were identical and the coding was conducted in early June 2024. As shown in Table 3, the prompt design consisted of seven components that helped the AI tools gain the comprehensive understanding needed to conduct the analysis. The prompts provided the AI tools with the necessary background information, instructions on the task to be performed, and the codebook detailing the codes and their corresponding definitions (see Table 1). Each component of the prompt design served to systematically introduce the task to the AI tools and maximise their output quality. For example, Zhang et al. (2024) stated that the role-playing prompt allows ChatGPT to focus effectively on completing a specific type of task. Consequently, we assigned the AI tools the role of a research expert in qualitative analysis. In subsequent prompts, we copied and pasted the discourse data into the input box of each AI tool episode by episode. Fig. 1 shows the output of ChatGPT in coding Episode 1 (Turns 1 to 12) in Lesson 1.

3.5. Data extraction and analysis

The coding outcomes generated by ChatGPT and GPT-4o were extracted and stored in two separate spreadsheets. This process involved

Table 2
Characteristics of the three lessons.

Lesson	Topic	Word counts	Talk turns	Episodes
1	Percentages	8282	161	11
2	Inequalities	8497	152	11
3	Coordinate geometry	7387	101	9
Total		24,166	414	31

Table 3
Components of the prompt design and instructions for the AI tools.

Component	Instruction
1. Role-playing	You are now a research expert in qualitative analysis.
2. Goal of the task	The goal is to code the classroom discourse text data based on a predefined codebook in order to understand the instructions and interactions within these lessons.
3. Background/Conceptual understanding	The dataset consists of Cantonese-language mathematics classes conducted online in Hong Kong. The lessons cover the topics of Percentages, Inequalities, and Coordinate Geometry.
4. Focus on analytical process	Please conduct a thematic analysis and code each talk turn using only one code from the codebook provided below. ... (Note. Table 1 was inserted here) ...
5. Transparency and traceability	After assigning each code, provide a brief explanation of your code.
6. Data format (inputs)	My input format looks like this: [Turn number] [Role] [Classroom talk]
7. Data format (outputs)	Your output format looks like this: [Turn number] [Your assigned code] [Your explanation]

identifying and recording the codes assigned by each AI tool to each talk turn. To address RQ1, the descriptive statistics of the code distribution were first used to provide an overview of their coding outcomes compared with the human standard. To understand the extent to which human can delegate specific tasks to AI, we conducted a quantitative analysis at two levels: the category level and the code level. At the category level, we examined whether the AI tools classified individual talk turns into correct categories (i.e., MDI, APT, and Other codes). At the code level, we assessed whether the AI tools assigned correct codes to individual talk turns. As suggested by Nili et al. (2020), Fleiss's Kappa

was used to determine the agreement level among multiple coders simultaneously (i.e., the human standard, ChatGPT, and GPT-4o). For pairwise comparisons, Cohen's Kappa was used to determine the agreement level between (1) the human standard and ChatGPT coding, (2) the human standard and GPT-4o coding, and (3) ChatGPT and GPT-4o coding. When interpreting the Kappa statistics, we used the benchmarks defined by Viera and Garrett (2005), in which 0 or below indicates poor agreement, 0.01 to 0.20 indicates slight agreement, 0.21 to 0.40 indicates fair agreement, 0.41 to 0.59 indicates moderate agreement, 0.60 to 0.79 indicates substantial agreement, and 0.80 to 0.99 indicates almost perfect agreement.

To address RQ2 regarding the alignment and discrepancies between the outcomes of the AI tools and the human standard in coding the classroom discourse, we followed the qualitative analysis procedure proposed by Braun and Clarke (2006). First, we familiarised ourselves with the coding outcomes obtained using ChatGPT and GPT-4o and how they compared with the human standard. Subsequently, we generated initial codes to categorise instances in which the AI tools performed well and those in which discrepancies emerged. We then searched for themes across these codes, focusing on identifying the competences and weaknesses of the AI tools. After reviewing these themes to ensure that they were consistent and relevant to RQ2, each theme was defined to capture the key findings, culminating in the identification of three major competences and three major weaknesses (see Section 4.2).



- 1 3.3 Guiding and administration: The teacher is providing information about the class, introducing themselves, and explaining how students can ask questions if they do not understand.
- 2 3.3 Guiding and administration: The teacher is discussing attendance and test materials with the students.
- 3 1.1 Examples: The teacher introduces the topic of percentages and explains its relevance to the DSE exam.
- 4 2.7 Add on: The teacher explains the benefits of participating in answering questions and invites students to engage in the process.
- 5 3.2 Fixing technical problems: The teacher addresses technical issues with students' microphones and instructs them on how to communicate if their microphone is not working.
- 6 3.3 Guiding and administration: The teacher instructs students to refer to their notebooks for notes on the lesson.
- 7 3.3 Guiding and administration: The teacher checks if students have the correct page in their notebooks for the lesson on percentages.
- 8 2.8 Agree/disagree: The teacher randomly selects a student and asks if they are present in the class.
- 9 3.3 Guiding and administration: The teacher checks if students have the correct page in their textbooks for the lesson.
- 10 4 Student responses: A student responds to the teacher's question about having the correct page in their textbook.
- 11 3.3 Guiding and administration: The teacher instructs students to find paper to write on if they do not have the required page in their textbook.
- 12 3.3 Guiding and administration: The teacher informs students about upcoming exercises and how to submit their answers for feedback.

Fig. 1. Coding output of ChatGPT for Episode 1 (Turns 1 to 12) in Lesson 1.

