LLM-based Multimodal Representations of Student Collaboration in Mathematics

Shan Zhang[1[0009-0003-3532-0661]], Hongming Li[1[0000-0002-6024-677X]], Seiyon Lee[1[0009-0001-9041-3313]], Xiner Liu[2[0009-0004-3796-2251]], Ji-Eun Lee[3[0000-0001-8521-8997]], Noah L. Schroeder[1[0000-0002-3281-2594]], and Anthony F. Botelho[1[0000-0002-7373-4959]]

[1] University of Florida, Gainesville, FL 32608, USA
[2] University of Pennsylvania, Philadelphia, PA 19104, USA
[3] University of Technology and Design, Singapore 487372, Singapore
zhangshan@ufl.edu

Abstract. Assessing the quality of students' collaboration at scale remains a key challenge in computer-supported collaborative learning, especially when integrating multimodal data sources. From a Quantitative Ethnography (QE) perspective, this challenge centers on how to systematically represent and measure the structure of collaborative meaning-making across heterogeneous data streams**.** This study examines how large language models (LLMs) interpret different combinations of behavioral logs and dialogue transcripts to evaluate collaboration quality in mathematics problem solving. Grounded in a QE-informed analytic framework, we developed a theoretically grounded rubric encompassing social, cognitive, and regulatory dimensions. We then compared six LLMs from the Gemini and GPT families on their ability to evaluate the quality of students' collaboration using interaction log, dialogue, and human-coded label-based data sources from 22 college students (11 dyads) collaboratively solving 49 algebra problems in a digital learning platform. Results show that adding dialogue to log data improved agreement with expert ratings, while human-annotated dialogue labels offered no additional benefit. Dimension-level findings revealed stronger alignment on explicit communicative behaviors than on implicit indicators. Our findings highlight both the potential and limitations of LLMs for collaboration assessment from multimodal data sources.

Keywords: Collaborative Learning, Multimodal Data, Large Language Models (LLMs), Dialogue Analysis

1 Introduction and related works

Collaborative learning (CL) has long been recognized as an effective instructional approach that promotes deeper understanding and critical thinking by engaging students

in joint problem solving [41]. Rather than learning in isolation, students in CL environments exchange ideas, reason together, and co-construct knowledge through shared inquiry [5]. Extensive research has shown that CL is a dynamic, multidimensional process involving behavioral, cognitive, motivational, and socio-emotional components [3, 10]. CL has also been associated with improvements in students' cognitive outcomes, such as conceptual understanding and reasoning [19], and non-cognitive outcomes, including communication [12], engagement [22], and problem-solving skills [23].

Within computer-supported collaborative learning, these interactions are increasingly captured through multimodal process data such as dialogue, interaction logs, and other digital traces [17]. This expansion of multimodal data has also raised methodological questions about how multiple forms of process data can be integrated while preserving the distinct relationships and structures associated with each modality [37]. More fundamentally, rich process data, by itself, does not resolve how the collaborative learning process should be represented and interpreted across data sources. From a QE standpoint, this challenge centers on how to transform heterogeneous traces of interaction into structured representations of collaborative meaning that can be systematically compared across learners and contexts [2]. In this sense, the problem is not only how collaboration can be systematically analyzed, but whether the analytic representation preserves the relational structure through which collaborative meaning is constructed [36].

Building on this perspective, the use of multimodal data introduces both opportunities and additional challenges for representing collaborative processes. Analyzing process data related to collaboration, particularly multimodal data, is not a trivial task. Early studies analyzed dialogue data to capture how students verbalized and negotiated knowledge during group discussions [15]. More recent work has incorporated multimodal evidence, such as gaze trajectories through eye-tracking [35] and interaction behaviors extracted from digital log data [26]. While these additional data streams provide rich insights into collaborative processes, they are often unstructured in raw form and require extensive preprocessing and interpretation to become analytically meaningful. This difficulty reflects the challenge of constructing representations that preserve meaningful relations among actions, discourse, and context rather than reducing them to isolated indicators.

To address this challenge of making multimodal data interpretable, researchers have developed coding schemes and rubrics that identify meaningful units of collaboration. In this sense, a rubric operationalizes a theoretical account of practice by specifying what counts as meaningful action and how those actions are systematically interpreted. For instance, Lee et al. applied a collaborative competency rubric to capture cognitive and social aspects of CPS, including constructing shared knowledge, negotiation and coordination, and maintaining team function [25]. This rubric builds on the generalized competency model proposed by Sun et al., designed to apply across diverse contexts [38]. Similarly, Herro et al. developed a coding rubric for CPS in K–12 STEAM activities, extending applicability beyond STEM domains. However, rubric-based approaches are inherently resource-intensive, requiring significant expert training and

manual labeling [14]. This labor demand limits scalability and raises questions about how well rubrics generalize across different conditions or contexts.

Given these limitations, recent work has started to explore automated approaches to support or extend qualitative analysis, particularly through the use of large language models (LLMs). Researchers have applied LLMs to assist with deductive coding [4], showing strong alignment with human coders for well-structured constructs [28], but challenges remain. Specifically, recent work has shown that abstract or context-dependent coding that requires nuanced interpretation [44] can be problematic with LLMs, and there may also be systematic biases that are not visible even when overall agreement appears high [34].

