Scaling Up Collaborative Dialogue Analysis: An AI-driven Approach to Understanding Dialogue Patterns in Computational Thinking Education

Stella Xin Yin*
Wee Kim Wee School of Communication and Information,
Nanyang Technological University
Singapore, Singapore
xin013@e.ntu.edu.sg

Zhengyuan Liu
Institute for Infocomm Research (I2R),
A*STAR
Singapore, Singapore
liu_zhengyuan@i2r.a-star.edu.sg

Dion Hoe-Lian Goh
Wee Kim Wee School of Communication and Information,
Nanyang Technological University
Singapore, Singapore
ASHLGoh@ntu.edu.sg

Choon Lang Quek
National Institute of Education,
Nanyang Technological University
Singapore, Singapore
choonlang.quek@nie.edu.sg

Nancy F. Chen
Institute for Infocomm Research (I2R),
A*STAR
Singapore, Singapore
nfychen@i2r.a-star.edu.sg

Abstract

Pair programming is a collaborative activity that enhances students' computational thinking (CT) skills. Analyzing students' interactions during pair programming provides valuable insights into effective learning. However, interpreting classroom dialogues is a challenging and complex task. Due to the simultaneous interaction between interlocutors and other ambient noise in collaborative learning contexts, previous work heavily relied on manual transcription and coding, which is labor-intensive and time-consuming. Recent advancements in speech and language processing offer promising opportunities to automate and scale up dialogue analysis. Besides, previous work mainly focused on task-related interactions, with little attention to social interactions. To address these gaps, we conducted a four-week CT course with 26 fifth-grade primary school students. We recorded their discussions, transcribed them with speech processing models, and developed a coding scheme and applied LLMs for annotation. Our AI-driven pipeline effectively analyzed classroom recordings with high accuracy and efficiency. After identifying the dialogue patterns, we investigated the relationships between these patterns and CT performance. Four clusters of dialogue patterns have been identified: Inquiry, Constructive Collaboration, Disengagement, and Disputation. We observed that Inquiry and Constructive Collaboration patterns were positively related to students' CT skills, while Disengagement and Disputation patterns were associated with lower CT performance. This study contributes to the understanding of how dialogue patterns relate to CT performance and provides implications for both research and educational practice in CT learning.

CCS Concepts

• **Applied computing → Collaborative learning**; **Interactive learning environments**; • **Computing methodologies → Natural language processing**.

Keywords

Collaborative learning, Computational thinking, Dialogue analysis, Large language models, Pair programming, Speech and language processing

ACM Reference Format:
Stella Xin Yin, Zhengyuan Liu, Dion Hoe-Lian Goh, Choon Lang Quek, and Nancy F. Chen. 2025. Scaling Up Collaborative Dialogue Analysis: An AI-driven Approach to Understanding Dialogue Patterns in Computational Thinking Education. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706474

1 Introduction

Computational Thinking (CT) is recognized as an integral component of digital and information literacy, and a fundamental skill for students to navigate and thrive in an increasingly technology-driven world [59]. CT includes not only the basic concepts of computer science but also the cognitive and social dimensions of problem-solving skills [40]. Given the growing importance of CT skills, researchers and educators have designed curricula and supported CT education across K-12 levels [9]. To better facilitate CT learning, collaborative learning is promoted as an effective teaching approach for imparting CT skills [66]. Among various CL strategies, pair programming (PP) stands out as one of the popular collaborative activities. In PP, two learners alternate roles as driver and navigator to solve CT tasks together on one computer. Benefits include improved programming skills [23], positive attitudes [69], and enhanced collaboration skills [16].

[*] Corresponding Author: Stella Xin Yin (xin013@e.ntu.edu.sg)

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706474

To gain insights into the mechanisms that contribute to effective collaboration, researchers have analyzed students' behaviors, emotions, and interactions during PP [56, 65]. Particularly, the verbal interchange between students has been seen as a central facet of collaborative efficacy [38]. Such verbal interchange is closely related to conceptual understanding and long-term knowledge retention, which have the greatest effect on learning outcomes [57, 61]. Therefore, identifying dialogue patterns that emerge in groups and investigating the relationship between these patterns and CT performance could improve collaboration quality and enhance CT learning experiences and outcomes. Further, productive collaborative talk does not occur naturally, and it requires facilitation through instruction, scaffolding, and prompting [48]. Consequently, teachers play a crucial role in the collaboration process by providing real-time guidance and support.

However, several challenges hinder the comprehensive understanding of collaborative dialogues in CT education. First, previous studies analyzing dialogues have mainly relied on manual transcription and coding, which is labor-intensive and time-consuming [56, 65]. With the advancements in speech and language processing techniques, such as Automatic Speech Recognition (ASR) [67], Voice Activity Detection (VAD) [12], and Large Language Models (LLMs) [1], significant improvements have been observed in dialogue understanding and generation [13, 50]. This enables the significant potential for automated classroom dialogue analysis, which allows efficient labeling and visualization of interactions. Second, previous work has mainly focused on task-related interaction, with little attention given to social interaction. However, social interaction is equally important for effective collaborative learning, as it supports task-related discussions and benefits students' socialization and academic achievement [27]. Third, little research has explored the sequence of dialogue patterns and their relationships with individual CT performance. According to Barr and Stephenson's [5] CT framework, CT is defined as a sequential problem-solving process that can be automated, transferred, and applied across subjects. It also provides a foundation for analyzing learning processes within CT tasks. We thus argue that the sequences of dialogue patterns closely relate to CT performance.

To address the above gaps, in this work, we aim to 1) build an AI-driven pipeline and evaluate the effectiveness of leveraging speech & language processing (e.g., ASR, VAD, LLMs) for collaborative dialogue analysis, 2) identify dialogue patterns during PP, and 3) examine the relationship between dialogue patterns and students' CT performance. Specifically, we answer the following questions:

- Q1: How effective are AI-driven techniques in automating collaborative dialogue analysis?
- Q2: What different patterns of collaborative dialogue emerged during PP?
- Q3: What are the relationships between dialogue patterns and students' CT skills?

2 Related works

2.1 Collaborative learning and pair programming in CT education

Collaborative learning is defined as a set of teaching and learning approaches that involve groups of learners working together to solve problems, complete tasks, or create products [17]. In collaborative settings, students are encouraged to ask questions, provide detailed explanations, present and defend ideas, generate new ideas, and formulate problem solutions [33]. Research has revealed the positive effects of collaborative learning on academic performance, motivation, cognition, and social skill development [29, 55]. In CT education, collaborative learning is widely recognized and promoted as an effective teaching approach for imparting CT skills [66]. It takes many forms in teaching and learning activities, varying from problem-based learning, peer reviewing, and PP.

PP has gained popularity in educational settings and demonstrated its effectiveness in teaching introductory programming concepts to novice learners. Recently, some schools have begun to integrate PP into K-12 CT practices [19]. Compared to problem-based learning and game-based learning, PP has its distinct advantages in delivering CT concepts and practices. First, PP derives from team-based learning but requires only two students to work together on the same computer, taking turns as the "driver" and "navigator" to jointly solve the assigned problem. This collaborative approach ensures both participants remain engaged actively in the tasks, leaving no room for slacking off [43]. In addition, this interactive nature of PP promotes more frequent and meaningful interactions and discussions between learners such as explaining concepts, questioning actions, and arguing or negotiating with partners, thereby enhancing effective collaboration [64]. Second, during PP, both students have the opportunity to practice and enhance their individual programming abilities [10]. This dual practice approach fosters a dynamic learning environment, wherein students can learn from each other's thought processes and mistakes, resulting in fewer errors and a deeper understanding of CT concepts. Moreover, research suggests that such learning experiences can have long-term effects on students' retention and programming skills [23]. These advantages align well with CL pedagogical guidelines, making PP an ideal approach for teaching CT skills.