4. Results

4.1. RQ1: what is the level of agreement between the human standard, ChatGPT, and GPT-4o in coding discourse data from synchronous online mathematics lessons?

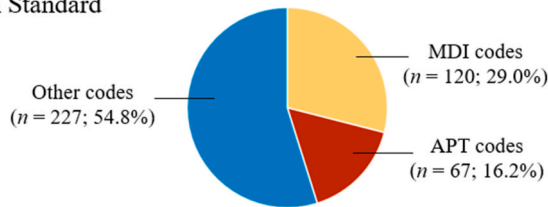
This section addresses RQ1 in two stages. First, we present the descriptive statistics of the code distribution to provide an overview of how each coder categorised the classroom discourse data. Second, we analyse the Kappa statistics to quantitatively compare the coding agreement between the human standard and the AI tools at the category and code levels.

4.1.1. Descriptive statistics of the code distribution

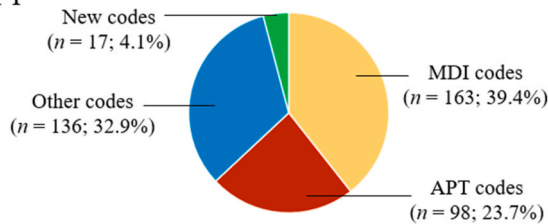
Fig. 2 presents the code distributions of the human standard and the ChatGPT and GPT-4o outputs by category. Compared with the code distribution of ChatGPT, that of GPT-4o more closely resembled that of the human standard. In both the human standard and GPT-4o output, the MDI category accounted for approximately 30% of the codes, whereas ChatGPT assigned nearly 40% of its codes from this category. Similarly, in both the human standard and GPT-4o output, the APT category accounted for around 15% of the codes, whereas ChatGPT assigned over 23% of its codes from this category. Table 4 provides a detailed breakdown of the frequency of each code assigned by ChatGPT and GPT-4o compared with that in the human standard. Within every code category, the AI tool outputs differed from the human standard in terms of their emphasis on each code. For example, regarding the APT code category, the human standard primarily used the code “Say more,” whereas ChatGPT and GPT-4o most frequently used “Explain others” and “Press for reasoning,” respectively.

In addition to the codes in the three main categories outlined in the codebook, both ChatGPT and GPT-4o generated new codes during their coding process. ChatGPT created five new codes, assigning them to 17 talk turns (4.1% of the total). These new codes were “Building rapport” ($n = 2$), “Clarification” ($n = 7$), “Clarifications” ($n = 3$), “Correction” ($n = 4$), and “Providing information” ($n = 1$). However, “Clarification” and

(a) Human Standard



(b) ChatGPT



(c) GPT-4o

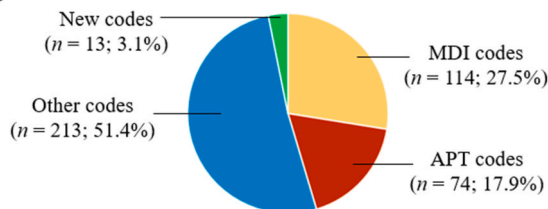


Fig. 2. Code distributions for (a) the human standard, (b) ChatGPT, and (c) GPT-4o by category.

Table 4

Frequency distribution of codes assigned by ChatGPT and GPT-4o compared with that of the human standard.

Code	Human standard	ChatGPT	GPT-4o
Category 1. MDI codes	120 (29.0%)	163 (39.4%)	114 (27.5%)
1.1. Examples	0 (0%)	22 (5.3%)	36 (8.7%)
1.2. Tasks	27 (6.5%)	102 (24.6%)	34 (8.2%)
1.3. Naming	27 (6.5%)	21 (5.1%)	16 (3.9%)
1.4. Legitimations	66 (15.9%)	18 (4.3%)	28 (6.8%)
Category 2. APT codes	67 (16.2%)	98 (23.7%)	74 (17.9%)
2.1. Say more	59 (14.2%)	1 (0.2%)	20 (4.8%)
2.2. Re-voice	3 (0.7%)	1 (0.2%)	0 (0%)
2.3. Press for reasoning	3 (0.7%)	2 (0.5%)	40 (9.7%)
2.4. Challenge	0 (0%)	10 (2.4%)	0 (0%)
2.5. Explain others	0 (0%)	40 (9.7%)	2 (0.5%)
2.6. Restate	0 (0%)	7 (1.7%)	0 (0%)
2.7. Add on	1 (0.2%)	32 (7.7%)	10 (2.4%)
2.8. Agree/disagree	1 (0.2%)	5 (1.2%)	2 (0.5%)
Category 3. Other codes	227 (54.8%)	136 (32.9%)	213 (51.4%)
3.1. Checking attention	27 (6.5%)	21 (5.1%)	29 (7.0%)
3.2. Fixing technical problems	3 (0.7%)	3 (0.7%)	5 (1.2%)
3.3. Guiding and administration	72 (17.4%)	24 (5.8%)	36 (8.7%)
3.4. Student responses	83 (20.0%)	29 (7.0%)	83 (20.0%)
3.5. Affirmation	42 (10.1%)	59 (14.3%)	60 (14.5%)
New codes created by AI coders	N.A.	17 (4.1%)	13 (3.1%)
Total	414 (100%)	414 (100%)	414 (100%)

“Clarifications” appeared to be redundant. GPT-4o created four new codes, assigning them to 13 talk turns (3.1% of the total). These new codes were “Checking understanding” ($n = 6$), “Introduction of new concept” ($n = 3$), “Open floor” ($n = 3$), and “Repetition” ($n = 1$). As can be seen, the set of new codes generated by ChatGPT was distinct from that generated by GPT-4o, indicating these AI tools’ divergent approaches in identifying additional codes beyond the prescribed codebook.

