Collaborative Problem-Solving Dialogue Analysis with Interpretable Temporal Clustering

Yeo Jin Kim[1], Daeun Hong[2], Wookhee Min[1], Snigdha Chaturvedi[3]
Cindy E. Hmelo-Silver[2], and James Lester[1]

[1] North Carolina State University, U.S.A
{ykim32, wmin, lester}@ncsu.edu
[2] Indiana University, U.S.A
{dh37, chmelosi}@iu.edu
[3] UNC Chapel Hill, U.S.A
snigdha@cs.unc.edu

Abstract. Student communication during collaborative problem solving centers on sharing and negotiating ideas, regulating problem-solving processes, and maintaining social interaction. These cognitively and socially driven processes help students consolidate knowledge, manage their actions, and engage effectively in collaborative learning environments. To enhance these environments, analyzing student dialogue is crucial for delivering adaptive scaffolding and fostering deeper engagement and collaboration. However, such analysis poses significant challenges due to the complexity of student interactions and the need for interpretable analysis. To address these challenges, we introduce a novel framework for analyzing collaborative problem-solving dialogue that integrates temporal clustering of dialogue patterns with LLM-generated explanations. Using video and chat log data from middle school student groups engaged in a collaborative game-based learning environment, we demonstrate that our framework effectively identifies collaborative problem-solving dialogue patterns. Furthermore, the LLM-generated interpretations enhance the interpretability of these clusters, to enable both collaborative problem-solving assessment and early prediction of learning outcomes. This work lays the foundation for enabling adaptive scaffolding, automated collaboration assessment, and improved learning processes within collaborative, game-based educational settings.

Keywords: Dialogue analysis · Temporal clustering · Large language model · Collaborative learning · Collaborative problem solving.

1 Introduction

Collaborative problem solving has emerged as a key 21st-century competency, requiring individuals to integrate knowledge, skills, and efforts to solve complex problems [10, 22]. Collaborative learning environments that feature collaborative problem solving have been shown to enhance student learning, engagement, and problem-solving outcomes [10, 17]. In these settings, students develop collaborative problem-solving skills by tackling complex, ill-structured problems [17], yet they often face difficulties in navigating and regulating collaborative processes [29]. Addressing these challenges requires expert analysis of student dialogues, a notably labor-intensive process. AI-driven methods offer promising alternatives by enabling the automatic monitoring and analysis of collaborative problem-solving behavior, thus supporting tailored interventions. Previous approaches have primarily focused on recognizing collaborative problem-solving dialogue acts [9, 18] or analyzing utterance-level transitions mapped to collaborative problem-solving categories [15] and actions [4, 20, 34, 31]. However, these approaches often struggle to capture higher-level, long-range patterns and to connect behaviors to learning outcomes. Deep sequential models like Long Short-Term Memory networks [14] can model temporal dependencies but tend to overfit due to the variability and unpredictability in student dialogue. Compounding the issue, their "black box" nature limits interpretability in linking students' collaborative problem-solving behaviors and their learning outcomes. Temporal clustering has shown potential to uncover patterns in dialogue data. However, its lack of interpretability has led many studies to focus on discrete collaborative problem-solving events or their transitions over broader temporal trends, typically using statistical methods [15, 31]. Large language models (LLMs) can generate intuitive interpretations, but their use in real-time learning environments is limited by computational cost, latency, and data privacy concerns.

To address these challenges, we propose a novel collaborative problem-solving dialogue analysis framework that integrates temporal clustering of collaborative problem-solving dialogue patterns with LLM-generated explanations: LLM-Enhanced Dialogue Clustering and Interpretation (LEDI). The LEDI collaborative dialogue analysis framework bridges the gap between temporal data analysis and human-readable, domain-specific interpretation. Crucially, it enhances scalability, reduces computational cost, and preserves data privacy by eliminating the need for real-time LLM inference at deployment. By increasing the transparency and usability of collaborative problem-solving dialogue analysis, LEDI enables automated collaboration assessment in collaborative game-based learning environments and lays the foundation for improving collaborative learning processes and outcomes. In this paper, we investigate three research questions: (RQ1) How effective are temporal clustering methods in identifying meaningful patterns within collaborative problem-solving dialogue?; (RQ2) How accurately can LLMs interpret and evaluate clustered dialogues to provide insightful and domain-specific explanations?; and (RQ3) Can the evaluation of clustered dialogue quality reliably predict students' learning outcomes?

2 Related Work

Collaborative learning encompasses both cognitive and social dimensions. The cognitive aspect focuses on problem solving, while the social aspect emphasizes interaction and cooperation [12, 22]. In computer-supported collaborative learning (CSCL) environments, collaborative learning has been strongly associated with improved learning outcomes, as it enables students to support and guide one another, fostering both individual and collective learning [26]. CSCL cultivates collaborative problem-solving skills through pedagogical strategies such as problem-based and inquiry-based learning [13, 28]. Notably, students in game-based learning environments tend to achieve better learning outcomes compared to those who learn individually [5]. However, understanding and effectively supporting collaborative learning behaviors in these environments remains a challenge due to their inherently complex and dynamic nature [16].