2.2 Interaction dynamics in collaborative learning

Collaborative learning is grounded in social constructivist learning theory, which emphasizes that learning is socially situated and knowledge is constructed through interaction with others [17, 52]. From this perspective, collaboration is a dynamic interactive process wherein learners engage actively with each other to construct knowledge and collectively solve problems [42]. The interaction among collaborative learners has been regarded as a “gold mine of information” on how learners acquire knowledge and skills together [24]. Further, the quality of collaborative interactions has direct influence on effective learning as they enhance the process of knowledge elaboration and shared understanding, foster group cohesiveness, promote a team-oriented mindset, and cultivate a strong sense of community among participants [30, 46]. According to previous research (e.g., [2, 27, 28]), collaborative interactions include diverse patterns of language use and practice, which can be classified into two forms.

The first form, **task-related interaction** involves different kinds of content-related helping behavior, such as questioning, argumentation, and elaboration [61]. Based on the interactive process between peers, Mercer [37] introduced a framework to distinguish conversation from less productive to more productive: *Disputational* (unproductive), *Cumulative* (less productive), and *Exploratory* (productive). *Disputational talk* refers to disagreement, often followed by constructive criticism or suggestions. *Cumulative talk* is characterized by positive responses to what partners have said, such as repetitions, confirmations, and elaborations. *Exploratory talk* involves participants engaging critically but constructively with each other's ideas, leading to improved reasoning and conceptual understanding [37]. Several studies that applied this framework to analyze students' conversations during PP showed that students who organically used *Exploratory talk* significantly achieved higher performance in CT tasks [60, 68].

Next, the **social interaction** form refers to social-emotional support during collaborative learning activities [28]. According to Bakhtiar et al. [3], social interaction is purposeful interchange among group members that shapes perceptions of emotions and socio-emotional climate within the group. Positive interaction, such as encouragement or positive feedback on group work, has been shown to facilitate productive collaboration [28]. In contrast, negative interaction, such as discouraging other students from participating or disrespecting other group members, can have adverse consequences on group cohesion, commitment, satisfaction, and performance [41]. Although social interaction is not directly related to learning content, it is an indispensable part to support content-related discussion and benefit students' socialization and academic achievement [27]. Therefore, fostering positive and supportive social interaction is essential for promoting productive collaboration and optimizing group performance.

Previous research found that these two forms of interaction alternate during the collaboration process, and both are key factors in successful collaborative learning [28, 61]. For example, Aksoy-Pekacar [2] focused on language classes and categorized task-related interaction into five types. This work not only sheds light on the different types of task-related interaction but also offers valuable insights for future language teaching and learning approaches. For social interaction, Huang and Lajoie [25] reviewed empirical findings on social interaction in collaboration contexts and identified effective strategies and practical suggestions for educators and instructors to facilitate and encourage students' social interaction, thereby enhancing the learning experience. In addition, Isohätälä et al. [28] investigated the fluctuation of cognitive (task-related) and socio-emotional interactions during collaboration activities. Their findings suggested that socio-emotional interaction tends to involve more active participation compared to cognitive interaction.

While research has demonstrated the effectiveness of the two forms of interaction in promoting effective collaboration across various educational contexts [11, 54, 57], there is still limited understanding of how different kinds of interactions emerge during PP and the relationship between the two forms of interaction and CT skills achievement. CT education differs from traditional subjects like language, math, and science, as it involves extensive practice with computers and digital devices. In CT classes, peer interactions go beyond mere engagement in problem-solving solutions; they also involve computer manipulation and the creation of artifacts. Given that PP is an emergent practice in CT classes at the K-12 level, both instructors and students are often unfamiliar with the specific forms and content of PP activities. Consequently, implementing PP in CT classes may lead to difficulties and challenges, including the need for guidance on group members' roles and the necessity for effective facilitation of the overall collaboration process [66].

2.3 AI-driven methods for dialogue analysis

Analyzing classroom dialogic interaction is a complex task involving understanding various aspects of spoken language, discourse structure, and dynamic interactions between speakers in a noisy environment [62]. Different methods are involved, such as Automatic Speech Recognition (ASR), Voice Activity Detection (VAD), and computational linguistic analysis.

Speech processing is the first step to analyzing spoken language and extracting meaningful information from audio data. In educational contexts, automated transcription of classroom recordings can significantly improve the efficiency of linguistic and pedagogical analysis. However, accurate ASR of child speech remains challenging [7] since acoustics and linguistic properties such as the spectral and temporal features of adults and children are different [63]. Moreover, in collaboration contexts, multiple small groups sit close to each other in a classroom resulting in challenges including noisy audio, multiparty chatter, and other ambient noise [36]. The state-of-the-art data-driven approaches (e.g., Google Speech-to-Text [22] and OpenAI Whisper [47]) show superior performance on children's speech compared to traditional ASR models [51]. In addition, VAD systems are designed to distinguish between speech and non-speech segments within an audio stream [12]. Integrating VAD with ASR systems can further enhance the accuracy of speech transcription in noisy and dynamic classroom environments [15, 26].

Coding the transcribed data is a necessary step for dialogue analysis in order to determine the prevalence of patterns. However, it is labor-intensive, time-consuming, and difficult to scale for timely feedback. While previous work applied statistical methods [53] and trained task-specific models [44], these approaches often require in-domain training data and may struggle with the nuances and contextual richness of natural dialogue. Recently, LLMs have demonstrated promising capabilities in natural language understanding, generation, and reasoning tasks [1]. This has opened new possibilities for efficient data annotation and linguistic analysis. Several studies have shown that LLMs can be applied to categorizing and labeling tasks with human-level performance but at lower annotation costs [13, 20, 35, 50].

Therefore, in this present research, we investigate the integration of advanced speech processing systems with LLMs to analyze dialogic interactions in collaboration settings. Our approach aims to evaluate the effectiveness of these automated and scalable methods for analyzing classroom interactions, enabling timely feedback in educational contexts.

3 Methodology

We developed a comprehensive pipeline to explore the role of dialogue patterns in collaborative CT tasks, involving data collection,

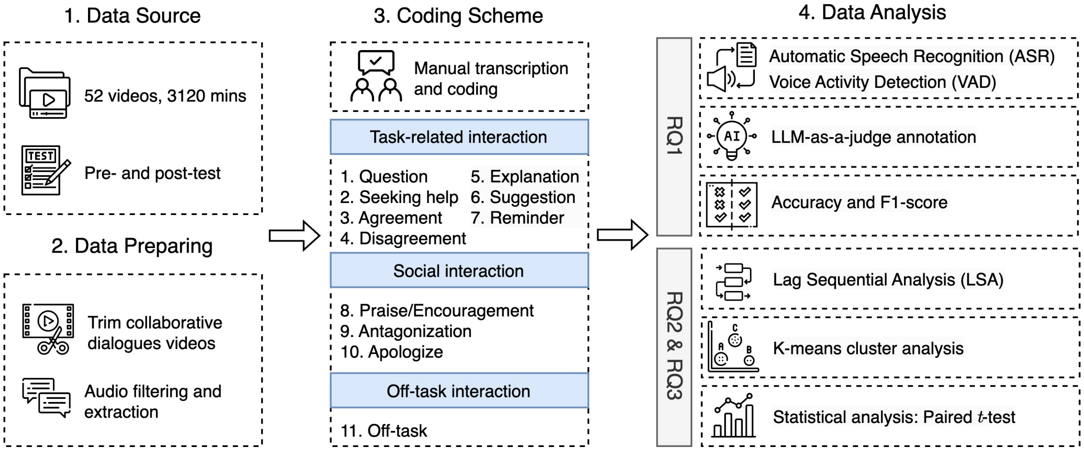


Figure 1: The overview of our AI-driven pipeline for analyzing collaborative dialogues.

preparation, developing the coding scheme, and dialogue analysis. Figure 1 outlines this pipeline.