Researchers have also begun applying LLMs to interpret behavioral traces and discourse in collaborative settings. For example, An et al. developed an LLM-empowered agent that monitored participation-frequency logs to prompt disengaged students [1], while Author et al. combined Gemini-2.0-flash and GPT-4o to transform mathematics problem-solving logs into interpretable narratives for a collaborative learning dashboard [24]. Other work has extended LLM-based analysis beyond logs to dialogue and written reflection. Yang et al. used GPT-4 to code student reflections, finding high reliability for concrete categories but lower consistency for codes requiring more complex reasoning [42]. Similarly, Lee et al. [25] and Otten et al. [33] leveraged GPT-based models for dialogue coding in collaborative problem solving, showing promising accuracy in detecting cognitive and social dimensions such as shared knowledge construction and transactivity. These studies suggest that LLMs can support the interpretation of behavioral and discourse data, but they do not yet establish how different representations of collaborative activity support rubric-based evaluations of group collaboration quality.

For rubric-based evaluations of collaboration quality, the key challenge is not only whether LLMs can automate expert-coded assessment, but whether the data provided to the model preserves the information needed to make those judgments. Behavioral logs, dialogue transcripts, and human-coded annotations each represent collaboration differently. Logs can show patterns of participation and tool use, while dialogue can provide evidence of reasoning, negotiation, and shared understanding. Human-coded annotations can make selected features of discourse more explicit, but they may also reduce the broader context available in the original interaction. From a QE perspective, these differences matter because each representation shapes what can be inferred about the quality of joint work. LLMs may support scalable approaches to expert-coded assessment because they can process unstructured, multimodal data within a unified semantic framework [32] and adapt across contexts with minimal feature engineering [25]. However, their usefulness depends on how well the input representation aligns with the collaborative processes being evaluated. This study explores how LLMs can be leveraged to scale the evaluation of pair collaboration based on log data, dialogue data, and human-coded annotations:

- RQ1: How does LLM grading performance on group collaboration quality vary across different combinations of input data (e.g., logs, logs with dialogue, and logs with dialogue and human-coded annotations)?

- RQ2: Which specific aspects of group-level collaboration quality are more accurately evaluated by the LLM under the optimal combination of input data?

2 Methods

2.1 Participants and study context

Participants in this study were 22 undergraduate students from a university in the northeastern United States. The group included 12 males (55%) and 10 females (45%). In terms of race and ethnicity, 14 students (64%) identified as White, 5 (23%) as Asian, and 3 (13%) as other categories. The majority of participants (95%, n = 21) reported majoring in STEM fields.

In this study, students worked in Graspable Math (GM, https://activities.graspablemath.com/), an interactive digital platform grounded in embodied cognition and perceptual learning theories. GM allows learners to manipulate math expressions directly through touch or mouse gestures. As shown in Figure 1, its shared whiteboard environment supports real-time collaboration, allowing peers to jointly construct, revise, and discuss math representations during problem solving.



Fig. 1. GM's collaborative whiteboard interface (*left*) and algebraic problem-solving process (drag & drop; *right*)

2.2 Measures

The pre-survey included two items assessing students' collaborative skills and prior experiences working with peers in mathematics, rated on a 4-point Likert scale (e.g., "What is your level of collaborative experience for learning mathematics?"). The post-survey comprised nine items evaluating collaborative learning satisfaction and overall perceptions of the group experience, measured on a 5-point Likert scale (e.g., "Collaborative problem-solving in mathematics is helpful."). For performance, collaborative problem-solving accuracy was scored dichotomously (0 = incorrect, 1 = correct) for each of the five problems, resulting in a total score from 0 to 5. The average group accuracy across all problems was 0.51.

2.3 Study procedure

This study was reviewed and approved by the Institutional Review Board. Participants were recruited through the university's research portal and received course credit for their voluntary participation. Students were randomly paired to form 11 dyads, each participating in a one-hour laboratory session facilitated by a research assistant. The session began with informed consent and a pre-survey capturing their demographic information and prior experience with collaborative learning. Subsequently, students viewed brief instructional videos introducing the GM platform and collaboratively completed two practice problems to become familiar with its interface and all interactive features. They then engaged in a 30-minute collaborative problem-solving activity involving five algebraic word problems, primarily focused on rate and mixture scenarios, within GM's shared whiteboard environment. Videos were recorded via Zoom for subsequent analysis. Each session concluded with a post-survey assessing students' perceptions of their collaborative problem-solving experience. In total, 11 collaborative groups completed 49 problems.