4.1.2. Kappa statistics comparing the human standard with ChatGPT and GPT-4o outputs

We first evaluated the coding agreement between the human standard, the ChatGPT output, and the GPT-4o output at the category level. There were three categories (MDI, APT, and Other codes; see Table 1) in the codebook, together with a “new codes” category comprising codes generated by the AI tools. Table 5 shows that the overall agreement between the three coders at the category level was moderate (Fleiss’s Kappa = 0.46). Pairwise comparisons revealed that the agreement between the human standard and GPT-4o output was substantial (Cohen’s Kappa = 0.69), indicating the capacity of GPT-4o to classify talk turns into correct categories. In contrast, the agreements between the human standard and ChatGPT output (Cohen’s Kappa = 0.33) and between the ChatGPT output and GPT-4o output (Cohen’s Kappa = 0.40) were fair. The Sankey diagrams in Figs. 3 and 4 visualise the performances of ChatGPT and GPT-4o, respectively, by mapping their coded classroom discourse data to the human standard. Compared with the ChatGPT output (Fig. 3), the GPT-4o output had fewer deviations from the human

Table 5

Coding agreement between the human standard, ChatGPT output, and GPT-4o output at the category level.

Comparison	Kappa statistics (level of agreement)
Overall comparison (Fleiss’s Kappa)	0.46 (moderate)
Pairwise comparison (Cohen’s Kappa)	
● Human standard vs. ChatGPT	0.33 (fair)
● Human standard vs. GPT-4o	0.69 (substantial)
● ChatGPT vs. GPT-4o	0.40 (fair)

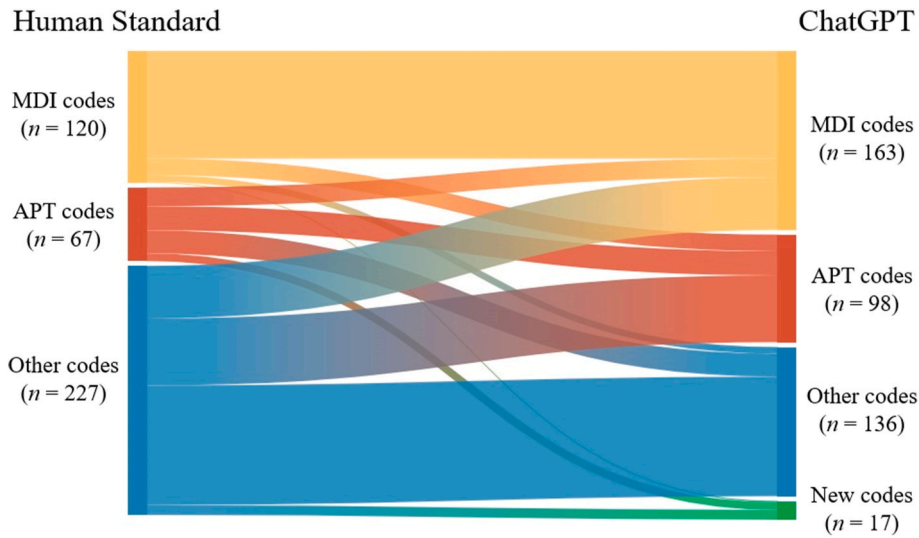


Fig. 3. Sankey diagram of the human standard (left) vs. ChatGPT coding (right) by category.

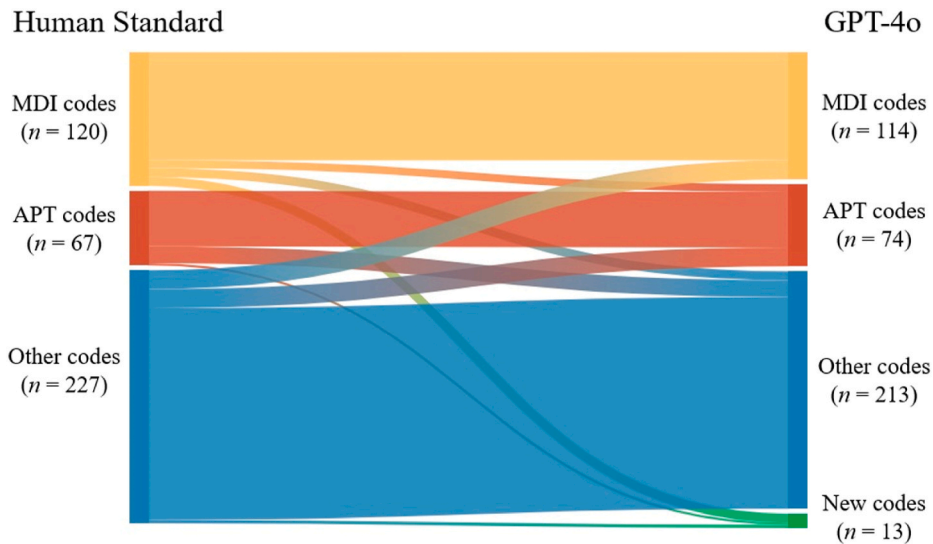


Fig. 4. Sankey diagram of the human standard (left) vs. GPT-4o coding (right) by category.

standard (Fig. 4).