3.1 Participants and contexts

We recruited 26 fifth-grade students (13 male and 13 female, *Mean age*=10.95) from a primary school to participate in a four-week CT course. After obtaining parental and participant consent [1], participants were paired up randomly and collaborated on CT tasks. Before the first week, participants took a pre-test to assess their prior knowledge of CT and programming experience. From week 1 to week 4, participants attended a 60-minute lesson each week to learn basic CT concepts and practice CT tasks. Each lesson began with a 10-minute introduction to CT concepts and guidelines for PP practice. Next, students had 45 minutes to collaboratively complete assigned tasks. In the final 5 minutes, students were encouraged to share their task solutions and receive feedback from peers and teachers. During PP practice, participants collaborated on tasks and alternated between "driver" and "navigator" roles. When in the "driver" role, students controlled the keyboard and mouse, constructing and editing code. Meanwhile, students in the "navigator" role closely observed their partner's work, identified errors, offered structural suggestions, and sought clarification if disagreements arose. The teacher acted as a facilitator, instructing students to switch roles after a predetermined period. After the four-week course, participants had learned basic CT concepts including sequencing, loops, and events, and were able to apply these concepts to create games. The course contents were adapted from Code.org (Course E), a learning platform offering educational resources and tools suitable for K-12 classrooms [14]. At the end of the course, participants completed the same CT test again.

3.2 Instruments

The *CT skill test* was adapted from Román-González et al. [49] (α = .79), utilizing 18 items relevant to our CT course on sequences, loops, and events. To ensure consistency and comparability, the same CT skill test was administered for both pre- and post-tests. This ensured that both tests had an equal level of difficulty [45]. In addition, the post-test question items and choices were shuffled to minimize any testing effects, ensuring that the students' performance on the post-test was not unduly influenced by the pre-test.

The *coding scheme* was developed from earlier studies (e.g., [28, 58, 68]) and consisted of two forms of interaction: task-based interaction and social interaction. Based on the original coding scheme, we manually coded two sample videos, and we found some unique patterns in the PP process that were not covered by the initial coding scheme, including 1) sharing uncertainties, and waiting for a partner's assistance; 2) speaking out the subsequent coding block names; 3) encouraging partners to run the program; 4) apologizing for giving incorrect suggestions; and 5) engaging in off-task interaction, such as mimicking program sounds. Consequently, we added those five new utterance behaviors to the initial scheme to form the final coding scheme, as shown in Table 1.

3.3 Data collection and analysis

We collected a total of 52 videos, with each about 60 minutes long. Prior to addressing our research questions, we randomly selected two videos, manually transcribed the dialogue, and coded each utterance according to our coding scheme. Any disagreements that arose were resolved through discussion between the coders. The raw agreement was 0.83, suggesting substantial reliability. The kappa values for each category are: Question (κ=0.901), Seeking help (κ=0.821), Agreement (κ=0.881), Disagreement (κ=0.799), Explanation (κ=0.786), Suggestion (κ=0.797), Reminder (κ=0.805), Praise/Encouragement (κ=0.867), Antagonization (κ=0.792), and Apologize (κ=0.765). These manually coded documents served as a baseline to evaluate the feasibility and effectiveness of automated methods.

To address RQ1, we applied AI-driven methods to automatically transcribe collaborative dialogues and annotate utterances. First, we combined ASR and VAD to convert audio files into textual utterances with speaker-level segmentation. This process involved 1) extracting vocal features using Titanet to build speaker embeddings; 2) applying Whisper for ASR, with timestamps generated

[1] The IRB approval was obtained from Nanyang Technological University. The approval number is IRB-2024-062.

Table 1: Coding scheme for collaborative dialogue patterns

Categories	Sub-categories	Descriptions	Examples
Task-related interaction	1. Question (QUE)	Any inquiry or expression of concern, often seeking clarification or confirmation.	“Where can I find the ‘if’ block?”
	2. Seeking help (SH)	Waiting for assistance.	“I don’t understand.”
	3. Agreement (AGR)	Agreement on any opinion or edits	“Yeah, I think so too.”
	4. Disagreement (DIS)	Disagreement on any opinion or edits.	“No, this should be here. No, I think...”
	5. Explanation (EXP)	Providing an explanation of the steps being taken or the edits being made.	“If I add ‘forward’ here, it is supposed to draw a line.”
	6. Suggestion (SUG)	Offering help or identifying the specific block before or during programming.	“Turn right and move forward.”
	7. Reminder (REM)	Reminding, redirecting or warning about performing specific steps or staying on task.	“Wait, wait, stop!”
Social interaction	8. Praise / Encouragement (PE)	Encouraging or praising the partner.	“I am so proud of you and myself.”
	9. Antagonization (ANT)	Making hurtful comments, putting down partner’s contributions, and showing annoyance with partner	“You are being ridiculous.”
	10. Apologize (APO)	Apologizing for providing incorrect suggestions.	“Sorry, it’s my mistake.”
Off-task interaction	11. Off-task (OT)	Any utterance that unrelated to the activity or imitating program sounds.	“It’s almost lunch time.” or “Uh-huh, uh”

by WhisperX; 3) utilizing MarbleNet for VAD to segment utterances and exclude silences. Next, we produced the final transcripts by aligning ASR and VAD results. To evaluate the transcription accuracy of ASR and VAD in our dataset, we manually sampled and transcribed 1000 utterances and calculated the word error rate (WER) of models on the transcripts. WER is a common metric of the performance of a speech recognition system and is calculated based on the Levenshtein Distance, which quantifies the difference between two utterances by counting the number of words needed to transform one into the other [31]. The WER metric ranges from 0 to 1, where 0 indicates that the compared pieces of text are identical, and 1 indicates that they are completely different with no similarity.

Second, we leveraged LLMs (i.e., LLM-as-a-judge) for utterance annotation. Based on previous work [34], we created scoring instructions by combining the coding scheme, scoring criteria, and utterance context. Next, we fed these instructions to the LLMs to predict utterance types. To evaluate the feasibility and efficacy of our AI-driven methods, we compared the predicted results to manual annotations.

To address RQ2 and RQ3, we employed lag sequential analysis (LSA), cluster analysis to identify dialogue patterns that emerged in pair groups, and paired *t*-test to investigate the relationship between dialogue patterns and CT performance.

4 Results

The video recordings of 13 groups were transcribed to a total of 22,243 utterances, with an average of 1,711 utterances per group. As shown in Figure 2, Group 7 was the least active group (speaking time = 6,969 seconds, utterance count = 1,032), while Group 2 was the most active group (speaking time = 8,760 seconds, utterance count = 2,561).

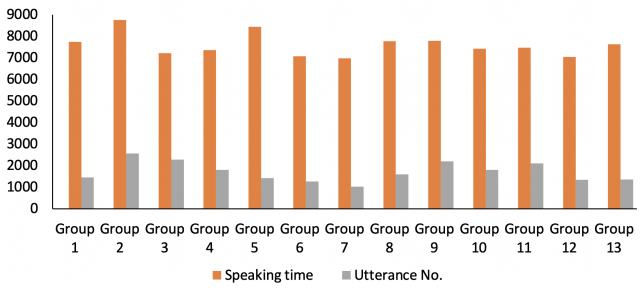
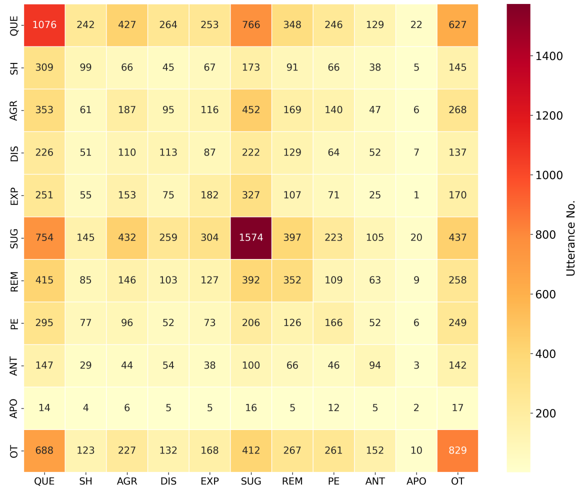


Figure 2: Speaking time and utterance number of each group.