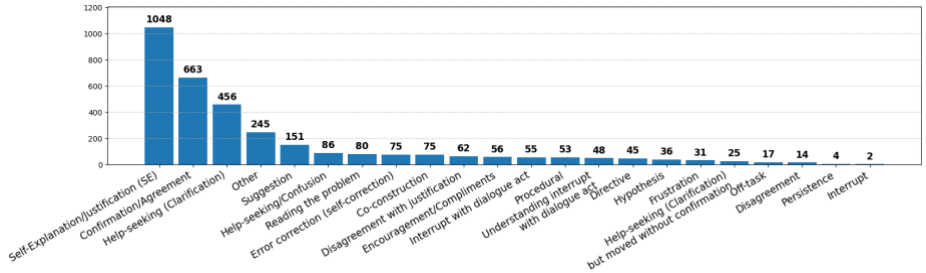


Fig. 2. Distribution of dialogue labels (N = 22)

2.4 Data analysis

We began by cleaning the transcripts, using the recorded video of students' problem-solving process to verify speaker turns and ensure each utterance was accurately segmented. To analyze how students interacted, we adapted a well-established dialogue coding framework from previous research [6, 43]. Two coders first reviewed one video together to clarify category definitions and establish a shared understanding of the dialogue acts. They then independently coded 20% of the data, achieving Cohen's kappa values above 0.75 across all categories. Any disagreements were resolved through social moderation, after which one coder completed the rest of the dataset [13]. As summarized in Figure 2, the finalized scheme includes 22 dialogue acts. These span cognitive processes such as Self-Explanation/Justification (SE), Hypothesis, and Suggestion; regulatory and metacognitive acts such as Help-Seeking (Clarification) and Error Correction (Self-Correction); and social-emotional expressions such as Confirmation/Agreement, Encouragement/Compliments, and Frustration. In total, students produced 3,327 coded utterances. The most frequent acts were Self-

Explanation/Justification (SE) (n = 1,048), Confirmation/Agreement (n = 663), and Help-Seeking (Clarification) (n = 456).

Table 1. Action definitions and frequencies

Action Type	Definition	Freq. (%)
Content Creation & Modification		
AlgebraAction	Performs an algebraic operation on an existing math expression (e.g., simplifying, expanding).	839 (37.09%)
CreateAction	Creates a new object on the canvas, such as a math expression (Derivation) or a text box (TextBox).	665 (29.40%)
TextEditAction	Modifies the text content within an existing text box.	393 (17.37%)
RewriteFirstLineAction	Intentional edit that rewrites the first line of a math expression.	7 (0.31%)
Deletion		
EraseAction	Uses the 'eraser' tool to remove strokes or parts of objects.	68 (3.01%)
DeleteAction	Selects and deletes one or more complete objects.	43 (1.90%)
RemoveDerivation	Removes one or more rows from a mathematical expression.	8 (0.35%)
Interface Interaction		
MoveAction	Moves the position of one or more objects on the canvas.	167 (7.38%)
CloneAction	Duplicates an existing object.	23 (1.02%)
PackAction	(Often automatic) Groups multiple elements together.	6 (0.27%)
ScrubAction	Interactively changes a value by dragging (uncommon).	1 (0.04%)
System & State Management		
MigrateCanvasAction	Background system action to update the canvas data version.	40 (1.77%)
LoadCanvasStateAction	Loads a previous canvas state (e.g., undo/redo operations).	2 (0.09%)
Total		2262 (100%)

Complementing the dialogue transcripts, we incorporated granular, time-stamped log data captured by the GM platform. This raw data, consisting of a chronological stream of complex, nested JSON objects for each student interaction, was processed through a rule-driven ETL (Extract, Transform, Load) [20] and feature engineering pipeline. The pipeline's goal was to transform these heterogeneous events into a canonical, semantically rich format for multimodal analysis. The process involved first mapping student names to unique identifiers through entity resolution. Subsequently, a rule-based parsing function canonicalized each event based on its action category (see Table 1 for description). Through domain-guided feature engineering, we extracted critical content, such as the specific mathematical expressions from creation events, or encoded the state transition by capturing the content before and after modification for editing actions. Interactions without direct content manipulation were simplified into

event markers (e.g., LoadCanvasStateAction). This resulted in a structured JSON array for each session where every event retains its original timestamp and student identifier, which allows for precise temporal alignment between a student's action and their concurrent utterances.

Regarding the measurement of collaboration quality, we used a collaboration rubric grounded in established theoretical frameworks, including Mercer's [29, 30] Exploratory Talk, Weinberger and Fischer's [39] Argumentative Knowledge Construction model, and Järvelä et al.'s [16] Socially Shared Regulation of Learning, and which has also been applied in our prior work ([43]; see Table 2).

The rubric captures collaboration quality across six dimensions-Participation Balance, Responsiveness & Listening, Idea Elaboration & Co-construction, Questioning & Clarification, Conflict Resolution & Negotiation, and On-task Focus & Shared Goals. Each dimension was rated on a three-point scale with an optional "N/A." Two coders jointly established coding consistency by independently rating 20% of the data and resolving discrepancies through discussion, achieving strong interrater reliability (Quadratic Weighted Kappa [QWK] = 0.84), after which one coder completed the remaining ratings. A total collaboration score per problem was computed by averaging across applicable dimensions.

To address RQ1, which examines how LLM performance varies across different input data combinations, we designed three multimodal input scenarios informed by theoretical and practical considerations in educational data analysis [7]. The Log Only scenario included only the canonicalized JSON array of student actions, simulating a fully automated data pipeline typical in learning analytics, where behavioral data from digital environments can populate dashboards [24] or inform educators without manual transcription [26**Error! Reference source not found.**]. As the most accessible and least resource-intensive data type, log data makes this scenario practical and scalable for large-scale classrooms [26]. Building on this baseline, the Log + Dialogue scenario combined action logs with time-stamped dialogue transcripts, reflecting research contexts where verbal data are collected to deepen insights into student sense-making and are grounded in corresponding actions. Finally, the Log + Dialogue + Precode scenario added human-annotated dialogue act codes (see Figure 2 for distribution) to test whether expert-level semantic information improves LLM assessment accuracy enough to justify the additional coding effort. For each of the 49 problem-solving sessions (11 groups, with most of the groups solving three or four problems), we constructed a tailored prompt describing the context and task, while providing the full collaboration rubric [45]. The complete prompt is available on the Open Science Framework (OSF) at https://osf.io/cesv6/overview?view_only=559728f9343848b78b1663955444db54.