Next, we explored the code level to examine the coding agreement between the human standard, ChatGPT output, and GPT-4o output in detail. There were 17 predefined codes (four MDI codes, eight APT codes, and five Other codes; see Table 1), together with nine new codes generated by the AI tools. Table 6 shows that the overall agreement between the three coders at the code level decreased to fair (Fleiss's Kappa = 0.31), indicating inconsistency in the assignment of individual codes to talk turns. Within the MDI and APT code categories, the three coders showed slight (Fleiss's Kappa = 0.07) and poor (Fleiss's Kappa = -0.07) agreement, respectively. The Others category of codes achieved fair agreement (Fleiss's Kappa = 0.40), reflecting that the AI tools had some capability to recognise key aspects of everyday classroom interactions. In pairwise comparisons, the overall agreement between the human standard and GPT-4o output was moderate (Cohen's Kappa = 0.48), whereas the agreements between the other two pairs were fair. Within both the MDI and APT code categories, the agreements between the human standard and the two AI tool outputs were slight. Within the Other code category, however, the agreement between the human standard and GPT-4o output was substantial (Cohen's Kappa = 0.71) and that between the human standard and ChatGPT output was fair

Table 6

Coding agreement between the human standard, ChatGPT output, and GPT-4o output at the code level.

Comparison	Codes			
	Kappa statistics (level of agreement)			
	All codes (n = 414)	MDI codes (n = 120)	APT codes (n = 67)	Other codes (n = 227)
Overall comparison (Fleiss's Kappa)	0.31 (fair)	0.07 (slight)	-0.07 (poor)	0.40 (fair)
Pairwise comparison (Cohen's Kappa)				
● Human standard vs. ChatGPT	0.21 (fair)	0.03 (slight)	0.03 (slight)	0.28 (fair)
● Human standard vs. GPT-4o	0.48 (moderate)	0.05 (slight)	0.06 (slight)	0.71 (substantial)
● ChatGPT vs. GPT-4o	0.27 (fair)	0.24 (fair)	0.08 (slight)	0.29 (fair)

(Cohen's Kappa = 0.28).

4.2. RQ2: how do the coding outcomes of the AI tools align with or differ from the human standard, and what are the underlying reasons for differences, if any?

This section explores the coding agreements and discrepancies between the human standard and the AI tool outputs. We examined (1) instances in which the AI tools performed well, (2) instances in which GPT-4o performed better than ChatGPT, and (3) instances in which the AI tools did not perform effectively. Table 7 provides an overview of the major identified competences and weaknesses of the AI tools in classroom discourse analysis.

4.2.1. Instances in which the AI tools performed well

In some talk turns, the codes assigned by ChatGPT and GPT-4o aligned with the human standard. Most of these codes ($n = 72$) pertained to everyday classroom interactions (Competence 1) within the Others category of codes, such as "Affirmation," "Student responses," and "Guiding and administration." For example, both AI tools correctly assigned the code "Affirmation" to the teacher's statement, "The Pythagorean theorem, yes, that is right ..." (Lesson 3 Turn 24). The human coders identified that this statement showed the teacher affirming a student's answer. ChatGPT described it as "the teacher affirms Student 2's response and praises their correct answer," whereas GPT-4o explained that "the teacher acknowledges and validates the student's correct response." In another example, Student 6 provided a short numerical answer, "153" (Lesson 1 Turn 93), and both AI tools categorised the turn as "Student responses." ChatGPT noted that "Student 6 responded with the answer 153," and GPT-4o remarked that "the student responded with the result 153." These examples illustrate that both AI tools were capable of accurately categorising certain everyday classroom interactions.

4.2.2. Instances in which GPT-4o performed better than ChatGPT

GPT-4o demonstrated superior competence in classifying broad categories (Competence 2). At the category level, the agreement between the human standard and GPT-4o output was substantial, whereas the agreement between the human standard and ChatGPT output was fair (Table 5). For example, in Lesson 1 Turn 122, the teacher asked a student to elaborate on the next step: "40% more than the females, so what should we do next?" In alignment with the human standard, GPT-4o assigned this talk turn the APT code "Say more." However, ChatGPT assigned this turn the MDI code "Tasks," describing it as "the teacher instructs on the next steps to be taken in solving the equation involving the male members." Similarly, regarding the talk turn "Okay. Following this, we have to subtract something. Which one should we subtract?" (Lesson 2 Turn 30), which was a "Say more" talk turn, GPT-4o assigned different APT code, "Press for reasoning." However, ChatGPT assigned it the MDI code "Tasks." Clearly, in both the talk turns, the teacher did not prompt the students to perform familiar operations or apply skills.

Table 7

Major competences and weaknesses identified in ChatGPT and GPT-4o for classroom discourse analysis.

Performance	Description	ChatGPT	GPT-4o
Competence 1	Identification of everyday classroom interactions	✓	✓
Competence 2	Recognition of main categories	×	✓
Competence 3	Recognition of speaker roles	×	✓
Weakness 1	Lack of pedagogical content knowledge	✓	✓
Weakness 2	Misinterpretation of code definitions	✓	✓
Weakness 3	Creation of unnecessary codes	✓	✓

Instead, he requested them to say more, which, in the second instance, GPT-4o at least correctly interpreted as an APT move to elicit students' responses.