	QUE	SH	AGR	DIS	EXP	SUG	REM	PE	ANT	APO	OT
QUE	1076	242	427	264	253	766	348	246	129	22	627
SH	309	99	66	45	67	173	91	66	38	5	145
AGR	353	61	187	95	116	452	169	140	47	6	268
DIS	226	51	110	113	87	222	129	64	52	7	137
EXP	251	55	153	75	182	327	107	71	25	1	170
SUG	754	145	432	259	304	1574	397	223	105	20	437
REM	415	85	146	103	127	392	352	109	63	9	258
PE	295	77	96	52	73	206	126	166	52	6	249
ANT	147	29	44	54	38	100	66	46	94	3	142
APO	14	4	6	5	5	16	5	12	5	2	17
OT	688	123	227	132	168	412	267	261	152	10	829

Figure 3: Heatmap of sequential frequency transfer matrix.

Table 2: Model comparison of LLM-as-a-judge for utterance annotation

	1-Shot Inference		3-Shot Inference		5-Shot Inference	
Model	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Llama3-8B-instruct	0.636	0.616	0.675	0.673	0.677	0.674
Llama3.1-8B-instruct	0.671	0.685	0.683	0.698	0.701	0.723
Gemma2-2B-instruct	0.578	0.601	0.629	0.641	0.659	0.678
Gemma2-9B-instruct	0.689	0.707	0.701	0.723	0.707	0.731
GPT-4o-mini	0.701	0.708	0.702	0.715	0.718	0.736

Table 3: The significance of utterance sequences

	QUE	SH	AGR	DIS	EXP	SUG	REM	PE	ANT	APO	OT
QUE	**7.53*****	**7.98*****	**2.80***	1.75	-2.20	-6.83	-3.76	-2.48	-2.21	0.98	-1.49
SH	**2.70****	**7.63*****	-2.54	-1.52	0.10	-3.39	-0.54	0.07	0.43	0.38	-0.71
AGR	-1.61	-3.12	**2.21***	-0.74	-0.48	**3.36*****	-0.51	**2.02***	-2.36	-0.66	-0.76
DIS	-1.06	-0.69	0.85	**6.39*****	1.28	-2.04	1.87	-1.42	1.79	0.98	-3.32
EXP	-2.28	-1.45	**3.18****	-0.15	**10.28*****	**2.12***	-2.28	-2.08	-3.55	-2.06	-3.01
SUG	-7.35	-5.63	**2.13***	0.64	0.48	**24.51*****	-1.88	-4.78	-4.92	0.25	-11.56
REM	0.12	-1.22	-2.43	-0.80	-0.42	-2.14	**12.90*****	-1.99	-0.96	0.21	-2.97
PE	1.01	1.54	-2.28	-2.84	-1.84	-5.82	-0.31	**8.83*****	0.62	0.12	**3.34****
ANT	-0.55	-1.15	-2.77	**2.11***	-1.61	-5.36	-0.58	-0.33	**13.74*****	-0.07	**3.07****
APO	-1.12	-0.12	-0.66	0.05	-0.35	-0.77	-1.24	**2.70****	1.09	**2.68****	1.06
OT	1.53	-2.65	-3.48	-3.69	-3.15	-12.58	-2.31	**4.26*****	**4.17*****	-1.00	**18.54*****

Note. QUE = Question, SH = Seeking help, AGR = Agreement, DIS = Disagreement, EXP = Explanation, SUG = Suggestion, REM = Reminder, PE = Praise/Encouragement, ANT = Antagonization, APO = Apologize, OT = Off-task $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

4.1 RQ1: AI-driven methods in automating collaborative dialogue analysis

We applied the WER metric to evaluate the transcription accuracy of ASR and VAD on our dataset. The average WER score was 0.095, suggesting that the speech recognition system correctly transcribed about 90.5% of the words. This level of accuracy is considered relatively high for classroom recording transcriptions and aligns with current benchmarks [18].

To demonstrate the feasibility and efficacy of automated annotation and compare the performance of LLMs, we prepared a human-annotated set as a reference, and results are presented in terms of correlation with human judgments, using accuracy and F1 scores. Here we selected and tested a list of representative models. As shown in Table 2, LLMs can provide reasonable results with only 1 in-context example (i.e., 1-shot inference), and the performance can be further improved in the 5-shot inference setting (adding more examples did not bring any improvement). The open LLMs can achieve state-of-the-art performance (e.g., Llama3.1-8B, Gemma2-9B), and are comparable to GPT-4o-mini (version 2024-07-18) in the few-shot setting. This demonstrates the feasibility of utilizing open LLMs to build automated and scalable utterance annotation. Moreover, we also calculate Cohen's kappa based on human labeled data; the result is 0.673 for Llama3.1-8B, 0.674 for Gemma2-9B, and 0.676 for GPT-4o-mini. This shows a substantial agreement between the machine and human raters.

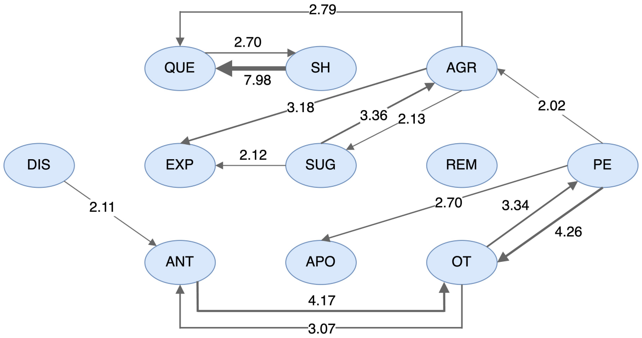


Figure 4: Utterance transition diagram of 13 groups.

4.2 RQ2: Collaborative dialogue patterns during pair programming

We employed LSA to investigate the sequential patterns of paired students' dialogues during PP. Our analysis aimed to identify whether specific utterances were statistically significant predictors of subsequent utterances. We created a heatmap to visualize the transition patterns between utterance types. This heatmap displays a matrix showing the frequency with which one type of utterance follows another, based on the 11 sub-categories defined in our coding scheme (see Figure 3). In this matrix, each column represents a current utterance, each row represents a subsequent utterance, and the values indicate the frequency with which the subsequent utterance occurs after the current one. Darker colors indicate higher frequencies, while lighter colors indicate lower frequencies. As shown in Figure 3, transfer among these three types - Question, Suggestion, and Off-task - occur most frequently.

Next, we calculated Z-scores to determine the significance of each sequence (see Table 3). Z-values exceeding 1.96, 2.58, and 3.29 indicate significance at the 0.05, 0.01, and 0.001 levels, respectively [21]. According to the Z-score results, we identified 14 significant utterance sequences (highlighted in red). To visualize these sequential patterns, Figure 4 presents a graphical representation of the significant utterance sequences. In this diagram, arrowheads indicate the direction of utterance transitions, while numerical values and arrow thickness denote Z-scores and significance levels, respectively. Our analysis revealed that Agreement and Praise/Encouragement exhibited the most connections with other utterance types. In contrast, Disagreement, Reminder, and Apologize demonstrated the fewest connections with other types.