To systematically analyze collaborative problem-solving behaviors, Liu et al. [21] proposed a discursive collaborative problem-solving framework for science education in collaborative learning environments. They emphasized that collaborative problem-solving skills are predominantly demonstrated through face-to-face and text-mediated dialogues. This framework delineates four core dimensions of collaborative problem solving: (1) *sharing ideas*, which is exchanging task-relevant information, ideas, and resources; (2) *negotiating ideas*, which is expressing agreement or disagreement with supporting evidence to reach consensus; (3) *regulating problem solving*, which is monitoring, reflecting, and coordinating collaborative efforts; and (4) *maintaining positive communication*, which is fostering a supportive and productive group dynamic. This framework has been foundational for numerous collaborative problem-solving investigations, including the present work.

Recent advances in natural language processing (NLP) techniques have significantly enhanced dialogue analysis in CSCL [1]. While earlier research focused on statistically analyzing sequential dialogue behaviors [7, 15] and problem-solving actions [4, 20, 34] to assess students' collaborative problem-solving competencies, recent studies have increasingly advanced NLP techniques, such as deep neural networks and Transformer-based methods, to tackle a broad range of collaborative learning analytics tasks. These include learning outcome prediction [11], detecting off-task behaviors [2], and identifying disruptive discourse [25, 8] through the analysis of students' interactions and dialogue data. More recently, large language models (LLMs) have gained attention for their generative capability and generalizability in education [30, 19], as they are trained on vast amounts of natural language data. Researchers have expanded dialogue analysis capabilities including dialogue act recognition [24, 18], problem behavior diagnosis [6], and inter-sentential relation analysis, such as temporal, causal, and dialogue relations [3]. Additionally, Markov Chain analysis has been integrated with LLM-based dialogue interpretation to examine transitions between two collaborative problem-solving events across multiple utterances [31]. Despite advances in NLP-driven collaborative problem-solving analysis, limited work has explored the role of consecutive temporal dynamics in collaborative problem-solving dialogue, which is crucial for capturing nuanced collaborative problem-solving patterns. Our proposed approach introduces a framework for understanding high-level temporal structures in collaborative problem-solving dialogue, which can lead to more effective adaptive support mechanisms in collaborative learning environments.

3 Crystal Island: EcoJourneys

The Crystal Island: EcoJourneys is a collaborative game-based learning environment designed to support middle school students in developing life science knowledge and collaborative problem-solving skills [28]. In the game, teams of three to four students investigate the cause of tilapia illness, communicating either in person or via an in-game chat. The game includes a tutorial and three quests, each featuring an individual Investigation phase, where students gather and analyze data on water quality and aquatic ecosystems, followed by collaborative problem-solving tasks. In the Deduce phase, students collaborate to answer multiple-choice questions, based on their findings, sharing ideas, negotiating, and resolving knowledge gaps, reaching a consensus before submitting their response. In the Talk, Investigate, Deduce, and Explain (TIDE) phase, students use a collaborative real-time collaboration tool to assess whether a given explanation is tenable by organizing evidence into consistent or inconsistent columns and discussing their reasoning. Through discussion, they evaluate and select relevant evidence to reach a consensus. This study analyzes students' collaborative dialogue patterns across the three quests, examining both in-person and in-game communication.

Data Collection and Data Sources. The classroom involved 67 consenting middle school students (grades 6-8, ages 11-14) distributed across 17 groups in seven science classes from four middle schools. Most of the participants (96%, 64 students) identified as White, and 46% identified as female and 49% as male, and the remaining participants selected other options. During gameplay, groups of three to four students collaborated using individual laptops. In-person and in-game conversations were captured through video and trace log data, respectively. After transcribing the in-person conversations across the three quests, we combined the transcribed and in-game dialogue, yielding 7,371 utterances. An utterance, $\boldsymbol{u}$, is defined as the sentences spoken or typed within a single speaker's turn, with examples shown in Table 1. On average, each group produced 433.6 utterances (SD=237.2), while each quest averaged 141.6 utterances (SD=44.3). A pre-and post-assessment on life science was administered on the first and last day of each implementation, respectively. A paired t-test shows a significant increase from the pre-test (mean=22.1) to the post-test (mean=24.6), $t(N-1)$=-3.83, p=0.0003, suggesting that the gameplay had an effect on the learning outcome. Data collection was conducted with Institutional Review Board (IRB) approval for research involving human subjects.

