Combining Log Data and Collaborative Dialogue Features to Predict Project Quality in Middle School AI Education

Conrad Borchers[1,†], Xiaoyi Tian[2,†], Kristy Elizabeth Boyer[3] and Maya Israel[3]

[1]*Carnegie Mellon University*
[2]*North Carolina State University*
[3]*University of Florida*

Abstract

Project-based learning plays a crucial role in computing education. However, its open-ended nature makes tracking project development and assessing success challenging. We investigate how dialogue and system interaction logs predict project quality during collaborative, project-based AI learning of 94 middle school students working in pairs. We used linguistic features from dialogue transcripts and behavioral features from system logs to predict three project quality outcomes: *productivity* (number of training phrases), *content richness* (word density), and *lexical variation* (word diversity) of chatbot training phrases. We compared the predictive accuracy of each modality and a fusion of the modalities. Results indicate log data better predicts productivity, while dialogue data is more effective for content richness. Both modalities modestly predict lexical variation. Multimodal fusion improved predictions for productivity and lexical variation of training phrases but not content richness. These findings suggest that the value of multimodal fusion depends on the specific learning outcome. The study contributes to multimodal learning analytics by demonstrating the nuanced interplay between behavioral and linguistic data in assessing student learning progress in open-ended AI learning environments.

Keywords

project-based learning, multimodal learning analytics, project quality prediction, dialogue, K-12

1. Introduction

Project-based learning (PBL) is crucial in STEM, especially in computing, where students collaboratively create artifacts [1, 2]. PBL fosters computational thinking, problem-solving [3, 4], and engagement [3, 5]. Effective assessment of student learning and project quality is essential for educators to provide targeted feedback [6]. However, because of the iterative, open-ended nature of project development (e.g., app creation), tracking progress in PBL is challenging. Analyzing behavioral data during PBL offers insights into student learning, collaboration, and refinement strategies. Prior work has explored predicting project outcomes such as final grades [7, 8, 9], task performance [10, 11], group satisfaction [12, 13], and engagement [14]. In computer science education specifically, there is a growing interest in understanding more granular aspects of the project, such as the completeness of the student code traces [15, 16], or the correctness of the steps within multi-step problem-solving episodes [17]. These finer-grained project quality outcomes capture learning progress during PBL, enabling educators and intelligent systems to identify struggling learners and provide timely support. Predicting project quality **proxies** *during* learning processes (in addition to conventional, post-completion learning outcomes) can serve two future research purposes: First, it may inform adaptive modules that provide additional instruction based on detected areas of improvement. Second, it can inform the study of effective student collaboration by providing insights through models capturing what contributes to desirable learning process outcomes.

CSEDM'25: 9th Educational Data Mining in Computer Science Education (CSEDM) Workshop, July 20, 2025, Palermo, Sicily, Italy
[†]Conrad Borchers and Xiaoyi Tian contributed equally to this work.
cborcher@cs.cmu.edu (C. Borchers); xtian9@ncsu.edu (X. Tian); keboyer@ufl.edu (K. E. Boyer); misrael@coe.ufl.edu (M. Israel)
0000-0003-3437-8979 (C. Borchers); 0000-0002-5045-0136 (X. Tian); 0000-0003-3434-3450 (K. E. Boyer); 0000-0003-0302-6559 (M. Israel)

Researchers have explored various data sources and analytical approaches to better understand and support student learning during PBL. A longstanding tradition in learning analytics involves inferring cognitive and metacognitive processes from log and language data [18]. While researchers have leveraged log interactions to model learning trajectories [17, 19], and language data to infer self-regulated learning strategies [20], these modalities are rarely used for joint prediction. Integrating dialogue-based and log-based features may offer a more holistic view of student collaborative learning analytics [21, 22]. While prior research has focused on predicting learning gains using multimodal collaboration features [23], less attention has been given to the temporal dynamics of learning [18]. We address this gap by predicting the quality of individual learning events in collaborative learning, that is, the quality of training phrases while students program a chatbot. Specifically, we compare the predictive capabilities of dialogue and log data in assessing the quality of student projects in an AI learning context where students collaboratively develop chatbots. We examine whether combining different modalities of data enhances the predictive accuracy of project quality compared to using them independently:

RQ1: How well can student project quality be predicted from single modalities (dialogue, log data)?

RQ2: To what extent does the multimodal fusion of these data sources enhance predictive accuracy?