Our model selection strategy evolved to balance technical feasibility and cross-model generalizability. The initial data processing pipeline generated extremely long inputs, with some sessions exceeding 200,000 tokens (well beyond most LLMs' context limits). We first selected Google's Gemini 2.5 models for their 1-million-token windows. However, to avoid bias toward a single architecture, we refined the pipeline to support cross-family comparison. By removing nested metadata and system traces while preserving meaningful action sequences and temporal order, we reduced input

length to a mean of 8,025 tokens (median = 6,093; range = 1,642–42,029), maintaining semantic richness. This optimization brought all sessions within the 128,000-token limit of other models, enabling inclusion of GPT as a second, structurally distinct family widely used in educational research [11].

The final analysis included six models spanning two families and three computational tiers each: Gemini-2.5-Pro, Flash, and Flash Lite; and GPT-5, Mini, and Nano. Each model's ratings were compared with expert scores across all three input scenarios. To address RQ2, we conducted a per-dimension rubric analysis, disaggregating agreement metrics by collaboration dimension within the best-performing input condition. This approach identifies the specific collaborative aspects LLMs evaluate most and least accurately, offering a nuanced understanding beyond aggregated performance scores.

To evaluate the agreement between LLM-generated and expert ratings, we computed two complementary metrics: macro F1 score for multi-class classification performance and Quadratic Weighted Kappa (QWK) for ordinal agreement with severity-weighted penalties [8]. To quantify the reliability of these estimates, we employed block bootstrap resampling (B = 2000) stratified by problem session to preserve within-group dependencies, resulting in 95% confidence intervals for all reported metrics. This approach accounts for the hierarchical structure of our data (ratings nested within collaborative sessions) and provides robust uncertainty estimates despite the relatively modest sample size (n = 294 ratings across 49 sessions). In addition, for treatment effect analysis, we computed pairwise performance deltas (Δ) between input scenarios using paired bootstrap to test whether observed differences exceed sampling variability.

Table 2. Collaboration quality rubric (from [45])

Dimensions	High (3)	Moderate (2)	Low (1)
Participation Balance	Equal contribution; ideas co-developed; turn-taking	Few dominant voices, others still contribute	One member dominates; others passive or silent
Responsiveness & Listening	Active listening; ideas built on and referenced	Some acknowledgment; limited topic building	Ideas ignored; no clear listening; disconnection
Idea Elaboration & Co-construction	Members elaborate, reason, and co-construct meaning	Some elaboration, mostly individual	Simple statements; little or no joint meaning-making
Questioning & Clarification	Frequent open-ended questions; collaborative clarification	Some simple or rhetorical questions; limited follow-up	Rare questioning; minimal clarification
Conflict Resolution & Negotiation	Differences explored respectfully; shared decisions reached	Conflicts noted but not fully resolved	Conflict avoided, suppressed, or escalated
On-task Focus & Shared Goals	Fully on-task; clear shared goals and collaboration	Mostly on-task; some distraction	Off-task; little evidence of shared goals or joint effort

Note. Each dimension includes an "N/A" option. Descriptions are abbreviated due to space limitations

3 Results

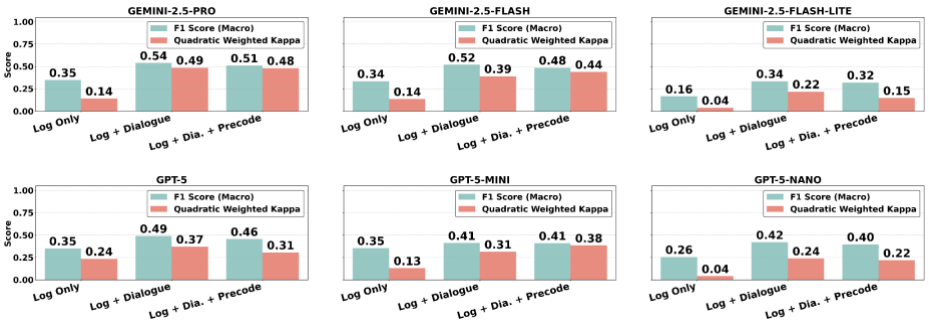


Fig. 3. Overall agreement metrics for all dimensions of six LLMs across three multimodal input scenarios

For RQ1, using expert ratings of collaboration quality as ground truth (M = 2.63; SD = 0.49; *Range* = 1.4-3), we evaluated LLM performance across the three input scenarios. Figure 3 presents the performance across all six models and three input scenarios. Performance in the Log Only scenario was uniformly poor, with models struggling to infer collaboration quality from behavioral traces alone. Gemini-2.5-Pro achieved a QWK of 0.14 and F1 of 0.35, while GPT-5 reached a QWK of 0.24 and F1 of 0.35. The lighter-weight models performed worse, with Gemini-2.5-Flash-Lite achieving QWK=0.04 and F1=0.16, barely exceeding chance-level performance. The wide confidence intervals, particularly for QWK, reflect the fundamental instability of assessments derived solely from action logs.