4 Method

The LEDI collaborative dialogue analysis framework (Fig. 1) involves four stages during the training phase: 1) temporal clustering of collaborative problem-solving dialogues, 2) LLM-based cluster interpretation, 3) cluster quality labeling, and 4) dialogue interpretation and prediction of learning outcomes. In the test phase,

Table 1. Examples of collaborative problem-solving utterances, dialogue sequences, cluster IDs, cluster quality labels, and cluster quality score.

u_i Student		Utterance	Dialogue Sequence	Cluster ID	Quality Label	Quality Score
u_1	Student1	The first one as compared to the bottom of the clean tank. That's the first question. There is dissolve oxygen at the bottom of the dirty tank.	$S_1{=}[u_1]$	c_1	Exemplary	2
u_2	Student2	Tha so, there's less at the bottom of the dirty tank.	$S_2{=}[u_1, u_2]$	c_2	Developing	1
u_3	Student1	Oh wait it out. So I thought it talks about the dirty tank. So, sort of comparing the dirty and the clean tank.	$S_3{=}[u_1, u_2, u_3]$	c_2	Developing	1
u_4	Student1	Yes, there's less dissolved oxygen because the bottom of the clean tank is 6.2 and the bottom of the dirty tank is 3.7.	$S_4{=}[u_1, \ldots, u_4]$	c_1	Exemplary	2
u_5	Student2	For the second one, it's photosynthesis?	$S_5{=}[u_1, \ldots, u_5]$	c_1	Exemplary	2

LLM-based cluster interpretation is omitted, since all clusters have already been interpreted; dialogue sequences are assigned cluster IDs and quality labels, and the average cluster score is calculated for prediction.

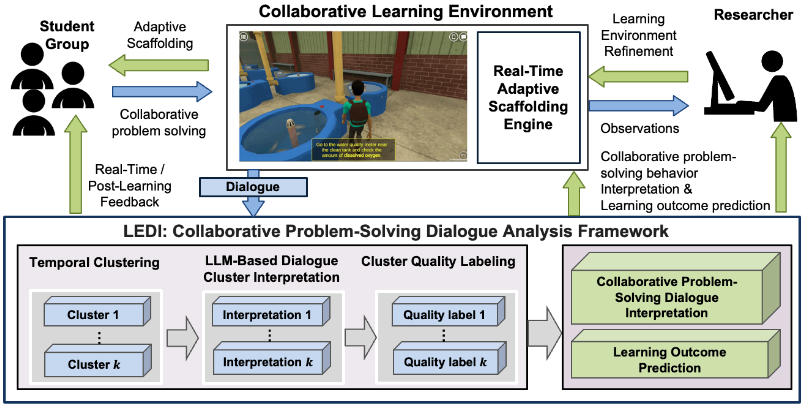
Student Group
Adaptive Scaffolding
Collaborative problem solving
Collaborative Learning Environment
Go to the water quality meter near the clean tank and check the amount of dissolved oxygen.
Real-Time Adaptive Scaffolding Engine
Learning Environment Refinement
Researcher
Observations
Collaborative problem-solving behavior Interpretation & Learning outcome prediction
Real-Time / Post-Learning Feedback
Dialogue
LEDI: Collaborative Problem-Solving Dialogue Analysis Framework
Temporal Clustering
Cluster 1
⋮
Cluster k
LLM-Based Dialogue Cluster Interpretation
Interpretation 1
⋮
Interpretation k
Cluster Quality Labeling
Quality label 1
⋮
Quality label k
Collaborative Problem-Solving Dialogue Interpretation
Learning Outcome Prediction

Fig. 1. The LEDI collaborative dialogue analysis framework.

Problem definition. For interpretable temporal clustering, given an input dialogue sequence, S_i, consisting of a sequence of utterances, our goal is to group S_i into a temporal cluster $c \in C$, denoted S_c, and provide a language description, $\mathcal{D}_c$, for every cluster for interpretation. Then, for cluster quality evaluation, each temporal cluster is assigned a quality label $q \in Q$ among Emerging, Developing, and Exemplary. Lastly, for early prediction, given a dialogue sequence S_i is classified into a higher- or lower-performing group.

Temporal Clustering of Dialogues. We train temporal clusters using dialogue sequences. To capture semantic similarity, we encode each utterance, $\boldsymbol{u}$, in a fixed-dimensional vector using Sentence Transformers [27]. These vectors are concatenated into l-length sequences, $S_i = [u_{i-l+1}, ..., u_{i-1}, u_i]$, to represent a single dialogue sequence as context. We then apply Dynamic Time Warping (DTW) to cluster $S_i \in \mathcal{S}$ into k groups, addressing variations in dialogue content, length, speed, and timing. DTW nonlinearly aligns utterances, preserving meaningful patterns for higher-level dialogue analysis. While clustering captures dialogue similarity, cluster meanings are interpreted in the next step.

LLM-Based Dialogue Cluster Interpretation. We interpret the learned clusters using an LLM. Algorithm 1 generates natural language descriptions of collaborative problem-solving dialogue clusters using an LLM. Given a subset of dialogue sequences, $\mathcal{S}_c$, from cluster c, it generates a text description, $\mathcal{D}_c$, which summarizes the cluster c's collaborative problem-solving traits identified by the clustering model. The process consists of three steps: (1) randomly sampling dialogue sequences from cluster c, (2) generating descriptions for each sample batch, and (3) repeating this process m times to derive a final summary and title of the cluster. The iterative approach enhances the representation of the trained clustering model by capturing broader dialogue patterns.