To further examine the variability in dialog patterns among groups, we conducted a K-means cluster analysis on the 14 significant utterance sequences identified in the previous step. We employed the elbow method to determine the optimal number of clusters, following the approach proposed by Bholowalia and Kumar [8]. This method involves plotting the Within-Cluster Sum of Squares (WCSS) against a range of K values (from 2 to 10). The optimal cluster number is at the “elbow” of the curve, where adding more clusters doesn’t significantly reduce WCSS. Our analysis revealed that the four-cluster solution corresponded to this elbow point, suggesting that four clusters effectively categorize the utterance sequences into homogeneous groups. We then analyzed the sequential patterns for each cluster and visualized the transition diagrams in Figure 5. In the figure, we highlight the dominant sequential dialog patterns in the figure by using shaded areas.

Cluster 1 includes group 6, 10, 12, and 13. This cluster is characterized by dominant patterns transitioning from Seeking help to Question, and bidirectional transitions between Explanation and Suggestion. These patterns indicate that students’ dialogues primarily initiate inquiries about problem-solving steps and requests for partner’s assistance, followed by the exchange of suggestions and explanations. Based on these characteristics, we name this cluster as the "*Inquiry Pattern*."

Cluster 2 includes group 3, 8, and 9. In this cluster, we observed that the most frequent utterance sequences occurred between Antagonization and Off-task. This pattern suggests that students in these groups were not actively engaged in CT tasks and that negative emotions frequently emerged during collaboration. Based on these characteristics, we name this cluster as the "*Disengagement Pattern*."

Cluster 3 includes group 1, 5, 7, and 11. This cluster exhibits the least diverse sequence patterns among the four clusters identified. Notably, while students expressed inquiries, concerns, or sought help and confirmation from partners (transitioning from Seeking help to Question), their partners frequently responded with disagreement regarding proposed thoughts and solutions (transitioning from Question to Disagreement). This pattern suggests that during collaboration, these groups often encountered disagreement and assertions, and challenges in reaching a consensus on solutions. Thus, we name this cluster as the "*Disputation Pattern*."

Cluster 4 includes group 2 and 4. This cluster reflects the most diverse sequence patterns among all clusters. The predominant transition paths occurred between Agreement and Suggestion, as well as between Praise/Encouragement and Off-task. Interestingly, we observed that when students realized they had provided incorrect suggestions, they promptly apologized and subsequently sought help from their partners (from Apologize to Seeking help). These patterns indicate that students in this cluster engaged in positive knowledge construction and reached consensus through productive discussions. Based on these characteristics, we name this cluster as the "*Constructive Collaboration Pattern*."

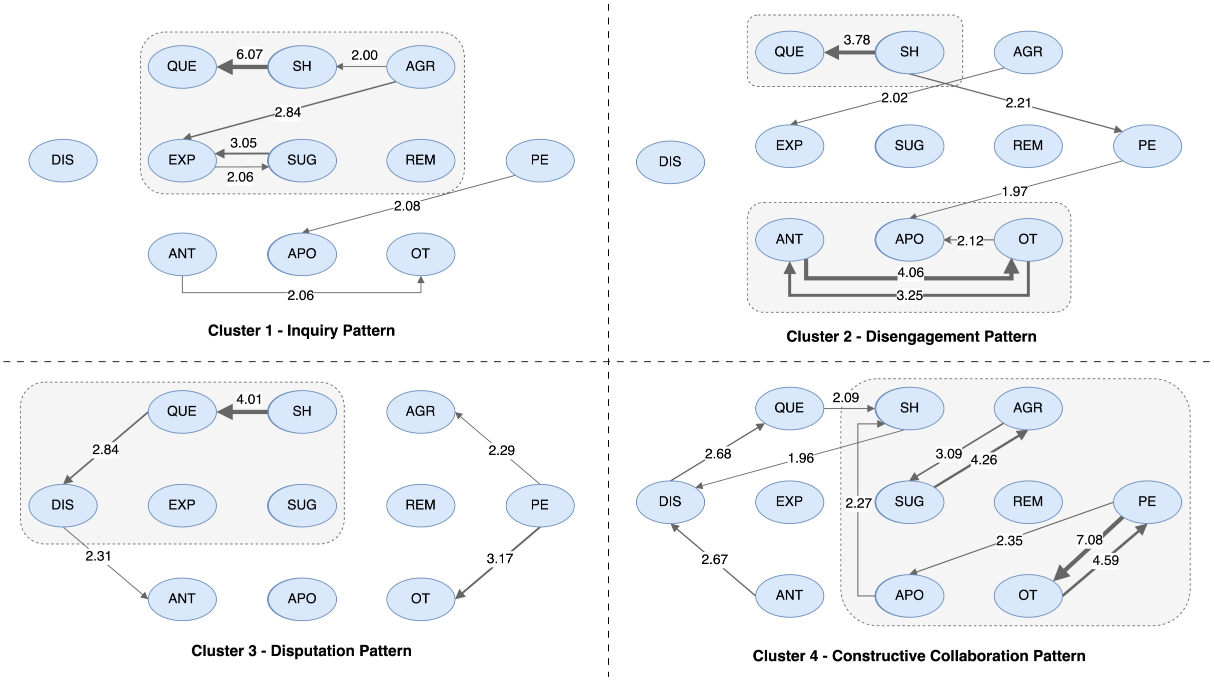
4.3 RQ3: The relationships between dialogue patterns and CT skills

To examine the relationships between four-cluster collaborative dialogue patterns on CT performance, we conducted paired t-tests to examine the mean differences between pre- and post-test scores following the 4-week intervention. Additionally, we calculated Cohen’s d to report effect sizes (see Table 4). For the entire class, students’ post-test scores were significantly higher than their pre-test scores across all clusters (t = -4.649, p < 0.001), indicating that PP activities significantly enhanced students’ CT skills. However, we observed distinct variations among the four clusters. Specifically, significant improvements in CT performance were found only in Cluster 1 ("Inquiry Pattern") (t = -3.752, p < 0.01) and Cluster 4 ("Constructive Collaboration Pattern") (t = -3.873, p < 0.05). In contrast, Cluster 2 ("Disengagement Pattern") (t = -31.719, p = 0.120) and Cluster 3 ("Disagreement Pattern") (t = -1.667, p = 0.194) showed no significant improvement in CT performance. These results suggest that collaborative dialogues characterized by inquiry-driven interactions and constructive collaboration patterns positively related to students’ CT performance. Conversely, dialogue patterns dominated by disengagement or persistent disagreement may exert an adverse effect on CT skill development.

5 Discussion

5.1 AI-driven dialogue analysis

In this study, we developed and evaluated an AI-driven pipeline that automates classroom dialogue analysis through speech recognition, segmentation, and pattern identification. Our experiment results demonstrate the outstanding performance of ASR models in recognizing children’s speech in collaboration contexts, as well as the feasibility of LLMs in annotating utterances. Moreover, this pipeline can scale efficiently to analyze hundreds of hours of CT classroom dialogues with minimal additional computational cost compared with human labeling, while maintaining consistent performance. However, several challenges remained to be addressed. We encountered difficulties in detecting children’s specific linguistic features, particularly when students mixed English with words or phrases from other languages such as Malay and Mandarin. Additionally, the CT tasks in our study involved a lot of sound effects alongside program execution, which created difficulties for AI models to accurately identify words amid background noise. Furthermore, the collaborative nature of the activities frequently results in multiparty noise, with two or more students speaking simultaneously. This poses additional challenges for the models to accurately detect and



Cluster 1 - Inquiry Pattern

Cluster 2 - Disengagement Pattern

Cluster 3 - Disputation Pattern

Cluster 4 - Constructive Collaboration Pattern

Figure 5: Utterance transition diagrams of four clusters.

Table 4: Mean difference of test scores among four clusters

Mean	Cluster 1 (N=4)	Cluster 2 (N=4)	Cluster 3 (N=2)	Cluster 4 (N=3)	All groups
Pre-test	13.375	14.700	15.000	11.750	13.885
Post-test	16.250	16.200	16.250	14.250	15.923
p-value	0.007**	0.120	0.194	0.030*	<0.001***
Cohen's *d*	1.327	0.544	0.833	1.936	0.912

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

transcribe speech. Thus, while our findings highlight the advancements of AI-driven systems in dialogue analysis, they also brought unsolved problems in special linguistic environments and complex collaborative settings. A potential solution would be to equip each student with a headset microphone, recording individual voices separately instead of group audio.