The introduction of dialogue data in the Log + Dialogue scenario yielded substantial improvements across all models. Gemini-2.5-Pro's performance increased from 0.14 to 0.49 on QWK and from 0.29 to 0.54 on F1. Similarly, GPT-5 moved from 0.24 to 0.37 on QWK and from 0.35 to 0.49 on F1. Notably, even the lightest models showed noticeable gains: Gemini-2.5-Flash-Lite's QWK improved from 0.04 to 0.22. However, contrary to initial expectations, adding human-generated precodes in the Log + Dialogue + Precode scenario did not yield further gains, in some cases causing even a slight decline in performance. For example, Gemini-2.5-Pro's QWK declined slightly from 0.49 to 0.48, while GPT-5 dropped from 0.37 to 0.31.

We conducted paired bootstrap analyses (B = 2000) to estimate the change in QWK (Δ) and corresponding 95% confidence intervals (CIs) for each pairwise scenario comparison. Results showed a clear positive impact of adding dialogue data to the log baseline. For three of six models, the 95% CI for Δ QWK (Log + Dialogue vs. Log Only) lay entirely above zero, providing statistical evidence of improvement. The largest gains were observed for Gemini-2.5-Pro (Δ = 0.35) and Gemini-2.5-Flash (Δ = 0.25). Even in models where the CI slightly overlapped zero (e.g., the GPT-5 family),

point estimates remained consistently positive. In contrast, adding expert-annotated precodes (Log + Dialogue + Precode vs. Log + Dialogue) yielded no consistent benefits: across all six models, 95% CIs broadly overlapped zero, with both positive and negative point estimates (e.g., Gemini-2.5-Pro $\Delta = -0.01$; GPT-5-Mini $\Delta = 0.07$). These results indicate that dialogue data substantially enhance LLM assessment accuracy, whereas human pre-coding adds no statistically significant advantage. Results can be found on OSF. Accordingly, subsequent dimension-level analyses focus on the Log + Dialogue scenario.

In addition to assessing accuracy, we also observed overall systematic rating biases and error analysis that varied with input scenarios. Details are available on OSF, but briefly, in the Log Only condition, all models consistently underestimated collaboration quality, with average rating differences (AI minus human) ranging from −0.27 (GPT-5) to −1.17 (Gemini-2.5-Flash-Lite). Adding dialogue reduced this bias substantially; Gemini-2.5-Pro's bias decreased from −0.42 to −0.16, and GPT-5 decreased from −0.27 to −0.13. The Log + Dialogue + Precode condition yielded near-zero bias for Gemini-2.5-Pro (0.00) and Gemini-2.5-Flash (−0.004), though precodes did not consistently eliminate bias across all models. Moreover, by dimension, error analysis in the Log + Dialogue condition revealed asymmetric misclassifications: severe underestimations (true=3, AI predicted=1) outnumbered overestimations (true=1, AI predicted=3) by 41 to 24 instances. Dimension-specific patterns showed underestimations concentrated in Conflict Resolution (29%), while overestimations affected On-task Focus (25%). Notably, 96% (53/55) of moderate overestimations (true=1, AI predicted=2) occurred in one dyad alone, suggesting specific interaction patterns, rather than model limitations, drive misclassification.

To address RQ2, we narrowed down to examine dimension-specific assessment capabilities. We analyzed model performance across the six collaborative constructs within the optimal Log + Dialogue scenario. This analysis reveals both individual model patterns and cross-family tendencies in assessing different aspects of collaboration (Figure 4). Model-specific analysis indicated considerable variability across all dimensions. For example, focusing on QWK (Figure 4, Right), Gemini-2.5-Pro achieved strong agreement on Idea Elaboration (0.61) and Responsiveness and Listening (0.58), while the comparably scaled GPT-5 reached only 0.45 and 0.44 on its strongest dimensions (Idea Elaboration and Questioning). Mid-tier models exhibited a similar cross-family pattern: Gemini-2.5-Flash peaked at 0.54 and 0.49 on these dimensions, whereas GPT-5-Mini reached 0.42 for both. Notably, GPT-5-Mini showed particularly poor performance on Responsiveness and Listening (−0.11), suggesting model-specific limitations. Despite these fluctuations, both model families showed broadly similar directional trends across the six dimensions. On average across all dimensions, the Gemini and GPT families exhibited a Spearman rank correlation of $\rho = 0.83$ ($p = 0.042$) when comparing dimension-level agreement, suggesting that the relative strengths and weaknesses across collaboration dimensions were largely consistent between the two families. In both cases, higher alignment appeared on more explicit dialogue dimensions, such as Questioning and Idea Elaboration, while lower or near-zero alignment was found for abstract dimensions like Conflict Resolution (−0.20 to 0.07) and On-task Focus (−0.08 to 0.09).

These results indicate general convergence rather than sharp divergence between LLM families. While individual models differ in magnitude, both families appear to share similar response patterns under the same collaborative context, implying that concrete, explicit communicative behaviors may currently generalize across large-language-model families, whereas implicit group-dynamic reasoning remains more challenging.