More specifically, in each iteration i, a random sample batch, $S_{c,i}$, of size r is drawn from the clustered sequences $\mathcal{S}_c$ (line 3). The LLM then extracts common characteristics using in-context learning with a prompt that includes only samples ($S_{c,i}$) or samples along with collaborative problem-solving knowledge ($\mathcal{K}$). We define three variants: (1) **LEDI-Base**, which includes samples only and excludes $\mathcal{K}$; (2) **LEDI-CPS**, which incorporates both samples and the collaborative problem-solving (CPS) framework description; and (3) **LEDI-Expert**, which includes samples, the CPS framework description, and expert-defined CPS behavior focuses (line 4). After m iterations, the algorithm summarizes the descriptions, $[D_{c,1}, ..., D_{c,m}]$ into a final summary D_c (line 6), enhancing the interpretability of the clustering model. This clustering interpretation step enables each collaborative problem-solving group's dialogue to be represented as a sequence of collaborative problem-solving contexts.

Algorithm 1 LLM-Enhanced Dialogue Interpretation Algorithm.

Input: clustered dialogue sequences $\mathcal{S}$, domain knowledge $\mathcal{K}$
Parameter: cluster number k, random sample size r, max iteration m
Output: cluster description, $\mathcal{D}$

```
1: for $c \in [1, k]$ do
2:     for $i \in [1, m]$ do
3:         $S_{c,i} \leftarrow$ randomSample($S_c, r$)
4:         $D_{c,i} \leftarrow$ extractCommonPatterns(LLM, $S_{c,i}$, $\mathcal{K}$)
5:     end for
6:     $\mathcal{D}_c \leftarrow$ summarize(LLM, $D_{c,1}, ..., D_{c,m}$)
7: end for
```

To define the focus of expert-informed CPS behavior, an educational expert identified eight key dialogue patterns distinguishing *exemplary* and *emerging* CPS behaviors. Emerging patterns reflect CPS behaviors with room for improvement: (1) insufficient participation: limited utterances that prevent meaningful engagement; (2) superficial negotiation: arguments that lack evidence or reasoning; (3) wheel-spinning conversations: repetitive or off-topic talk with unclear goals; and (4) expedited task completion: prioritizing finishing tasks quickly over quality dialogue. In contrast, exemplary patterns include: (1) comprehensive contribution: active discussion of diverse task aspects; (2) divergent idea sharing: exchange of diverse and relevant ideas; (3) evidence-based negotiation: reasoned dialogue supported by evidence; and (4) socially shared regulation: group-level planning, strategies, and reflection. Based on these patterns, we constructed a prompt for LEDI-Expert to interpret clusters and support its cluster quality labeling process, described in the next paragraph.

Cluster Quality Labeling. To systematically label the collaborative problem-solving quality of dialogue clusters, we employ two approaches: an LLM-based method that evaluates cluster descriptions (LEDI-Base/LEDI-CPS/LEDI-Expert), and a data-driven method, **LEDI-CR** (Cluster Ratio), that assigns quality labels based on each cluster's prevalence in higher- and lower-performing groups. In the LLM approach, a prompt provides each cluster's description obtained from the previous stage, asking an LLM to assign a quality label: Emerging, Developing, and Exemplary. For the data-driven approach, we devise a metric based on the ratio of clustered dialogue sequences in higher- and lower-performing groups to assign labels as follows: $Exemplary : r_H > r_L + \epsilon$, $Developing : (r_H > r_L - \epsilon) \,\&\, (r_H \leq r_L + \epsilon)$, and $Emerging : r_H \leq r_L - \epsilon$ where r_H and r_L represent the ratios of cluster instances from higher- and lower-performing groups, respectively. The parameter ϵ defines a buffer zone in which a cluster's occurrence frequency remains similar between the two groups, classifying it as Developing.

Learning Outcome Prediction. Evaluated clusters can be used for early prediction of learning outcomes by assessing the average quality of n-cumulative utterances from the start of each collaborative problem-solving task. In this work, we explore $n = [10, 20, 30, 50, 100, 200, all]$ to examine how different context lengths impact prediction performance. We frame the task as a binary classification, grouping dialogues into higher- and lower-performing categories. Higher-performing groups are defined as those whose average learning gain exceeds the median of group-level learning gain of the training data, while the remaining groups are classified as lower-performing. The quality labels, Emerging, Developing, and Exemplary, are assigned to collaborative problem-solving scores of 0, 1, 2, respectively. The average quality score of a sequence of utterances (i.e., target dialogue) is calculated as the mean of cluster quality scores. If the score exceeds the median score identified from the training data, the target dialogue is predicted as higher-performing; otherwise, it is predicted as lower-performing.

5 Experiments

We conduct 5-fold group-level cross-validation across four stages: temporal clustering, cluster interpretation, cluster quality labeling, and learning outcome prediction. For **temporal clustering of dialogues**, we produce utterance embeddings using Sentence Transformer (MiniLM, 384 dimensions) [33] and perform hyperparameter optimization via expert evaluation with grid search, using the number of clusters k=[5, **10**, 20] and dialogue sequence length l=[5, 10, **20**, 30] for DTW, along with the number of dialogue samples for interpretation r=[20, **30**, 40], with the optimal hyperparameters marked in bold.