GPT-4o also demonstrated superior capacity in recognising roles (Competence 3). GPT-4o correctly identified the student and teacher roles in the turns, whereas ChatGPT misinterpreted the roles in some turns. For example, in Lesson 3 Turns 68 to 69, the teacher asked a student, "How did you get the 4?" The student responded, "Add 6." GPT-4o assigned the same codes to these turns as the human standard, categorising Turn 68 as "Press for reasoning" and the next turn as "Student responses." It explained that "the teacher asks the student to explain how they derived the new y-coordinate" (Turn 68) and "the student explains that the new y-coordinate is found by adding 6 to the original y-coordinate" (Turn 69). In contrast, ChatGPT incorrectly categorised Turn 68 as "Student responses," stating that "Student 4 explains that the value changes by adding 6," and assigned the next turn the code "Affirmation," explaining that "the teacher acknowledges Student 4's explanation and thanks them for their contribution." This finding suggests that ChatGPT occasionally had difficulty distinguishing roles within a conversation.

4.2.3. Instances in which the AI tools did not perform well

We identified three major weaknesses of the AI tools: (1) lack of pedagogical content knowledge, (2) misinterpretation of code definitions, and (3) creation of unnecessary codes. First, although the AI tools correctly identified a range of mathematical concepts, they lacked an understanding of how these concepts were presented in teaching practices. This limited pedagogical content knowledge led to inaccuracies in assigning the MDI codes. For example, in Lesson 2 Turn 54, the teacher instructed the students, "If the two inequalities are in the same direction, that is, x is greater than a and simultaneously x is greater than b , to satisfy both conditions, there can only be one range. So, the final solution would take the one in which x is greater than b ." The human coders coded this talk turn as "Legitimations," suggesting that "the teacher was justifying concepts associated with compound inequalities." However, ChatGPT coded this talk turn as "Tasks," explaining that "the teacher introduces another case for compound inequalities with two inequalities in the same direction, explaining the solution process for 'and' situations." Similarly, GPT-4o mistakenly coded this talk turn as "Examples," describing it as "the teacher is presenting another example to explain a different case with 'and' conditions, showing the overlap and how to represent the solution." Despite demonstrating an understanding of the mathematical concept, the AI tools failed to comprehend the teacher's purpose of justifying concepts within compound inequalities (i.e., "Legitimations") rather than requiring students to perform "tasks," as indicated by ChatGPT, or presenting "Examples," as indicated by GPT-4o. This gap between their knowledge of pedagogical content and its application in teaching practices led to issues in their assignment of codes.

Second, both AI tools misinterpreted some of the code definitions. For example, in Lesson 3 Turn 40, the teacher checked a student's attention by asking, "Okay, this is just a random question. There is no preference for whom to ask. I will ask ... is Rebecca [a pseudonym for the student] here?" Instead of categorising this talk turn as "Checking attention," both ChatGPT and GPT-4o coded this turn as "Add on." ChatGPT described the turn as "the teacher randomly asks a student, Rebecca [a pseudonym of the student], if they are present in the class for a question," whereas GPT-4o explained it as "the teacher randomly selects a student to answer the question about the coordinates of B'." According to the codebook, "Add on" means prompting other student(s) to participate further. However, the teacher's question served as a prompt to check a student's attention rather than to elicit further responses from other students. These findings indicated that the AI tools were limited in their ability to fully interpret the code definitions, leading to inaccurate code assignments.

Third, the AI tools generated additional codes that were very similar

to those defined in the codebook. ChatGPT and GPT-4o produced five and four new codes, respectively. For example, in Lesson 2 Turn 80, the teacher asked, “Yes, can you further explain one more time?” ChatGPT categorised this talk turn as “Clarification,” explaining that “the teacher asks Student 4 to repeat their explanation for clarity.” GPT-4o assigned this talk turn the code “Repetition,” stating that “the teacher asks the student to repeat their suggestion, ensuring clarity.” These codes and descriptions reflected their understanding of the talk turn in which the teacher requested the student to repeat or clarify by providing further explanation. However, their new codes, “Clarification” and “Repetition,” were similar to the ATP code “Say more.” These findings suggested that the AI tools overlooked relevant codes covered in the prescribed codebook and created unnecessary new codes.

5. Discussion

This study examined the performance of ChatGPT and GPT-4o in analysing classroom discourse data. In this section, we first highlight the areas in which these AI tools can assist researchers and teacher educators in the context of teacher PD. Next, we explore the current shortcomings of these AI tools in this domain. Finally, we discuss the implications and limitations of our study and provide recommendations for future research.

5.1. Areas in which AI tools can assist in analysing classroom discourse data

Recognising everyday classroom interactions: Both of the AI tools, particularly GPT-4o, demonstrated competence in recognising and coding everyday classroom interactions (see Section 4.2.1). This competence is significant because it allows for the efficient analysis of classroom routines. Consistent with findings in other research areas (e.g., Morgan, 2023; Sen et al., 2023), the AI tools were able to qualitatively analyse the meaning of the classroom discourse data. Both AI tools understood Cantonese and provided explanations for their coding decisions (Zambrano et al., 2023; Zhang et al., 2024). These competences can aid teacher educators in PD contexts by automating the coding of everyday classroom interactions. Thus, despite these AI tools’ limitations in understanding deeper pedagogical content knowledge (see Section 4.2.3), their ability to handle routine interaction coding can free up teacher educators’ time, allowing them to focus on interpreting PD-specific data related to how teachers present knowledge and engage students in academically productive discussions.