2. Related Work

2.1. Multimodal Fusion and Prediction

Multimodal approaches to mining data and modeling are increasingly common in educational data mining (EDM) [24, 25]. A recent review by Chango et al. [26] traces emerging trends in the use of multimodal in typical EDM contexts, including online learning environments, classroom studies, and blended learning contexts. Data collected in these sites can range from text data to log data, physiological learner data such as eye-tracking, and spatiotemporal data such as teacher positions [25].

Natural language data, especially, is increasingly used for foundational EDM prediction tasks. For example, Zhang et al. [20] utilized large language model (LLM) embeddings to predict self-regulated learning stages from think-aloud protocols, illustrating the potential of text-based data in understanding cognitive processes. Scarlatos et al. [27] explored using LLM for knowledge tracing from learners' dialogue. Borchers et al. [19] fused data from peer tutoring chat collaborations with tutoring system problem-solving logs to model learning rates in relation to different dialogue acts. However, these approaches have rarely combined language data with log data to jointly model learning.

The present study builds on this body of work by applying text mining techniques based on automated transcriptions and embedding [20] with log data for a novel prediction task: predicting learner project quality in AI education. Specifically, we incorporate transcriptions of students' collaboration as a data source in EDM, combining them with established log data to enhance predictive accuracy. By testing the combined predictive power of log data and transcription-derived features, we test whether multimodal fusion improves the prediction of learning processes during collaboration. We contribute to the ongoing discourse on the utility of multimodal data in enhancing educational predictions and insights.

2.2. Mining Collaborative Learning Systems

Collaborative learning systems offer opportunities to mine multimodal data from learning-system and learner-learner interactions [28, 29].[1] These systems provide insights into learning patterns through modeling tutoring rates and collaborative behaviors [19], which have been used for feedback tools [31, 22]. However, selecting meaningful outcomes and learning constructs for multimodal analysis remains challenging. **The present study embraces this challenge in a novel predictive context,** where students learn to program chatbots using natural language inputs. We predict the quality of

[1]We define collaboration as the coordinated effort to jointly engage in problem-solving and knowledge construction [30].

student-created training phrases during learning. Insights from this research have implications for other collaborative systems and performance prediction from collaborative characteristics [19, 23, 32, 33].

In addition to peer collaboration interactions, research has demonstrated how dialogue can be leveraged to support learner-system interactions. For instance, intelligent agents can adapt instruction based on detected learning patterns [34]. In contrast, teachable agents provide practice problems and assessments [35] and build rapport with learners through lexical adaptation toward learners [36]. Similarly, adaptive dialogue systems such as EER-tutor have demonstrated improved learning gains by considering prior student errors [37]. Finally, Gaze Tutor [38] leveraged gaze data to detect disengagement and adapt dialogue to re-engage students. The present study builds on this foundation by seeking to develop meaningful predictors of student learning performance during collaboration, making progress toward detectors that could improve adaptivity in future collaborative learning systems.

3. Study Setup and Dataset

3.1. Study Context and Data Collection

The study was conducted in a public middle school in the southeastern United States in Spring 2024. This IRB-approved study included 128 students across six class periods in a science class. Out of these, 100 consented to participate in the research, and 97 reported their demographic information: 49 identified as girls, 46 as boys, one as non-binary, and one preferred not to disclose. Further, 38 students identified as Asian, 34 as White, 20 as Black/African American, six as Hispanic/Latinx, three as Native American, five as self-described, and three preferred not to disclose. The average age was 11.7 years (SD = 0.48).

The primary objective of this study was to engage students in core concepts of artificial intelligence (AI) and computer science (CS) through the hands-on development of conversational agents [39]. The classroom study contained a 10-hour learning module covering AI fundamentals, hands-on activities, and chatbot development. The instructional content was aligned with AI Five Big Ideas in K12 [40]. Specific learning goals included (1) understanding the role of data in AI systems and determining appropriate datasets for specific AI applications; (2) describing how datasets create representations of the world for reasoning tasks; (3) training and evaluating classification and prediction models while examining their accuracy on new inputs; and (4) creating chatbots and developing natural interactions. Each learning module was designed to build students' AI literacy progressively.