For **LLM-based dialogue cluster interpretation**, LEDI-Base, LEDI-CPS, and LEDI-Expert generate interpretations for each cluster. As a preliminary analysis suggested that LEDI-Expert produced the most reliable interpretations, a domain expert conducted an additional human evaluation of LEDI-Expert's interpretations to further examine its accuracy and coherence.

For **cluster quality labeling**, we assign a quality label to each cluster using LEDI-Base, LEDI-CPS, LEDI-Expert, and LEDI-CR. To further compare these approaches, we measure the cross-method similarity among cluster quality labels assigned by a human expert (human labeling based on samples per cluster), LEDI-Expert (the best interpretation-based approach), and LEDI-CR (data-driven). For LEDI-CR, we search ϵ of [0.005, 0.01, **0.02**, 0.03]. Similarity is calculated as the ratio of the clusters with matching quality labels to the total number of clusters between any two clustering methods.

Lastly, for **learning outcome prediction**, we compare our clustering-based approaches, variants of LEDI, with two non-clustering baselines (i.e., directly analyzing utterances): LSTM and LLM-Expert, whose hyperparameters are also optimized via grid search. For LSTM, we use Sentence Transformer embeddings (MiniLM, 384 dimensions) [33] and explore the following hyperparameters: sequence length = [20, 30, **40**, 50], hidden unit size = [16, **32**, 64, 128], and batch size =[32, 64, **128**, 256]. For LLM-Expert, we apply few-shot in-context learning, using a target dialogue sequence with [1,3,**5**] dialogue examples from higher- and lower-performing groups, incorporating expert knowledge, $\mathcal{K}$. GPT-4o-mini [23] is used as the base LLM for both LLM-Expert and LEDI-based methods.

6 Results

Temporal Clustering. The optimal number of clusters is determined through expert evaluation on LLM-generated interpretations to distinguish their characteristics. With 5 clusters, multiple collaborative problem-solving dialogue patterns were grouped into a single cluster, while 20 clusters led to overlap in interpretation, making 10 the optimal balance. For dialogue sequence length, 5-utterance sequences lacked context, whereas 30-utterance sequences often combined multiple situations, reducing clarity. Both 10- and 20-utterance sequences were viable, but domain experts favored 20 for a more comprehensive analysis. In the 5-fold training datasets, 10 clusters with 20-utterance sequences yielded cluster sizes ranging from 19 to 1,705 utterance sequences, with a mean of 576.8.

LLM-Based Dialogue Cluster Interpretation. Table 2 presents examples of LLM interpretation for collaborative problem-solving dialogue clusters. The expert assessed these interpretations by analyzing dialogue samples within each cluster, highlighting both strengths and limitations of LLMs. For strengths, LLMs effectively clarify the overall context and dialogue quality, particularly for emerging-level clusters, by identifying unproductive collaborative problem-solving behaviors and irrelevant scenarios. However, as limitations, LLMs often lack descriptive precision, using vague terms like 'varying participation' or 'conflicts' that obscure details, making descriptions difficult to understand or apply in specific contexts. LLMs also struggle with understanding the learning environment's task-related elements (e.g., moving notes to represent an answer) and misinterpreting them as out-of-domain content.

Table 2. Examples of LEDI-Expert interpretation of CPS dialogue clusters.

Dialogue Cluster	LLM-Driven Interpretation
Imbalanced participation with superficial negotiation	Engage in problem solving with varying participation. Some contribute meaningfully while others hesitate, leading to superficial idea negotiation. Frustration and impatience occasionally surface, hindering deeper exploration.
Hesitation and disputes in problem solving	Participate inconsistently with hesitation, causing confusion and disputes. A mix of reasoning attempts exists, but not all contributions are well-supported, leading to misunderstandings.
Evidence-based collaboration with dynamic engagement	Engage in dynamic problem solving on scientific concepts, referencing learning materials with varying participation. Emphasizing reasoning and evidence, they strive for consensus despite frustrations, maintaining respectful communication.

Cluster Quality Labeling. Table 3 presents the collaborative problem-solving quality labeling results from human expert (H-Expert), LEDI-Expert, and LEDI-CR in the first set of training data (14 groups). The random guess is 33.3% among three labels. The highest similarity (67%) was observed between H-Expert and LEDI-Expert, indicating a close alignment between human expert evaluation and LLM-generated interpretations, followed by LEDI-Expert and LEDI-CR (50%). In contrast, H-Expert and LEDI-CR exhibited the lowest similarity (33%), highlighting a divergence in clustering labeling approaches. While H-Expert and LEDI-Expert rely on qualitative descriptions, LEDI-CR uses a quantitative, data-driven method based on behavior frequency in higher- and lower-performing groups. This methodological distinction led to notable discrepancies, particularly in Exemplary clusters.

Fig. 2 illustrates average collaborative problem-solving dialogue quality scores for higher- and lower-performing groups across three quests, evaluated by H-Expert and LEDI-Expert. In the H-Expert evaluation, both groups start with similar quality scores but diverge over time, with higher-performing groups improving and lower-performing groups slightly declining. LEDI-Expert, which ap-

Table 3. Cluster quality label examples of collaborative problem-solving dialogue.