5.2 Collaborative dialogue patterns for effective learning

5.2.1 Inquiry and Constructive collaboration pattern. Our findings align with previous research showing that groups using more Exploratory Talk tend to solve problems more successfully and achieve better learning outcomes in subsequent tasks [38]. Additionally, our study extends this evidence to the context of CT education. In our study, students in Clusters 1 and 4 were actively engaged in Exploratory Talk, including seeking help from partners, sharing knowledge, challenging ideas, evaluating solutions, and expressing disagreement with alternative solutions. These patterns are considered essential components of effective collaboration [39]. The positive outcomes observed in Clusters 1 and 4 highlight the necessity of incorporating Exploratory Talk strategies into PP activities to enhance CT learning.

5.2.2 Positive social interaction pattern. Praise/encouragement was observed across all clusters, and this positive social interaction pattern influences the effectiveness of collaborative learning. Our findings align with previous research, demonstrating that when students engage in positive social interactions, they were more willing to support each other and could express disagreement more easily, which encouraged critical, divergent, and exploratory opinions [28]. Interestingly, we found a significant utterance transition from Praise/encouragement to Apologize in Clusters 1 and 4. This is probably because students in these clusters, who engaged more deeply in Exploratory Talk and critical thinking, were more likely to recognize and reflect on their mistakes. For example, when Student A appreciated their partner's help by saying, "Wow, thanks for helping me debug it!", Student B might suddenly realize their mistake and respond, "Oh, sorry, it's my mistake. It should be '3 times'." In such conversations, students may feel comfortable taking risks, proposing ideas, and admitting when they are wrong.

This behavior is considered individual accountability and is one of the strategies to develop trust within a group [32]. Moreover, the transition from praise to apology suggests a dynamic and iterative learning process. Students in Clusters 1 and 4 seem to cycle through phases of encouragement, critical evaluation, and reflection. Thus, we argue that this transition pattern may contribute to their higher performance in CT tasks, as it encourages continuous refinement of ideas and solutions.

5.2.3 *Disputation Pattern.*

Previous research has shown that Disputational Talk [37] often increases disengagement, tension and anxiety within groups , ultimately leading to lower performance [25, 28]. In our study, we observed that students in Cluster 3 primarily engaged in Disputational Talk, such as disagreement and individualized decision-making. They rarely offered constructive criticism or made suggestions; instead, most of their utterances were brief and oppositional, such as “Don’t do that!” and “No, it’s not!” However, we found an unexpected outcome in this cluster. Despite their unproductive interaction patterns, their test scores were relatively high in both pre-and post-tests. There are two potential explanations. First, in some groups, a high-ability student was paired with a low-ability student. In such cases, the high-ability student might express frustration or disagreement when encountering errors in their partner’s work. Second, both students in some groups were high-ability learners who complete required tasks quickly. Consequently, they may engage in off-task interactions while waiting for lower-paced peers in the class to catch up.

5.2.4 *Off-task pattern.*

Our findings suggest that off-task interaction is associated with less engagement and lower academic performance, which consistent with research by Beserra et al. [6]. In our study, we observed the same dominant utterance type, Off-task, in both Cluster 2 and Cluster 4. However, the sequential patterns involving Off-task utterances had contrasting effects in these clusters. In Cluster 2, students initiated off-task conversations when they felt annoyed with partners or disliked the tasks. In contrast, in Cluster 4, off-task interactions mainly consisted of imitating sound effects of CT tasks, which fostered relationships and increased the enjoyment of completing tasks. Given these observations, we argue that while off-task interaction may impede problem-solving, it could have positive roles in learning if connected with positive social interaction, as it relates to the socio-affective aspects of learning. This argument is supported by previous research [4] suggesting that there is no clear distinction between ‘off’ and ‘on’ task talk and that learners might weave between ‘on’ and ‘off’ task conversation while remaining engaged with learning tasks. Such interweaving of interaction may contribute to a more relaxed and productive learning environment, potentially improving learning outcomes.

6 Conclusion

This research not only advances our understanding of the relationship between dialogue patterns and CT performance in collaborative learning contexts but also demonstrates the effectiveness and efficiency of the AI-driven pipeline in supporting educational research and dialogue analysis. We identified 14 significant sequential patterns during PP and classified them into four clusters of dialogue patterns: Inquiry, Disengagement, Disputation, and Constructive collaboration. These findings provide a foundation for developing more effective teaching strategies and interventions to enhance CT skills. We discuss theoretical and practical implications below.

Theoretically, we contribute to CT education research in collaborative contexts by providing insights into the underlying factors influencing CT performance and PP experiences. Next, we extended and refined the collaborative dialogue coding scheme by categorizing interactions into three dimensions: task-related interaction, social interaction, and off-task interaction. This refined framework enhances our understanding of the dialogue patterns that emerge from PP and explains the characteristics of sequential patterns in improving CT performance. Moreover, we explored the effectiveness of leveraging speech and language processing to support educational research and practice in dialogue analysis, opening up opportunities for future research to overcome limitations in analyzing large volumes of dialogue data.

Practically, our results pave the way for more targeted interventions and pedagogical strategies to foster effective collaborative learning in CT education. As Mercer [37] posits, productive collaborative talk does not naturally occur but needs facilitation through instruction, scaffolding, and prompting [48]. Thus, teachers play a key role in orchestrating effective collaboration in PP. Based on our findings, we recommend that teachers provide guidance to promote Inquiry and Constructive collaboration talk among students. Second, teachers should offer timely support when students get stuck and intervene when off-task interactions become excessive. Third, teachers should encourage respectful behavior and positive social interactions within the group to create a supportive learning environment. Last, developing AI-based tools for real-world classrooms should prioritize user-friendliness and effectiveness, enabling teachers to efficiently analyze and respond to dialogue patterns as they emerge. This would allow teachers to visualize and analyze students’ interactions promptly, thereby facilitating timely interventions for groups that fall into Disengagement or Disputation patterns.

6.1 Limitations and future work

This study has several limitations that could be addressed in future research. First, our participants were recruited from one primary school, which may limit the generalizability of our findings to other contexts. To address this limitation, future studies can increase the sample size by including multiple schools. Second, we randomly paired students for PP. As discussed earlier, the ability levels of students in pairs may influence their dialogue patterns. We recommend future research group students based on ability levels to examine whether same-ability or mixed-ability groups produce different dialogue patterns and how these relate to CT performance. Third, the intervention was only 4 weeks, which was a challenge for us to track long-term progress in dialogue patterns and their relationships with CT skills. Future studies should consider longitudinal designs that span a full semester or academic year to capture the evolution of collaborative dialogue patterns over time and their sustained effects on CT development. Lastly, regarding off-task interaction, we grouped unrelated utterances and imitations of program sounds into one type. However, our findings imply that off-task utterances may have both positive and negative effects on CT performance depending on their transitions. Thus, we recommend future research split off-task interactions into sub-types to gain better insights into specific impacts on learning outcomes. This could inform the development of targeted scaffolding strategies to facilitate students' CT learning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Kadriye Aksoy-Pekacar. 2024. Task-related collaborative behaviours in task-based oral peer interactions. *The Language Learning Journal* 52, 4 (July 2024), 397–413. https://doi.org/10.1080/09571736.2023.2193577

[3] Aishah Bakhtiar, Elizabeth A. Webster, and Allyson F. Hadwin. 2018. Regulation and socio-emotional interactions in a positive and a negative group climate. *Metacognition and Learning* 13, 1 (2018), 57–90. https://doi.org/10.1007/s11409-017-9178-x

[4] Khaled Barkaoui, Margaret So, and Wataru Suzuki. 2015. Is it Relevant? The Role of Off-task Talk in Collaborative Learning. *Journal of Applied Linguistics and Professional Practice* 5, 1 (Sept. 2015), 31–54. https://doi.org/10.1558/japl.v5i1.31

[5] Valerie Barr and Chris Stephenson. 2011. Bringing computational thinking to K-12. *ACM Inroads* 2, 1 (Feb. 2011), 48–54. https://doi.org/10.1145/1929887.1929905

[6] Vagner Beserra, Miguel Nussbaum, and Macarena Oteo. 2019. On-Task and Off-Task Behavior in the Classroom: A Study on Mathematics Learning With Educational Video Games. *Journal of Educational Computing Research* 56, 8 (2019), 1361–1383. https://doi.org/10.1177/0735633117744346

[7] Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B. Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic Speech Recognition (ASR) systems for children: A systematic literature review. *Applied Sciences* 12, 9 (April 2022), 4419. https://doi.org/10.3390/app12094419

[8] Purnima Bholowalia and Arvind Kumar. 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* 105, 9 (2014).