Classifying broad categories: When classifying broad categories of classroom discourse, GPT-4o demonstrated substantial agreement with the human standard, showcasing more advanced capabilities than ChatGPT (see Table 5). Specifically, GPT-4o effectively classified talk turns into the three prescribed categories, namely MDI (Adler & Ronda, 2015), APT (Resnick et al., 2010), and Others. This observation aligns with the findings of Misiejuk et al. (2024), who found that GPT-4 was reasonably reliable in coding students’ online written discourse. In their study, GPT-4 showed substantial agreement with human coding for various coding categories, such as “Theory”, “Integration”, and “Personal.” In the current study, we found that GPT-4o excelled in identifying whether a talk turn was from a teacher or a student (see Section 4.2.2), whereas ChatGPT occasionally confused these roles, consistent with the findings of Sen et al. (2023). GPT-4o’s ability to correctly classify talk turns into broad categories and identify roles can significantly facilitate the initial analysis of classroom discourse data. This automated initial analysis by GPT-4o can provide teacher educators with an overview of the distribution of talk turns by category, making it a valuable tool in the context of teacher PD.

5.2. Areas in which the AI tools cannot yet assist in analysing classroom discourse data

Inadequate performance at the code level: Although GPT-4o can assist teacher educators in category-level coding, it is still significantly limited in its performance at the code level. The results of this study indicate that both ChatGPT and GPT-4o struggled with assigning correct, specific MDI and APT codes (see Table 6). These findings contrast with the more encouraging results reported in the literature regarding the AI tools’ ability to conduct thematic analysis (e.g., Misiejuk et al., 2024; Zambrano et al., 2023). The current study provided evidence that at the code level, the performance of both AI tools was less promising, as they failed to understand how pedagogical content knowledge was presented in teaching practices (see Section 4.2.3). This finding aligns with that of Wang et al. (2023), who noted that ChatGPT demonstrated only fair agreement with human codes when assessing pedagogy in student feedback on mathematics lessons.

Generating unnecessary new codes: Another drawback observed in using ChatGPT and GPT-4o for classroom discourse analysis was their tendency to generate unnecessary new codes, despite clear instructions to strictly adhere to the prescribed codebook (see Table 3). This finding is consistent with concerns highlighted by Yan et al. (2024), whose participants noted instances of ChatGPT generating hallucinated results (i.e., results that are not based on fact and are misleading) and fabricated codes. In the current study, we found that GPT-4o also introduced redundant or duplicate codes that did not align with the predefined MDI and APT frameworks for classroom discourse analysis. Such deviations may confuse teachers and divert attention from the intended goal of improving targeted teaching behaviour, such as using APT turns to engage students in class discussions (see Chen et al., 2020 for a review).

5.3. Implications, limitations, and recommendations for future research

Building on the seminal work by Zhang et al. (2024), the current study generated an enriched seven-component framework for prompt design and instructions for AI tools (see Table 3). This framework guides AI tools in generating relevant outputs by providing contextual information and clearly structuring human–AI interactions. Researchers can modify the prompts to suit their own contexts and specific needs, thereby ensuring the utility and performance of AI tools across diverse settings.

Based on our comparison of the performance of ChatGPT and GPT-4o, we recommend using GPT-4o to facilitate classroom discourse analysis for teacher educators in PD contexts. Our proposed approach involves a structured two-stage process to exploit the capabilities of GPT-4o without requiring advanced AI knowledge or techniques. First, teacher educators can use GPT-4o for category-level analysis, thereby providing teachers with timely resources for reflection. However, it is essential to validate any new AI-generated codes before presenting the analytics for teachers to review at this stage. Following this, “a hybrid approach that integrates human judgement” (Misiejuk et al., 2024, p. 1) should be applied. At this stage, teacher educators can use GPT-4o to conduct code-level analysis and update the coding outcomes to support deeper teacher reflection. While the current study focused on teacher PD, the potential utility of the two-stage approach extends to other contexts involving qualitative analysis. For example, Misiejuk et al. (2024) demonstrated that AI tools could assist in classifying discussion forum contributions into main themes, such as group engagement. As a potential next step, researchers could collaborate with AI tools to identify sub-themes, such as “constructive criticism,” “compliments for previous postings,” or “questions that lead to more discussions” (p. 5), thereby further enriching analysis.

Finally, this study had several limitations that must be acknowledged. First, our investigation was restricted to synchronous online mathematics instruction. Although this enabled us to garner valuable insights into how ChatGPT and GPT-4o handles classroom discourse

data in this subject domain and context, it limits the generalisability of our results. Different disciplines and instructional environments may require distinct coding frameworks and pedagogical understanding, which could influence the performance of the AI tools. Second, the study was conducted within a Chinese classroom context. Cultural and contextual factors unique to this setting may have impacted the AI tools' performance. Third, we established and utilised a specific codebook based on the literature to guide discourse analysis. However, the AI tools' performance may vary with the use of different codebooks tailored to other educational frameworks or subjects. To address the first three limitations, future research should incorporate a broader range of subjects and educational contexts to provide a more comprehensive evaluation of the capabilities of AI tools in classroom discourse analysis. Fourth, we did not examine multiple outputs from the AI tools and only refined queries between the pilot and main studies to enhance output quality. Future research can explore the consistency and accuracy of varied responses generated by the same AI tools and engage interactively with these tools during a coding process.

6. Conclusion

This study evaluated the performance of ChatGPT and GPT-4o in analysing classroom discourse data within the context of synchronous online mathematics instruction. Our findings revealed that GPT-4o outperformed ChatGPT and demonstrated potential utility for classifying broad categories of talk turns. However, significant limitations in these AI tools' outputs were identified at the code level, and they generated unnecessary new codes. Based on these insights, we recommend utilising GPT-4o for initial category-level analysis, followed by manual verification and refinement of the output at the code level to ensure accuracy and relevance. While this two-stage approach may not necessarily improve the coding quality of teacher educators, it can accelerate and enhance the efficiency of the coding process. Specifically, an AI tool reduces the burden of categorising broad themes and allows teacher educators to focus on refining and verifying AI-generated codes rather than requiring them to prepare codes from scratch. Most importantly, this two-stage process can enhance the effectiveness of teacher PD, as it first provides teachers with a timely initial overview and then facilitates their deeper reflection on teaching practices. Future studies should extend this research beyond the mathematics domain and Chinese classroom context to explore the applicability and performance of these AI tools in diverse subject disciplines and educational settings.