During collaborative chatbot development, students were randomly paired to create chatbots on science topics of their choice (e.g., the water cycle, climate, and living organisms). Each pair collaborated over three class sessions (40 minutes each) to develop their chatbot. We captured audio/video recordings of these collaborative sessions using recording software from their laptop. Humans transcribed the recordings through the Rev service. The resulting dialogue dataset contains 121 collaboration sessions from 47 student pairs (94 individuals). Each session contains an average of 278 utterances (SD = 108.7).

3.2. Chatbot Development Environment

The learning environment, AMBY (Figure 1), is a graphical interface designed for middle school students to create conversational agents and learn about AI [41]. The environment enables students to test example agents, edit agents, or create new agents. For instance, a student group might create an agent on the topic of "climate change," where they define *intents* (e.g., "impact on oceans"), input *training phrases* (e.g., "Does climate change impact the oceans?"), and write corresponding *responses* (e.g., "Climate change impacts oceans through sea-level rise and ocean acidification."). Students can test their agents by having conversations and adjusting the chatbot's voice to personalize the interaction.

Students construct and refine intents (i.e., basic chatbot development units representing the goal or purpose of a user message). In the development panel (Figure 1, left), students can add, edit, or delete training phrases and responses. They submit training requests by clicking "train the AI," which updates the chatbot's intent classification model based on the provided training data. Students worked with

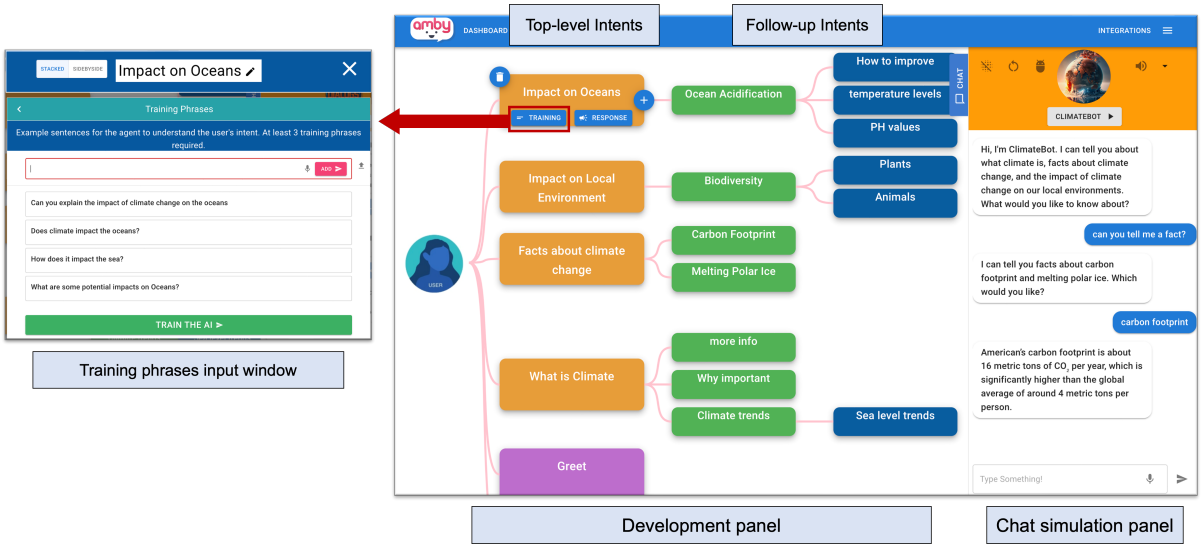

Figure 1: Chatbot development environment AMBY

a partner at the same computer on iterating on the chatbot design, following the pair programming paradigm [42], prompting spontaneous collaboration, dialogue, and joint decision-making.

The system collects 23 types of timestamped **user interaction logs** (e.g., adding or deleting training phrases, submitting AI training requests, chat messages). Across 121 recorded collaborative sessions, students submitted an average of seven intent training requests per session, and we used the content of training phrases they submitted each time as measures of project quality (Section 3.4).

3.3. Dataset Preparation

To model relationships between system logs, student collaborative dialogue, and project outcomes, we prepared a multimodal dataset (Figure 2). Following Borchers et al. [18], we synchronized dialogue transcripts with log data by aligning timestamps between the datasets, yielding an accuracy of 1 second. For each intent submission, we extracted its submitted training phrases for outcome measurement and defined ***intent-working segment***, which is the time window from when students began working on the intent until submission (based on logs). Based on an ad-hoc decision, we manually examined the segments for accuracy and removed segments that were excessively long (> 12.5 minutes, 90 percentile of the dataset) based on considerations of context length during embedding. This preprocessing step is not expected to introduce bias, as our three outcome measures were virtually uncorrelated with dialog length as measured in the total number of words ($|r| < 0.05$) and total number of turns ($|r| < 0.07$). The average segment duration is 193 seconds. We then extracted the synchronized student dialogues and system logs within these windows for downstream analysis (Section 4).