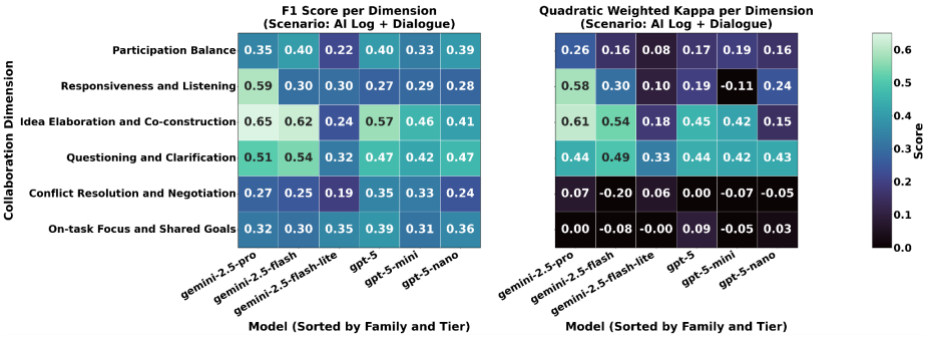


Fig. 4. Per-dimension agreement (*left*: F1 score; *right*: QWK) for six LLMs in the log + dialogue scenario

4 Discussion

This study examined how LLMs can be used to represent and interpret collaboration quality across various combinations of input data, and how different LLM families converge or diverge in their overall and dimension-specific performance. Our findings reveal a clear hierarchy in data requirements for LLM-based assessment. Behavioral traces alone produced poor agreement with the human coder (QWK=0.04–0.24). This result shows that while log data records what students did, they may not be sufficient to capture how students interacted, why they made such choices, or how ideas were shared and negotiated, which underpin social and regulatory dimensions [7]. From a QE standpoint, this finding indicates that representations built only from behavioral traces tend to separate actions from their interactional context, which limits the ability to model meaning-making processes. It also has implications for broader QE work, as models derived solely from behavioral log data (e.g., [18, 27, 40]) may not fully capture the interactional context in which actions take place or how those actions relate to broader processes, and they should therefore be interpreted with caution.

Adding dialogue transcripts resulted in higher agreement, with QWK values ranging from 0.22 to 0.49. This aligns with prior research showing that conversational data provide important insight into students' reasoning, negotiation, and shared meaning-making processes [31]. It also reinforces the central argument that the choice of representation shapes what becomes available for interpretation. In contrast, adding human-coded dialogue acts did not provide a clear additional benefit, with estimated changes in QWK overlapping zero across models. This finding should not be interpreted as evidence that human coding is unnecessary. Rather, it suggests that the

particular dialogue-act annotations used in this study may not have added information that improved LLM agreement with expert-coded collaboration ratings beyond what was already available in the raw transcripts. One possible explanation is that LLMs can extract some relevant semantic and interactional features directly from transcripts. Another possibility is that the human annotations made selected discourse features explicit while reducing access to the broader context needed to evaluate higher-level collaboration quality.

While no major differences were observed between the two model families, an analysis of computational efficiency revealed distinct processing patterns. Specifically, Gemini models consumed 26% more input tokens (1.49M vs. 1.18M) yet produced 74% fewer average output tokens (62K vs. 244K) compared to ChatGPT variants. Yet, these divergent tokenization strategies showed no relationship to assessment accuracy (Spearman ρ=0.83 for dimension rankings), indicating that computational efficiency and performance are independent optimization targets. Furthermore, cost analysis revealed substantial efficiency differences across models. Gemini-2.5-Flash achieved QWK=0.39 at 24% of Gemini-2.5-Pro's cost (QWK=0.49) and 15% of GPT-5's cost (QWK=0.37). For large-scale classroom analytics involving hundreds of session evaluations, this cost difference with a modest performance trade-off between Gemini-2.5-Flash and Gemini-2.5-Pro suggests that lower-cost models may be practical for some scalable applications. However, these findings should be interpreted in relation to the larger validity question. Efficiency gains are meaningful for QE only when the underlying representation still supports defensible claims about relationships, meanings, and patterns of practice.

Dimension-level analysis exposed systematic performance patterns. Models achieved moderate agreement on explicit communicative dimensions like Idea Elaboration (QWK ranging from 0.15 to 0.61) and Questioning (QWK ranging from 0.33 to 0.49). Performance decreased for dimensions requiring implicit inference, Conflict Resolution (QWK ranging from -0.20 to 0.07), and On-task Focus (QWK ranging from -0.08 to 0.09). Moreover, error analysis revealed asymmetric failure modes: Conflict Resolution accounted for 29% of severe underestimations (failing to detect high-quality negotiation), while On-task Focus explained 25% of severe overestimations (misinterpreting off-task behavior as engaged collaboration). This asymmetry suggests current LLMs can detect explicit communicative acts but struggle with temporal dynamics and implicit group processes that require inferring meaning beyond literal dialogue, a limitation consistent with broader evidence of shallow social reasoning in large language models [44]. This observation is consistent with earlier work showing that constructs grounded in explicit discourse are easier to operationalize from available traces [9]. On the other hand, constructs that depend on inference across turns, shared regulation, or evolving group context might be more difficult to represent consistently using the same data.

