

Assessing Student Collaboration with Multimodal Data and Large Language Models

Shan Zhang, University of Florida, zhangshan@ufl.edu

Hongming Li, University of Florida, hli3@ufl.edu

Seiyon Lee, University of Florida, leeseiyon@ufl.edu

Ji-Eun Lee, Singapore University of Technology and Design, jieun_lee@sutd.edu.sg

Noah L. Schroeder, University of Florida, schroedern@ufl.edu

Anthony F. Botelho, University of Florida, abotelho@coe.ufl.edu

Abstract: We introduce a multimodal evaluation approach that leverages large language models to assess group collaboration quality in computer-supported collaborative learning environments. Using synchronized log data, dialogue transcripts, and human-coded annotations from undergraduate dyads working in the Graspable Math platform, we develop a theoretically grounded rubric capturing cognitive, social, and regulatory dimensions of collaboration, and examine the extent to which LLM-generated evaluations align with expert human judgments.

Introduction

Collaborative learning (CL) supports deeper understanding through shared reasoning and knowledge co-construction (Dillenbourg, 1999). In computer-supported collaborative learning (CSCL), researchers increasingly rely on multimodal data (e.g., logs, dialogue) to capture these complex processes (Sharma & Giannakos, 2020). However, evaluating collaboration quality remains challenging, as rubric-based approaches require extensive human coding and are difficult to scale (Lee et al., 2025). Recent advances in large language models (LLMs) offer a promising alternative for automating the analysis of unstructured educational data, including dialogue and log traces (Liu et al., 2024). While prior work has shown that LLMs can approximate human coding for well-defined constructs, their ability to evaluate group-level collaboration quality using multimodal inputs remains underexplored. To address this gap, 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. Specifically: 1) RQ1: What aspects of collaboration quality should be considered when developing a group-level collaboration rubric grounded in the theoretical framework? 2) RQ2: How does the LLM perform in grading group collaboration quality using combined input data (logs, dialogue, and precoded annotations)?

Methods

This study involved 22 undergraduate students (11 dyads) who collaboratively solved algebra problems using the Graspable Math platform (see Figure 1). We recorded both log and dialogue data, providing synchronized multimodal records of student interactions. To address RQ1, we developed a theoretically grounded collaboration rubric based on prior frameworks, Mercer's Exploratory Talk Framework (1995), Weinberger and Fischer's (2006) Argumentative Knowledge Construction model, and Järvelä and Hadwin's (2013) Socially Shared Regulation of Learning perspective, capturing cognitive, social, and regulatory dimensions of collaboration. Dialogue transcripts were manually coded using an established scheme (Zhang et al., 2024) with strong inter-rater reliability ($\kappa > 0.75$), and log data were processed into structured, time-aligned action sequences. Details about dialogue, and log coding schemes are available on the Open Science Framework: https://osf.io/mnxts/overview?view_only=6c0769b0bfc44998b5a29b6eb184f42. To address RQ2, we constructed multimodal inputs combining logs, dialogue, and human-coded annotations, and evaluated LLM performance (e.g., GPT-5) in grading collaboration quality against expert ratings. Model outputs were assessed using macro F1 and Quadratic Weighted Kappa (QWK).

Figure 1

(a) GM's Collaborative Whiteboard Interface (a) Algebraic Problem Solving Process(b)



(a)

$$\frac{1}{3} + \frac{1}{2}$$

$$\frac{2}{2 \cdot 3} + \frac{1}{2}$$

$$\frac{2}{2 \cdot 3} + \frac{3}{2 \cdot 3}$$

(b)

Results

The rubric in Table 1 comprised six dimensions capturing key aspects of effective collaboration: Participation Balance, Responsiveness & Listening, Idea Elaboration & Co-construction, Questioning & Clarification, Conflict Resolution & Negotiation, and On-task Focus & Shared Goals. These six dimensions were organized into cognitive, social, and regulatory aspects, reflecting the multifaceted nature of collaborative learning. Each dimension was rated on a three-point scale, with an “NA” option when not applicable. Two coders jointly reviewed 20% of transcripts and video recordings to calibrate their understanding of the rubric. Inter-rater reliability was established at 0.84 QWK, after which one coder completed the remaining ratings. After scoring all 11 groups across their attempted problems ($n = 49$), we computed a total collaboration score per problem by summing ratings across all six dimensions and dividing by the number of applicable dimensions (excluding any marked “NA”). For RQ2, using these expert ratings as ground truth, we evaluated LLM performance under the input data of log, dialogue, and precoded. GPT-5 achieved an F1 score of 0.46 and a QWK of 0.31. This suggests that GPT-5 demonstrates some ability to clarify collaboration quality, but its alignment with expert judgments is still relatively modest. Our findings contribute to the ongoing discourse of using AI for qualitative analyses and provide evidence on the extent to which LLMs can approximate expert evaluations of collaborative processes.

Table 1

Collaboration Quality Rubric

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 here due to space limitations

References

- Dillenbourg, P. (1999). *What do you mean by “collaborative learning”?*
- Mercer, N. (1995). *The Guided Construction of Knowledge: Talk Amongst Teachers and Learners*. Multilingual Matters.
- Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51(5), 1450–1484. <https://doi.org/10.1111/bjet.12993>
- Järvelä, S., Järvenoja, H., Malmberg, J., & Hadwin, A. F. (2013). Exploring Socially Shared Regulation in the Context of Collaboration. *Journal of Cognitive Education and Psychology*, 12(3), 267–286.
- Lee, Y., Kwon, C., Seoh, S., Gweon, G., Stamper, J., & Rosé, C. P. (2025). Capturing Collaborative Competency with GPT-4o and ENA. In *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning-CSCL 2025*, pp. 72-80. International Society of the Learning Sciences.
- Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R. S. (2024). Assessing the Potential and Limits of Large Language Models in Qualitative Coding. In Y. J. Kim & Z. Swiecki (Eds.), *Advances in Quantitative Ethnography* (pp. 89–103). Springer Nature Switzerland.
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1), 71–95.
- 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.

Acknowledgements

We would like to thank the National Science Foundation (#2331379), Institute of Education Sciences (#R305B230007), Gates Foundation (#078981), Support from the Learning Engineering Tools Competition, and other anonymous philanthropy.