3.4. Outcome Measures

Our unit of analysis is the *intent*, a core chatbot element that handles the recognition of user queries. Students averaged seven intent training submissions per 30-minute collaborative session and 17 total submissions. We measured intent quality through three key outcomes chosen based on the learning objectives, learning curve analyses, and correlation with final project scores.

Training Phrase Count measures the number of phrases input by students for training the chatbot to recognize an intent. This metric represents productivity and engagement in the iterative training process: more phrases typically indicate greater effort in improving the chatbot's performance. The relationship between chat-based activity and engagement has also been observed in past research [19]. ***Lexical Density*** captures the proportion of content words (nouns, adjectives, verbs, and adverbs) to total words in the provided phrases, which assesses linguistic richness since effective training requires

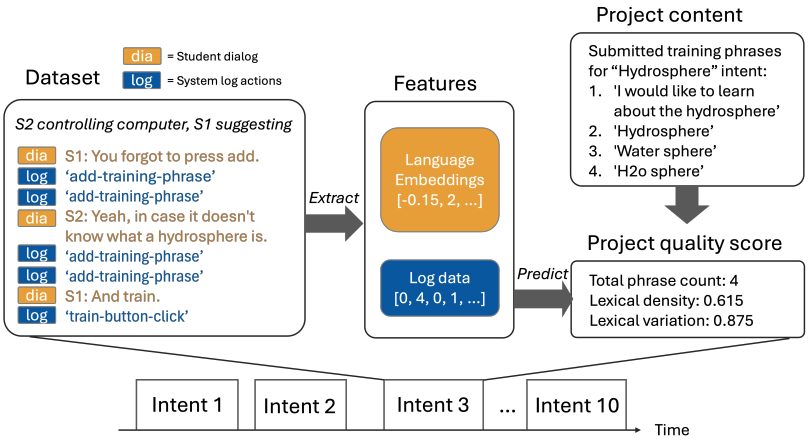

Figure 2: Overview of our multimodal predictive model, an example dataset and project outcome

substantive content over filler words [43, 44]. ***Lexical Variation*** measures language diversity through the ratio of unique content words to total content words [43], important because varied phrases help chatbots recognize different user expressions. Higher variation indicates more robust training input.

These project quality metrics align with key AI learning objectives [40, 39]. For instance, students learn that simply adding similar word permutations (i.e., high training phrase quantity with little information richness, such as `['hat', 'hats', 'HATS']`) is ineffective for chatbot programming. As students iteratively generate, refine, and diversify training phrases, they engage in active problem-solving and knowledge construction [45, 46]. Lexical diversity is also an important predictor of academic performance in tasks requiring writing and speaking [47, 48].

Validation. We validated our outcomes through learning curve analyses [49] and correlations with three final project grades (i.e., expert-rated score aggregated across 11 dimensions using a rubric, expert-rated training phrase score, and end-user satisfaction evaluated by three independent annotators). **Learning curves** can demonstrate whether students improve on our outcome measurements as they receive indirect feedback from AMBY on their chatbot testing performance. Figure 3 shows improvements across all metrics over time as students learned to add more phrases and improve lexical choices in the training data for better intent recognition (with the clearest improvement in lexical variation). For **correlations**, while none of our intent-level measures significantly *correlated* with overall expert ratings–likely due to the broad scope of the overall assessment, *training phrase count* significantly correlated with expert-rated training phrase scores ($r(45) = 0.5$, $p < .001$), and *lexical density* correlated with end-user satisfaction score ($r(45) = 0.36$, $p = .013$), indicating richer linguistic content improves end-user experience. Our measures showed appropriate separation: moderate correlation between lexical density and variation ($r = 0.28$) but minimal correlation with training phrase count ($r = 0.01$ and $r = -0.03$, respectively).