Size	Title	H-Expert	LEDI-Expert	LEDI-CR
293	Imbalanced participation with superficial negotiation	Emerging	Developing	Developing
928	Dominant voices and tension in negotiation	Developing	Developing	Developing
1518	Scientific dialogue with varying clarity	Developing	Developing	Emerging
907	Hesitation and disputes in problem solving	Emerging	Emerging	Emerging
1082	Evidence-based collaboration with dynamic engagement	Exemplary	Exemplary	Emerging
1099	Active in sharing information and regulating problem	Exemplary	Developing	Emerging

plies stricter criteria (higher=0.8, lower=0.5) compared to H-Expert (higher=1.1, lower=1.0), assigns consistently lower scores than H-Expert. Both groups show a decline in dialogue quality over the quests, though higher-performing groups maintain better scores than lower-performing ones. This suggests that H-Expert and LEDI-Expert use different criteria but both distingush performance level.

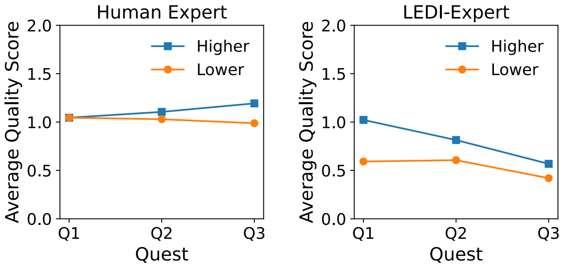

Fig. 2. Average Quality Score of dialogue sequences in higher- and lower-performing groups across quests, evaluated by human expert (H-Expert) and LEDI-Expert.

Learning Outcome Prediction Table 4 presents prediction accuracy as the utterance window expands from 10 to all utterances within a collaborative problem-solving task. Among non-clustering-based approaches, LLM-Expert outperforms the LSTM baseline by an average of 2.5% points. Among the cluster-based approaches, LEDI-CPS produces 46.9% accuracy, underperforming relative to the LSTM baseline. However, by incorporating expert-derived collaborative problem-solving behavior focuses, LEDI-Expert (71.7%) improves prediction accuracy by an average of 15.4% points over LSTM, while LEDI-CR achieves the highest accuracy (75.4%) in this evaluation. In sum, the clustering-based methods with expert knowledge outperform the non-clustering approaches in prediction while also enhancing the interpretability of dialogue sequences.

Table 4. The average accuracy of learning outcome early prediction by increasing utterances within a collaborative problem-solving task. * indicates a significant difference from the baseline LSTM based on a Wilcoxon rank sum test ($p < 0.05$).

Utterances	10	20	30	50	100	200	All	Mean	Diff. (%)
LSTM	58.4	50.4	**56.2**	58.2	56.6	**59.5**	**54.9**	56.3	0
LLM-Expert	**64.7**	**64.7**	52.9	**58.8**	**58.8**	58.8	52.9	**58.8**	2.5
LEDI-Base	51.7	45.0	45.0	51.7	45.0	45.0	45.0	46.9	-9.4
LEDI-CPS	61.7	*61.7	55.0	55.0	*55.0	55.0	*55.0	56.9	0.6
LEDI-Expert	*71.7	*71.7	*65.0	*71.7	*65.0	***78.3**	***78.3**	*71.7	15.4
LEDI-CR	***78.3**	***78.3**	***78.3**	**73.3**	***73.3**	73.3	73.3	***75.4**	19.1

7 Discussion

Our findings address the three research questions. For RQ1 (How effective are temporal clustering methods in identifying meaningful patterns within collaborative problem-solving dialogue?), temporal clustering effectively identifies distinct collaborative problem-solving dialogue patterns, as validated by domain expert evaluations and by the high accuracy of learning outcome predictions based on the resulting clusters. Regarding RQ2 (How accurately can LLMs interpret and evaluate clustered dialogues, providing insightful and domain-specific explanations?), LEDI-Expert, an expert knowledge-augmented approach, significantly enhances interpretability, achieving alignment 67% with human expert assessments. Notably, when collaborative problem-solving behaviors were quantified in quality scores based on cluster interpretation, a significant difference was observed between higher- and lower-performing groups. In response to RQ3 (Can the evaluation of clustered dialogue quality reliably predict students' learning outcomes?), given clustered collaborative problem-solving dialogues, LEDI-CR (75.4%) and LEDI-Expert (71.7%) successfully predict learning outcomes, outperforming competitive baselines (LSTM and LLM-Expert). Moreover, since analyzing new utterances relies only on light-weighted Sentence Transformers for embedding extraction, eliminating cost-ineffective LLM use during testing, our framework retains its analytical benefits while improving collaborative problem-solving analysis scalability in educational settings where LLM use is limited.

In the prediction task, the superior performance of the cluster-ratio-based method (LEDI-CR) over the interpretation-based methods (LEDI-Expert and LEDI-CPS) stems from several factors. First, while collaborative problem-solving behavior correlates with learning gains, they are not the sole determinant. Effective collaborative problem solving requires both collaboration and problem-solving skills, incorporating various factors such as group dynamics and prior knowledge in constructing appropriate solutions [32]. Second, LLM prompting can be improved by including an indicator of active participation in collaborative problem solving (e.g., the amount of group dialogue), which serves as a key collaborative problem-solving metric. Third, LLMs' reliance on general terms to describe common dialogue characteristics reduces cluster specificity and increases ambiguity, highlighting a limitation of natural language in conceptualizing data. These findings suggest that integrating data-driven metrics with dialogue interpretation offers greater synergy than relying solely on one approach.