CRedit authorship contribution statement

Simin Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Xiaowei Huang:** Writing – review & editing, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Chung Kwan Lo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization. **Gaowei Chen:** Writing – review & editing. **Morris Siu-yung Jong:** Writing – review & editing.

Statements on open data and ethics

The intervention was approved by an ethics committee (ID: A2020-2021-0110). Informed consent was obtained from all participants, and their privacy was strictly preserved. The data can be obtained by emailing a request to the corresponding author.

Declaration of competing interest

There is no potential conflict of interest in this study.

Acknowledgments

This work described in this paper was substantially supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China (Project No. EdUHK 28604623) and by Department of Mathematics and Information Technology (Departmental Research Grant; MIT/DRG02/24–25), The Education University of Hong Kong.

References

- Adler, J., & Ronda, E. (2015). A framework for describing mathematics discourse in instruction and interpreting differences in teaching. *African Journal of Research in Mathematics, Science and Technology Education*, 19(3), 237–254. <https://doi.org/10.1080/10288457.2015.1089677>
- Blomberg, G., Renkl, A., Gamoran Sherin, M., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal for Educational Research Online*, 5(1), 90–114. <https://doi.org/10.25656/01:8021>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* <https://doi.org/10.48550/arXiv.2303.12712>
- Campbell, J. L., Quincy, C., Osseman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42(3), 294–320. <https://doi.org/10.1177/0049124113500475>
- 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. K., Chan, K. K. H., Clarke, S. N., & Resnick, L. B. (2020). Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. *The Journal of the Learning Sciences*, 29(4–5), 642–680. <https://doi.org/10.1080/10508406.2020.1783269>
- Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. Palo Alto, CA.
- Drápal, J., Westermann, H., & Savelka, J. (2023). Using large language models to support thematic analysis in empirical legal studies. *arXiv preprint arXiv:2310.18729* <https://doi.org/10.48550/arXiv.2310.18729>
- Essien, A. A. (2021). Understanding the choice and use of examples in mathematics teacher education multilingual classrooms. *ZDM—Mathematics Education*, 53(2), 475–488. <https://doi.org/10.1007/s11858-021-01241-6>
- Gamoran Sherin, M., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, 60(1), 20–37. <https://doi.org/10.1177/0022487108328155>
- Gandolfi, A. (2024). GPT-4 in Education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *Advance online publication International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00403-3>
- Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review*, 16, 41–67. <https://doi.org/10.1016/j.edurev.2015.06.001>
- Gutentag, T., Orner, A., & Asterhan, C. S. C. (2022). Classroom discussion practices in online remote secondary school settings during COVID-19. *Computers in Human Behavior*, 132, Article 107250. <https://doi.org/10.1016/j.chb.2022.107250>
- Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International journal of qualitative methods*. Advance online publication. <https://doi.org/10.1177/16094069231201504>
- Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research*, 34(7), 595–606. <https://doi.org/10.1177/10497323231217392>
- Islam, R., & Moushi, O. M. (2024). GPT-4o: The cutting-edge advancement in multimodal LLM. *Authorea Preprints*. Retrieved from <https://easychair.org/publications/preprint/download/z4TJ>
- Jaekel, A. K., Fütterer, T., & Göllner, R. (2023). Teaching characteristics in distance education—associations with teaching quality and students' learning experiences. *Teaching and Teacher Education*, 132, Article 104174. <https://doi.org/10.1016/j.tate.2023.104174>
- Korkmaz Guler, N. K., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research*, 9(2), Article em0188. <https://doi.org/10.29333/pr/14145>
- Larison, S., Richards, J., & Sherin, M. G. (2024). Tools for supporting teacher noticing about classroom video in online professional development. *Journal of Mathematics Teacher Education*, 27(2), 139–161. <https://doi.org/10.1007/s10857-022-09554-3>
- Lo, C. K., Hew, K. F., & Jong, M. S. Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, 219, Article 105100. <https://doi.org/10.1016/j.compedu.2024.105100>