[9] Stefania Bocconi, Augusto Chioccariello, Panagiotis Kampylis, Valentina Dagienė, Patricia Wastiau, Katja Engelhardt, Jeffrey Earp, Milena Anna Horvath, Eglė Jasutė, Chiara Malagoli, Vaida Masiulionytė-Dagienė, and Gabrielė Stupurienė. 2022. *Reviewing computational thinking in compulsory education.* Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/126955

[10] Grant Braught, L Martin Eby, and Tim Wahls. 2008. The effects of pair-programming on individual programming skill. *ACM SIGCSE Bulletin* 40, 1 (Feb. 2008), 200–204. https://doi.org/10.1145/1352322.1352207

[11] Rebeca Cerezo, Miguel Sánchez-Santillán, M. Puerto Paule-Ruiz, and J. Carlos Núñez. 2016. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers and Education* 96 (2016), 42–54. https://doi.org/10.1016/j.compedu.2016.02.006

[12] Shuo-Yiin Chang, Bo Li, Gabor Simko, Tara N Sainath, Anshuman Tripathi, Aäron van den Oord, and Oriol Vinyals. 2018. Temporal modeling using dilated convolution and gating for voice-activity-detection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5549–5553.

[13] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. (2023). http://arxiv.org/abs/2306.14924

[14] Code.org. 2021. About Us | Code.org. https://code.org/about

[15] Samuele Cornell, Jee-Weon Jung, Shinji Watanabe, and Stefano Squartini. 2024. One Model to Rule Them All ? Towards End-to-End Joint Speaker Diarization and Speech Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 11856–11860. https://doi.org/10.1109/ICASSP48485.2024.10447957

[16] Caio Matheus Campos de Oliveira, Edna Dias Canedo, Henrique Faria, Luis Henrique Vieira Amaral, and Rodrigo Bonifacio. 2018. Improving student's learning and cooperation skills using coding dojos (In the wild!). In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–8. https://doi.org/10.1109/FIE.2018.8659056

[17] Pierre Dillenbourg. 1999. *What do you mean by "collaborative learning"?* Oxford: Elsevier, 1–19.

[18] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models. *arXiv preprint arXiv:2406.10507* (2024).

[19] Association for Computing Machinery, Code.org, Computer Science Teachers Association, Cyber Innovation Center, National Math, and Science Initiative. 2016. *K-12 Computer Science Framework.* http://www.k12cs.org.

[20] Ryan Garg, Jaeyoung Han, Yixin Cheng, Zheng Fang, and Zachari Swiecki. 2024. Automated Discourse Analysis via Generative Artificial Intelligence. In *Proceedings of the 14th Learning Analytics and Knowledge Conference.* ACM, New York, NY, USA, 814–820. https://doi.org/10.1145/3636555.3636879

[21] Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* 10, 2 (2012), 486.

[22] Google. 2024. Turn speech into text using Google AI. https://cloud.google.com/speech-to-text?hl=en

[23] Anja Hawlitschek, Sarah Berndt, and Sandra Schulz. 2023. Empirical research on pair programming in higher education: a literature review. *Computer Science Education* 33, 3 (July 2023), 400–428. https://doi.org/10.1080/08993408.2022.2039504

[24] Miguel Ángel Herrera-Pavo. 2021. Collaborative learning for virtual higher education. *Learning, Culture and Social Interaction* 28 (March 2021), 100437. https://doi.org/10.1016/j.lcsi.2020.100437

[25] Xiaoshan Huang and Susanne P. Lajoie. 2023. Social emotional interaction in collaborative learning: Why it matters and how can we measure it? *Social Sciences & Humanities Open* 7, 1 (Jan. 2023), 100447. https://doi.org/10.1016/j.ssaho.2023.100447

[26] Zili Huang, Desh Raj, Paola García, and Sanjeev Khudanpur. 2023. Adapting Self-Supervised Models to Multi-Talker Speech Recognition Using Speaker Embeddings. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10097139

[27] Jaana Isohätälä, Hanna Järvenoja, and Sanna Järvelä. 2017. Socially shared regulation of learning and participation in social interaction in collaborative learning. *International Journal of Educational Research* 81 (Jan. 2017), 11–24. https://doi.org/10.1016/j.ijer.2016.10.006

[28] Jaana Isohätälä, Piia Näykki, and Sanna Järvelä. 2020. Cognitive and socio-emotional interaction in collaborative learning: Exploring fluctuations in students' participation. *Scandinavian Journal of Educational Research* 64, 6 (2020), 831–851. https://doi.org/10.1080/00313831.2019.1623310

[29] Heisawn Jeong, Cindy E Hmelo-Silver, and Kihyun Jo. 2019. Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational research review* 28 (2019), 100284.

[30] Hanna Järvenoja, Sanna Järvelä, and Jonna Malmberg. 2020. Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction* 70, June 2017 (2020), 101090. https://doi.org/10.1016/j.learninstruc.2017.11.004

[31] Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 1-2 (2002), 19–28.

[32] Kyungbin Kwon, Ying Hsiu Liu, and Lashaune P. Johnson. 2014. Group regulation and social-emotional interactions observed in computer supported collaborative learning: Comparison between good vs. poor collaborators. *Computers and Education* 78 (2014), 185–200. https://doi.org/10.1016/j.compedu.2014.06.004

[33] Marjan Laal and Mozhgan Laal. 2012. Collaborative learning: what is it? *Procedia - Social and Behavioral Sciences* 31, 2011 (2012), 491–495. https://doi.org/10.1016/j.sbspro.2011.12.092

[34] Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F. Chen. 2024. Scaffolding Language Learning via Multi-modal Tutoring Systems with Pedagogical Instructions. In *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 1258–1265. https://doi.org/10.1109/CAI59869.2024.00223

[35] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 626–642. https://doi.org/10.18653/v1/2024.emnlp-main.37

[36] Yingbo Ma, Joseph B. Wiggins, Mehmet Celepkolu, Kristy Elizabeth Boyer, Collin Lynch, and Eric Wiebe. 2021. *The Challenge of Noisy Classrooms: Speaker Detection During Elementary Students' Collaborative Dialogue.* 268–281. https://doi.org/10.1007/978-3-030-78292-4_22

[37] Neil Mercer. 2000. *Words and Minds: How We Use Language to Think Together and Get Things Done.*

[38] Neil Mercer and Christine Howe. 2012. Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction* 1, 1 (March 2012), 12–21. https://doi.org/10.1016/j.lcsi.2012.03.001

[39] Neil Mercer and Karen Littleton. 2007. *Dialogue and the Development of Children's Thinking: A Sociocultural Approach - Neil Mercer, Karen Littleton - Google Books.*

[40] Anders Mørch and Yasmin Kafai. 2022. Computational thinking as a social movement. *KI - Künstliche Intelligenz* 36, 1 (mar 2022), 87–90. https://doi.org/10.1007/s13218-022-00754-w

[41] Piia Näykki, Sanna Järvelä, Paul A. Kirschner, and Hanna Järvenoja. 2014. Socio-emotional conflict in collaborative learning—A process-oriented case study in a higher education context. *International Journal of Educational Research* 68 (2014), 1–14. https://doi.org/10.1016/j.ijer.2014.07.001

[42] A. Sullivan Palincsar. 1998. Social constructivist perspectives on teaching and learning. *Annual Review of Psychology* 49, 1 (Feb. 1998), 345–375. https://doi.org/10.1146/annurev.psych.49.1.345

[43] Laura Plonka, Judith Segal, Helen Sharp, and Janet van der Linden. 2011. *Collaboration in pair programming: Driving and switching*. Vol. 77 LNBIP. Springer Verlag, 43–59. https://doi.org/10.1007/978-3-642-20677-1_4

[44] Samuel L Pugh, Arjun Rao, Angela EB Stewart, and Sidney K D'Mello. 2022. Do speech-based collaboration analytics generalize across task contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 208–218.