There are some limitations to the current study that may be addressed through future research. The small sample of 11 dyads solving mathematics problems in a controlled setting may limit the generalizability of the findings to larger groups and diverse domains. Given the rapid evolution of LLMs, continuous benchmarking will be essential to maintain validity over time, including the incorporation of additional

process data to examine the extent to which LLMs can explain variations in collaboration quality. Future work should explore whether integrating multimodal data (e.g., gaze, gesture, and physiological signals) from larger collaborative groups can further enhance the validity and scalability of LLM-based assessments of collaboration. In addition, considering current trends and the unique strengths and limitations of each model, future researchers could adopt a multi-agent framework in which different models specialize in specific tasks, such as analyzing dialogue, interpreting behavior, or applying rubrics, to work collaboratively toward a more comprehensive and reliable assessment. More broadly, future work should examine how alternative forms of segmentation, coding schemes, and multimodal representation affect what aspects of collaboration can be interpreted consistently and validly.

5 Conclusion

Overall, our findings suggest that LLM-based analysis combining log and dialogue data can be a useful tool for initial screening or analytic triage, particularly for identifying sessions that show stronger or weaker collaborative patterns through explicit communicative indicators. However, current models are not yet ready for fully automated evaluation of collaboration quality, especially for dimensions that require inference about regulation, conflict, or the evolving context of group interaction. For researchers interested in integrating LLMs into QE workflows, we recommend an AI-in-the-loop approach, where AI plays a supportive role and human judgment remains central to the evaluation process. More broadly, these findings highlight that differences in data representation shape what can be interpreted about collaboration. Maintaining alignment among data, coding, model input, and interpretation is therefore essential for producing valid accounts of collaborative meaning-making.

Acknowledgments. This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grants DRL-2229612, DRL-2048480, R305B230007, and #2331379, as well as the Gates Foundation (#078981), support from the Learning Engineering Tools Competition, and other anonymous philanthropy. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

1. An, S., Li, Y., Ma, Y., Cheng, G., Xu, G.: Developing an LLM-empowered agent to enhance student collaborative learning through group discussion. In: *International Conference on Computers in Education* (2024).
2. Arastoopour Irgens, G., Eagan, B.: The foundations and fundamentals of quantitative ethnography. In: *International Conference on Quantitative Ethnography*, 3–16 (2022).
3. Barron, B.: When smart groups fail. *Journal of the Learning Sciences* 12(3), 307–359 (2003).

4. Chew, R., Bollenbacher, J., Wenger, M., Speer, J., Kim, A.: LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv* arXiv:2306.14924 (2023).
5. Dillenbourg, P.: What do you mean by "collaborative learning"? (1999).
6. Earle-Randell, T. V., Wiggins, J. B., Ruiz, J. M., Celepkolu, M., Boyer, K. E., Lynch, C. F., ... Wiebe, E.: Confusion, conflict, consensus: Modeling dialogue processes during collaborative learning with hidden markov models. In: *International Conference on Artificial Intelligence in Education*, 615–626 (2023).
7. Esterhazy, R., Kaliisa, R., Sanchez, D., Langford, M., Damsa, C.: Advancing multimodal collaboration analytics: A scoping review. *J. of Learning Analytics* 12(2), 105–124 (2025).
8. Fleiss, J. L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3), 613–619 (1973).
9. Goldfarb Cohen, S., Yanai, J. V., Dishon, G.: Modeling group discourse with epistemic network analysis: Unpacking connections, perspectives, and individual contributions. *Journal of Science Education and Technology* 34(5), 1080–1092 (2025).
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. Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Hassan, S. Z., Shoman, M., Wu, J., Mirjalili, S., Shah, M.: Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Preprints* (2024).
12. Hasibuan, R. K., Chandra, D. A., Wahyudi, S., Setiawan, A.: Collaborative learning to improve communication skills in information and communication technology subjects. *Jurnal Mandiri IT* 11(4), 132–137 (2023).
13. Herrenkohl, L. R., Cornelius, L.: Investigating elementary students' scientific and historical argumentation. *Journal of the Learning Sciences* 22(3), 413–461 (2013).
14. Herro, D., Quigley, C., Andrews, J., Delacruz, G.: Co-measure: Developing an assessment for student collaboration in STEAM activities. *Int. J. of STEM Education* 4(1), 26 (2017).
15. Hurme, T., Järvelä, S.: Students' activity in computer-supported collaborative problem solving in mathematics. *International Journal of Computers for Mathematical Learning* 10(1), 49–73 (2005).
16. Järvelä, S., Järvenoja, H., Malmberg, J., Hadwin, A. F.: Exploring socially shared regulation in the context of collaboration. *Journal of Cognitive Education and Psychology* 12(3), 267–286 (2013).
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. Karumbaiah, S., Baker, R., Barany, A., Shute, V.: Using epistemic networks with automated codes to understand why players quit levels in a learning game. In: *International Conference on Quantitative Ethnography*, 106–116 (2019).
19. Khosa, D. K., Volet, S. E.: Productive group engagement in cognitive activity and metacognitive regulation during collaborative learning: Can it explain differences in students' conceptual understanding? *Metacognition and Learning* 9(3), 287–307 (2014).
20. Kimball, R., Caserta, J.: *The data warehouse ETL toolkit* (2004).
21. Klang, N., Karlsson, N., Kilborn, W., Eriksson, P., Karlberg, M.: Mathematical problem-solving through cooperative learning—The importance of peer acceptance and friendships. *Frontiers in Education* 6, 710296 (2021).