Evaluation was conducted on a MacBook Pro (Apple M1 Max, 32GB RAM, no GPU), with total training time of approximately 15 minutes. At runtime, the model responds in under 0.02 seconds, about 50 faster than typical LLM inference (≈1 second). LEDI uses ≈115K tokens during the design phase for cluster interpretation and evaluation via LLM, with no LLM usage at runtime. In contrast, runtime LLM usage (450 tokens per request) would reach 115K tokens after 255 calls, highlighting the model's efficiency for real-time deployment.

8 Conclusion

Analyzing student dialogue in collaborative learning environments is essential for evaluating collaborative problem-solving behaviors, predicting learning outcomes, and enabling adaptive feedback. However, the inherent complexity of students' dialogue makes effective analysis challenging. To address this, we propose a collaborative dialogue analysis framework that integrates temporal clustering with LLM-generated interpretations to analyze collaborative problem-solving dialogues. Our findings highlight the potential of combining LLMs with expert insights to automate the interpretation and assessment of collaborative problem-solving dialogue, reducing reliance on labor-intensive manual evaluations. Moreover, our framework demonstrates that clustered collaborative problem-solving dialogues can predict student groups' learning outcomes with high accuracy, outperforming competitive baselines. These results underscore the promise of AI-driven methods in collaborative problem-solving assessment and learning outcome prediction, offering valuable methods for educators and researchers to better understand and support collaborative learning.

Future research should explore combining LLMs with data-driven approaches to explain the specific relationship between dialogue patterns and learning outcomes to gain deeper and interpretable insights. Additionally, exploring alternative prompt engineering techniques could increase interpretability by incorporating broader context information such as tasks in the game and students' pretest knowledge. Applying this framework to game-based learning environments could further enhance adaptive learning systems, improving students' collaborative problem-solving experience through real-time feedback. Ultimately, by advancing our understanding of collaborative problem-solving dialogue, this work contributes to the development of more effective adaptive collaborative learning environments.

Acknowledgments. This research was supported by funding from the National Science Foundation under grants DRL-2112635. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Borchers, C., Yang, K., Lin, J., Rummel, N., Koedinger, K.R., Aleven, V.: Combining dialog acts and skill modeling: What chat interactions enhance learning rates during ai-supported peer tutoring? In: PaaÃŸen, B., Epp, C.D. (eds.) Proceedings of the 17th International Conference on Educational Data Mining. pp. 117–130. International Educational Data Mining Society, Atlanta, Georgia, USA (July 2024). https://doi.org/10.5281/zenodo.12729784
2. Carpenter, D., Emerson, A., Mott, B.W., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Lester, J.C.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Proceedings, Part I. p. 55–66. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-52237-7_5
3. Chan, C., Jiayang, C., Wang, W., Jiang, Y., Fang, T., Liu, X., Song, Y.: Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In: Graham, Y., Purver, M. (eds.) Findings of the Association for Computational Linguistics: EACL 2024. pp. 684–721. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024)
4. Chang, C.J., Chang, M.H., Chiu, B.C., Liu, C.C., Chiang, S.H.F., Wen, C.T., Hwang, F.K., Wu, Y.T., Chao, P.Y., Lai, C.H., et al.: An analysis of student collaborative problem solving activities mediated by collaborative simulations. Computers & Education **114**, 222–235 (2017)
5. Chen, C.H., Law, V.: Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. Comput. Hum. Behav. **55**(PB), 1201–1212 (Feb 2016)
6. Chen, P., Fan, Z., Lu, Y., Xu, Q.: Pbchat: Enhance student's problem behavior diagnosis with large language model. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 32–45. Springer Nature Switzerland, Cham (2024)
7. von Davier, A.A., Hao, J., Liu, L., Kyllonen, P.: Interdisciplinary research agenda in support of assessment of collaborative problem solving: Lessons learned from developing a collaborative science assessment prototype. Computers in Human Behavior **76**, 631–640 (2017)
8. Earle-Randell, T.V., Wiggins, J.B., Ruiz, J.M., Celepkolu, M., Boyer, K.E., Lynch, C.F., Israel, M., Wiebe, E.: Confusion, conflict, consensus: Modeling dialogue processes during collaborative learning with hidden markov models. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 615–626. Springer Nature Switzerland, Cham (2023)
9. Glass, M., Kim, J., Parham, J., Banks, J.: Machine identification of collaboration dialogue acts in typed-chat collaborative problem-solving. In: EDULEARN20 Proceedings. p. 6089. 12th International Conference on Education and New Learning Technologies, IATED (6-7 July, 2020 2020). https://doi.org/10.21125/edulearn.2020.1597
10. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the science of collaborative problem solving. Psychological science in the public interest **19**(2), 59–92 (2018)
11. Gupta, A., Carpenter, D., Min, W., Mott, B., Glazewski, K., Hmelo-Silver, C.E., Lester, J.: Enhancing stealth assessment in collaborative game-based learning with multi-task learning. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 304–315. Springer Nature Switzerland, Cham (2023)

12. Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P.: A framework for teachable collaborative problem solving skills. Assessment and teaching of 21st century skills: Methods and approach pp. 37–56 (2015)
13. Hmelo-Silver, C.E., Chernobilsky, E.: Understanding collaborative activity systems: the relation of tools and discourse in mediating learning. In: Proceedings of the 6th International Conference on Learning Sciences. p. 254–261. ICLS '04, International Society of the Learning Sciences (2004)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (nov 1997)
15. Hong, D., Feng, C., Zou, X., Hmelo-Silver, C., Glazewski, K., Wang, T., Mott, B., Lester, J.: Examining coordinated computer-based fixed and adaptive scaffolds in collaborative problem-solving game environments. In: Proceedings of the Seventeenth International Conference on Computer-Supported Collaborative Learning. pp. 43–50. Buffalo, New York (2024)
16. Hur, P., Bosch, N., Paquette, L., Mercier, E.: Harbingers of collaboration? the role of early-class behaviors in predicting collaborative problem solving. In: Proceedings of the 13th International Conference on Educational Data Mining (EDM). pp. 104–114 (2020)
17. Jeong, H., Hmelo-Silver, C.E., Jo, K.: Ten years of computer-supported collaborative learning: A meta-analysis of cscl in stem education during 2005–2014. Educational research review **28**, 100284 (2019)
18. Kim, Y.J., Acosta, H., Min, W., Rowe, J., Mott, B., Chaturvedi, S., Lester, J.: Dual process masking for dialogue act recognition. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 15270–15283. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
19. Li, X., Henriksson, A., Duneld, M., Nouri, J., Wu, Y.: Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals: Using textbook content as context for retrieval-augmented generation with large language models. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 118–132. Springer Nature Switzerland, Cham (2024)
20. Lin, P.C., Hou, H.T., Chang, K.E.: The development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: An analysis of learners' performance and their cognitive process in discussion. Interactive Learning Environments **30**(7), 1273–1290 (2022)
21. Liu, L., Hao, J., von Davier, A.A., Kyllonen, P., Zapata-Rivera, J.D.: A tough nut to crack: Measuring collaborative problem solving. In: Handbook of research on technology tools for real-world skill development, pp. 344–359. IGI Global (2016)
22. OECD: Pisa 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving. PISA (2017)
23. OpenAI: Gpt-4o-mini (2023), https://platform.openai.com, accessed: 2025-01-26
24. Pande, J., Min, W., Spain, R.D., Saville, J.D., Lester, J.: Robust team communication analytics with transformer-based dialogue modeling. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 639–650. Springer Nature Switzerland, Cham (2023)
25. Park, K., Sohn, H., Min, W., Mott, B., Glazewski, K., Hmelo-Silver, C.E., Lester, J.: Disruptive talk detection in multi-party dialogue within collaborative learning environments with a regularized user-aware network. In: Lemon, O., Hakkani-Tur, D., Li, J.J., Ashrafzadeh, A., Garcia, D.H., Alikhani, M., Vandyke, D., Dušek, O. (eds.) Proceedings of the 23rd Annual Meeting of the Special Interest Group on

Discourse and Dialogue. pp. 490–499. Association for Computational Linguistics, Edinburgh, UK (Sep 2022). https://doi.org/10.18653/v1/2022.sigdial-1.47

26. Patil, P.V., T S, A., Rajendran, R.: Fostering interaction in computer-supported collaborative learning environment. In: Proceedings of the 16th International Conference on Educational Data Mining (EDM) (2023)
27. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing (2019), https://api.semanticscholar.org/CorpusID:201646309
28. Saleh, A., Hmelo-Silver, C.E., Glazewski, K.D., Mott, B., Chen, Y., Rowe, J.P., Lester, J.C.: Collaborative inquiry play: A design case to frame integration of collaborative problem solving with story-centric games. Information and Learning Sciences **120**(9/10), 547–566 (2019)
29. Savery, J.R.: Overview of problem-based learning: Definitions and distinctions. Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows **9**(2), 5–15 (2015)
30. Schmucker, R., Xia, M., Azaria, A., Mitchell, T.: Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 75–90. Springer Nature Switzerland, Cham (2024)
31. Snyder, C., Hutchins, N.M., Cohn, C., Fonteles, J.H., Biswas, G.: Analyzing students collaborative problem-solving behaviors in synergistic stem+c learning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. p. 540–550. LAK '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3636555.3636912
32. Sun, C., Shute, V.J., Stewart, A.E., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N., D'Mello, S.K.: The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. Computers in Human Behavior **128**, 107120 (2022). https://doi.org/https://doi.org/10.1016/j.chb.2021.107120
33. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
34. Zhou, Y., Kang, J.: Enriching multimodal data: A temporal approach to contextualize joint attention in collaborative problem-solving. Journal of Learning Analytics **10**(3), 87–101 (2023)