- Lo, C. K., & Liu, K. Y. (2022). How to sustain quality education in a fully online environment: A qualitative study of students' perceptions and suggestions. *Sustainability*, 14, 5112. <https://doi.org/10.3390/su14095112>
- Lo, C. K., Xu, S., & Chen, G. (in press). An exploratory study of using AI tools to analyse classroom discourse data. In K. Nakamatsu, R. Kountcheva, & S. Patnaik (Eds.), *Recent Trends of AI Technologies and Virtual Reality: Proceedings of 8th International Conference on Artificial Intelligence and Virtual Reality (AIVR 2024)*. Singapore: Springer.
- 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>
- Mercer, N., Hennessy, S., & Warwick, P. (2019). Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research*, 97, 187–199. <https://doi.org/10.1016/j.ijer.2017.08.007>
- Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, 27, 283–297. <https://doi.org/10.1007/s11217-007-9071-1>
- Misiejuk, K., Kaliisa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse: A question of reliability. *Computers & Education: Artificial Intelligence*, 6, Article 100216. <https://doi.org/10.1016/j.caeai.2024.100216>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), Article 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International journal of qualitative methods*. Advance online publication. <https://doi.org/10.1177/16094069231211248>
- Ng, O. L., Ni, Y., Shi, L., Chen, G., & Cui, Z. (2021). Designing and validating a coding scheme for analysis of teacher discourse behaviours in mathematics classrooms. *Journal of Education for Teaching*, 47(3), 337–352. <https://doi.org/10.1080/02607476.2021.1896340>
- Nili, A., Tate, M., Barros, A., & Johnstone, D. (2020). An approach for selecting and using a method of inter-coder reliability in information management research. *International Journal of Information Management*, 54, Article 102154. <https://doi.org/10.1016/j.ijinfomgt.2020.102154>
- Office of the Chief Information Officer. (2024). Use of ChatGPT at EdUHK. Retrieved from <https://www.edu.hk/ocio/eduhk-chatgpt>.
- OpenAI. (2022). Introducing ChatGPT. Retrieved from <https://openai.com/index/chatgpt/>.
- OpenAI. (2023). Gpt-4 technical report. Retrieved from <https://cdn.openai.com/papers/gpt-4.pdf>.
- OpenAI. (2024). Introducing GPT-4o and more tools to ChatGPT free users. Retrieved from <https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/>.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th international conference on neural information processing systems* (pp. 27730–27744). New York, NY: ACM. <https://dl.acm.org/doi/10.5555/3600270.3602281>.
- Plevris, V., Papazafeiropoulos, G., & Jiménez Ríos, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and google bard. *AI Artificial intelligence*, 4, 949–969. <https://doi.org/10.3390/ai4040048>
- Resnick, L. B., Michaels, S., & O'Connor, M. C. (2010). How (well-structured) talk builds the mind. In D. D. Preiss, & R. J. Sternberg (Eds.), *Innovations in educational psychology: Perspectives on learning, teaching, and human development* (pp. 163–194). New York, NY: Springer.
- Roberts, S. A., & Olarte, T. R. (2023). Enacting multilingual learner core practices: A PST's approximations of practice of mathematics language routines. Advance online publication *Journal of Mathematics Teacher Education*. <https://doi.org/10.1007/s10857-023-09600-8>.
- Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., & Mello, R. F. (2024). Assessing the quality of automatic-generated short answers using GPT-4. *Computers & Education: Artificial Intelligence*, 7, Article 100248. <https://doi.org/10.1016/j.caeai.2024.100248>
- Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and open-AI models: A preliminary review. *Future Internet*, 15(6), 192. <https://doi.org/10.3390/fi15060192>
- Sen, M., Sen, S. N., & Sahin, T. G. (2023). A new era for data analysis in qualitative research: ChatGPT. *Shanlax International Journal of Education*, 11(S1), 1–15. <https://doi.org/10.34293/education.v11iS1-Oct.6683>
- Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., Mannuru, A., & Batool, L. (2024). Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Preprints*, 2024, Article 2024061635. <https://doi.org/10.20944/preprints202406.1635.v1>
- Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE-Life Sciences Education*, 13(4), 624–635. <https://doi.org/10.1187/cbe.14-06-0108>
- Thanheiser, E., & Melhuish, K. (2019). Leveraging variation of historical number systems to build understanding of the base-ten place-value system. *ZDM—Mathematics Education*, 51, 39–55. <https://doi.org/10.1007/s11858-018-0984-7>
- Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine*, 37(5), 360–363. <https://pubmed.ncbi.nlm.nih.gov/15883903/>.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wang, D., & Chen, G. (2024). Are perfect transcripts necessary when we analyze classroom dialogue using AIoT? *Internet of Things*, 25, Article 101105. <https://doi.org/10.1016/j.iot.2024.101105>
- Wang, R. E., Wirawarn, P., Goodman, N., & Demszy, D. (2023). Sight: A large annotated dataset on student insights gathered from higher education transcripts. *arXiv preprint arXiv:2306.09343*. <https://doi.org/10.48550/arXiv.2306.09343>
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. <https://doi.org/10.1109/JAS.2023.123618>
- Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Proceedings of 28th international conference on intelligent user interfaces* (pp. 75–78). New York, NY: ACM. <https://doi.org/10.1145/3581754.3584136>.
- Yan, L., Echeverría, V., Nieto, G. F., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., & Martínez-Maldonado, R. (2024). The human-AI collaboration in thematic analysis using ChatGPT: A user study and design recommendations. In F. F. Mueller, P. Kyburz, J. R. Williamson, & C. Sas (Eds.), *Extended abstracts of the CHI conference on the human factors in computing systems*. New York, NY: ACM. <https://doi.org/10.1145/3613905.3650732>. article no. 191).
- Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & Nasir, N. (2023). From nCoder to ChatGPT: From automated coding to refining human coding. In G. Arastoopour Irgens, & S. Knight (Eds.), *Advances in quantitative ethnography. ICQE 2023. Communications in computer and information science* (Vol. 1895, pp. 470–485). Cham: Springer. https://doi.org/10.1007/978-3-031-47014-1_32.
- Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2024). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint arXiv:2309.10771*. <https://doi.org/10.48550/arXiv.2309.10771>.
- Zhu, N., Zhang, N., Shao, Q., Cheng, K., & Wu, H. (2024). OpenAI's GPT-4o in surgical oncology: Revolutionary advances in generative artificial intelligence. *European Journal of Cancer*, 206, Article 114132. <https://doi.org/10.1016/j.ejca.2024.114132>