22. Law, Q., Chung, J., Leung, L., Wong, T.: Perceptions of collaborative learning in enhancing undergraduate education students' engagement in teaching and learning English. *US-China Education Review* 7(2), 89–100 (2017).
23. Lazakidou, G., & Retalis, S.: Using computer supported collaborative learning strategies for helping students acquire self-regulated problem-solving skills in mathematics. *Computers & Education* 54(1), 3–13 (2010).
24. Lee, S. M., Li, H., Zhang, S., Zhong, Z., Lee, J. E., & Botelho, A. F. (2025, July). So What? Unpacking the Complexities in Collaborative Problem Solving with AI-Augmented Sense-Making. In *International Conference on Artificial Intelligence in Education* (pp. 431-438). Cham: Springer Nature Switzerland.
25. Lee, Y., Kwon, C., Seoh, S., Gweon, G., Stamper, J., Rosé, C. P.: Capturing collaborative competency with GPT-4o and ENA. In: *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning (CSCL 2025)*, 72–80 (2025).
26. Li, H., Zhang, S., Lee, S., Lee, J. E., Zhong, Z., Weitnauer, E., & Botelho, A. F. (2024). Math in Motion: Analyzing Real-Time Student Collaboration in Computer-Supported Learning Environments. *International Educational Data Mining Society*.
27. Liu, X., Hussein, B., Barany, A., Baker, R. S., Chen, B.: Decoding player behavior: Analyzing reasons for player quitting using log data from puzzle game *Baba Is You*. In: *International Conference on Quantitative Ethnography*, 34–48 (2023).
28. Liu, X., Zhang, J., Barany, A., Pankiewicz, M., Baker, R. S.: Assessing the potential and limits of large language models in qualitative coding. In: Kim, Y. J., Swiecki, Z. (eds.) *Advances in Quantitative Ethnography*, 89–103 (2024).
29. Mercer, N.: *The guided construction of knowledge: Talk amongst teachers and learners*. Multilingual Matters (1995).
30. Mercer, N.: *Words and minds*. Routledge (2002).
31. Mercer, N., Howe, C.: Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction* 1(1), 12–21 (2012).
32. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* 16(5), 1–72 (2025).
33. Otten, D. M., Papadopoulos, P. M., Haeri, M. A., Endedijk, M. D.: Automatic analysis of transactivity in collaborative problem solving dialogues: The relationship of novelty and reference. In: *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning (CSCL 2025)*, 321–325 (2025).
34. Ruis, A. R., Cai, Z., Shaffer, D. W.: Subgroup fairness in multilingual text classification. In: *International Conference on Quantitative Ethnography*, 85–101 (2025).
35. Salminen-Saari, J. F. A., Garcia Moreno-Esteva, E., Haataja, E., Toivanen, M., Hannula, M. S., Laine, A.: Phases of collaborative mathematical problem solving and joint attention: A case study utilizing mobile gaze tracking. *ZDM – Mathematics Education* 53(4), 771–784 (2021).
36. Shaffer, D. W.: *Quantitative ethnography*. Lulu.com (2017).
37. Shaffer, D. W., Wang, Y., Ruis, A.: Transmodal analysis. *Journal of Learning Analytics* 12(1), 271–292 (2025).
38. Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. *Computers & Education* 143, 103672 (2020).
39. Weinberger, A., Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education* 46(1), 71–95 (2006).

40. Wu, M., Zhang, J., Barany, A.: Understanding detectors for SMART model cognitive operation in mathematical problem-solving process: An epistemic network analysis. In: *International Conference on Quantitative Ethnography*, 314–327 (2022).
41. Xu, E., Wang, W., Wang, Q.: The effectiveness of collaborative problem solving in promoting students' critical thinking: A meta-analysis based on empirical literature. *Humanities and Social Sciences Communications* 10(1), 16 (2023).
42. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–32 (2024).
43. Zhang, S., Earle-Randell, T. V., Shen, Q., Botelho, A. F., Israel, M., Boyer, K. E., ... & WIEBE, E. (2024, November). Predicting and Analyzing Students' Higher-Order Questions in Collaborative Problem-Solving. In *International Conference on Computers in Education*. APSCE.
44. Zhang, S., Earle-Randell, T. V., Ganapathy Prasad, P., Liu, Z., Shi, Y., Bhat, S., ... & Botelho, A. F. (2026, February). Examining Students' Code Comprehension with LLMs in Block-and Text-Based Programming. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V. 2* (pp. 1601-1602).
45. Zhang, S., Li, H., Lee, S., Lee, J.-E., Schroeder, N. L., & Botelho, A. F. (2026). Assessing student collaboration with multimodal data and large language models. In C. Rosé, G. Gweon, C. Asterhan, & A. Keune (Eds.), *Proceedings of the 19th International Conference on Computer-Supported Collaborative Learning (CSCL 2026)* (pp. 767–769). International Society of the Learning Sciences.