[45] Diana Pérez-Marín, Raquel Hijón-Neira, Adrián Bacelo, and Celeste Pizarro. 2020. Can computational thinking be improved by using a methodology based on metaphors and scratch to teach computer programming to children? *Computers in Human Behavior* 105 (April 2020), 105849. https://doi.org/10.1016/j.chb.2018.12.027

[46] Muhammad Asif Qureshi, Asadullah Khaskheli, Jawaid Ahmed Qureshi, Syed Ali Raza, and Sara Qamar Yousufi. 2023. Factors affecting students' learning performance through collaborative learning and engagement. *Interactive Learning Environments* 31, 4 (May 2023), 2371–2391. https://doi.org/10.1080/10494820.2021.1884886

[47] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

[48] Sylvia Rojas-Drummond and Neil Mercer. 2003. Scaffolding the development of effective collaboration and learning. *International Journal of Educational Research* 39, 1–2 (Jan. 2003), 99–111. https://doi.org/10.1016/S0883-0355(03)00075-2

[49] Marcos Román-González, Juan-Carlos Pérez-González, and Carmen Jiménez-Fernández. 2017. Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behavior* 72 (July 2017), 678–691. https://doi.org/10.1016/j.chb.2016.08.047

[50] Jennifer Santoso, Kenkichi Ishizuka, and Taiichi Hashimoto. 2024. Large Language Model-Based Emotional Speech Annotation Using Context and Acoustic Feature for Speech Emotion Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2024), 11026–11030. https://doi.org/10.1109/ICASSP48485.2024.10448316

[51] Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D'Mello, and Jacob Whitehill. 2024. Automatic Speech Recognition Tuned for Child Speech in the Classroom. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12291–12295. https://doi.org/10.1109/ICASSP48485.2024.10447428

[52] Gerry Stahl, Timothy Koschmann, and Daniel Suthers. 2014. *Computer-supported collaborative learning*. Cambridge University Press, 479–500. https://doi.org/10.1017/CBO9781139519526.029

[53] Florence R Sullivan and P Kevin Keith. 2019. Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. *British Journal of Educational Technology* 50, 6 (2019), 3047–3063.

[54] Zhong Sun, Chin Hsi Lin, Minhua Wu, Jianshe Zhou, and Liming Luo. 2018. A tale of two communication tools: Discussion-forum and mobile instant-messaging apps in collaborative learning. *British Journal of Educational Technology* 49, 2 (2018), 248–261. https://doi.org/10.1111/bjet.12571

[55] Yao-Ting Sung, Je-Ming Yang, and Han-Yueh Lee. 2017. The effects of mobile-computer-supported collaborative learning: Meta-analysis and critical synthesis. *Review of educational research* 87, 4 (2017), 768–805.

[56] Jinbo Tan, Lei Wu, and Shanshan Ma. 2024. Collaborative dialogue patterns of pair programming and their impact on programming self-efficacy and coding performance. *British Journal of Educational Technology* 55, 3 (2024), 1060–1081. https://doi.org/10.1111/bjet.13412

[57] Harriet R. Tenenbaum, Naomi E. Winstone, Patrick J. Leman, and Rachel E. Avery. 2020. How effective is peer interaction in facilitating learning? A meta-analysis. *Journal of Educational Psychology* 112, 7 (2020), 1303–1319. https://doi.org/10.1037/edu0000436

[58] Jennifer Tsan, Jessica Vandenberg, Zarifa Zakaria, Danielle C. Boulden, Collin Lynch, Eric Wiebe, Kristy Elizabeth Boyer, and Kristy Elizabeth Boyer. 2021. Collaborative Dialogue and Types of Conflict: An Analysis of Pair Programming Interactions between Upper Elementary Students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, Vol. 7. ACM, New York, NY, USA, 1184–1190. https://doi.org/10.1145/3408877.3432406

[59] UNESCO, Fengchun Miao, and Wayne Holmes. 2023. *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

[60] Jessica Vandenberg, Zarifa Zakaria, Jennifer Tsan, Anna Iwanski, Collin Lynch, Kristy Elizabeth Boyer, and Eric Wiebe. 2021. Prompting collaborative and exploratory discourse: An epistemic network analysis study. *International Journal of Computer-Supported Collaborative Learning* 16, 3 (Sept. 2021), 339–366. https://doi.org/10.1007/s11412-021-09349-3

[61] Essi Vuopala, Pirkko Hyvönen, and Sanna Järvelä. 2016. Interaction forms in successful collaborative learning in virtual learning environments. *Active Learning in Higher Education* 17, 1 (2016), 25–38. https://doi.org/10.1177/1469787415616730

[62] Deliang Wang, Yang Tao, and Gaowei Chen. 2024. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research* 123, July 2023 (2024), 102275. https://doi.org/10.1016/j.ijer.2023.102275

[63] Fei Wu, Leibny Paola Garcia, Daniel Povey, and Sanjeev Khudanpur. 2019. Advances in automatic speech recognition for child speech using factored time delay neural network. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2019-Septe (2019), 1–5. https://doi.org/10.21437/Interspeech.2019-2980

[64] Stelios Xinogalos, Maya Satratzemi, Alexander Chatzigeorgiou, and Despina Tsompanoudi. 2017. Student perceptions on the benefits and shortcomings of distributed pair programming assignments. In *2017 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1513–1521. https://doi.org/10.1109/EDUCON.2017.7943050

[65] Weiqi Xu, Yajuan Wu, and Fan Ouyang. 2023. Multimodal learning analytics of collaborative patterns during pair programming in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (Feb. 2023), 8. https://doi.org/10.1186/s41239-022-00377-z

[66] Stella Xin Yin, Dion Hoe-Lian Goh, and Choon Lang Quek. 2024. Collaborative learning in K-12 computational thinking education: A systematic review. *Journal of Educational Computing Research* 62, 6 (2024), 1440–1474. https://doi.org/10.1177/07356331241249956

[67] Dong Yu and Lin Deng. 2016. *Automatic speech recognition*. Vol. 1. Springer.

[68] Zarifa Zakaria, Jessica Vandenberg, Jennifer Tsan, Danielle Cadieux Boulden, Collin F. Lynch, Kristy Elizabeth Boyer, and Eric N. Wiebe. 2022. Two-computer pair programming: Exploring a feedback intervention to improve collaborative talk in elementary students. *Computer Science Education* 32, 1 (2022), 3–29. https://doi.org/10.1080/08993408.2021.1877987/FORMAT/EPUB

[69] Baichang Zhong, Qiyun Wang, and Jie Chen. 2016. The impact of social factors on pair programming in a primary school. *Computers in Human Behavior* 64 (nov 2016), 423–431. https://doi.org/10.1016/j.chb.2016.07.017