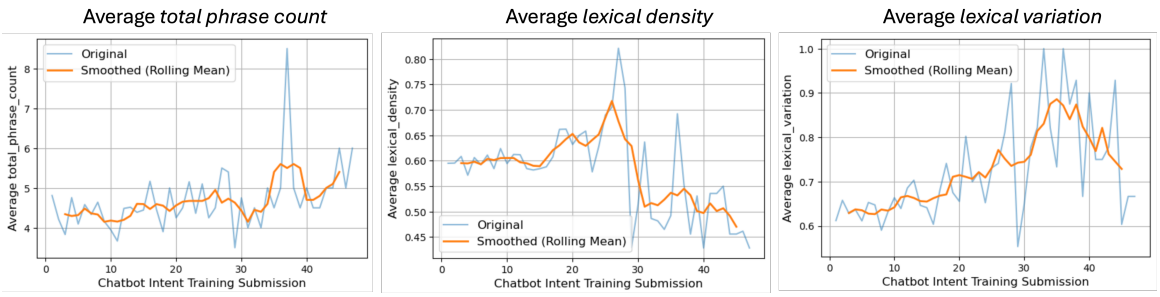

Figure 3: Learning curves for training phrase count, lexical density, and lexical variation, showing student improvements over time through smoothed running averages. Trends were less consistent for students with high submission counts (>30 intents vs. average 17), likely representing outliers.

4. Analysis Methods

4.1. Preprocessing and Feature Engineering

Log Data Features To capture students' system interactions during chatbot development, we engineered log features reflecting engagement, timing, and strategy as inputs for our machine learning models. We computed timing-related features (six features) capturing temporal dynamics: mean, standard deviation, minimum, maximum, median, and IQR of time between actions. Event-based features (2124 features) characterized interaction patterns through event type counts and bigram/trigram sequences of the 31 unique event types (e.g., `[test-chatbot, create-new-intent, add-phrase]`). This approach of using log patterns to predict process outcomes aligns with common EDM methods where log-based learning rates relate to instructional events [50, 51].

Language Features in Dialogue Transcripts To capture student dialogue differences, we computed numerical representations of student dialogue transcripts within a certain segment using state-of-the-art embedding models. Our early experiments showed Sentence-BERT (SBERT) embeddings [52] from `bert-base-uncased` performed comparably to OpenAI's `text-embedding-3-large` model [53]. Given computational efficiency and open-source considerations, we used SBERT's 768-dimensional embeddings as input features for predictive modeling, following established EDM methods [20].

4.2. Model Architecture, Training, and Evaluation

We employed a feedforward neural network for regression tasks [20], aiming to predict training phrase quality based on log data features and embeddings. The model architecture consisted of 2-4 hidden layers with ReLU activations and dropout regularization. Hyperparameters were optimized using grid search with five-fold student-level cross-validation over the following parameter space: Hidden layer configurations: *[256,128], [512,256,128], [1024, 512, 256,128]* and Dropout rates: *0.0, 0.025, 0.05, 0.1, 0.3, 0.5*. Model training was conducted using the Adam optimizer, and performance was evaluated using Mean Absolute Error (MAE) and AUC. Early stopping was implemented with a patience threshold of two epochs to prevent overfitting. The code is in the study's digital appendix.[2]

After model training terminated, we evaluated models on a held-out test set (33% of the data), reporting predictive performance across three feature configurations: **log only**, **dialogue only**, and **combined**. Using bootstrapped resampling, we computed 95% AUC confidence intervals (CIs). Based on average cross-validation AUC across folds, the best model was evaluated on the held-out test set.

5. Results

5.1. RQ1: Can student project quality be predicted from dialogue or log data?

We evaluated models trained on *dialogue data* and on *log data* to predict training phrase quality. Table 1 summarizes the holdout AUC. *Training phrase count* was better predicted by log-based features (AUC = 0.8053) than dialogue-based features (AUC = 0.5971), suggesting system interaction logs (e.g., frequency and timing of edits) are stronger indicators of phrase production. Conversely, *lexical density* was better captured by the dialogue-only model (AUC = 0.6551 vs. 0.5112), indicating student talk features capture the richness of their written training phrases. For *lexical variation*, both models performed modestly (log: AUC = 0.6016, dialogue: AUC = 0.5260), with log features showing a slight advantage.

5.2. RQ2: Does multimodal fusion improve predictive accuracy?

To answer RQ2, we performed early fusion, concatenating log, and dialogue features before passing them to the model [24]. The results are in Table 1 ("combined" rows). Comparing the best single-modality

[2] https://github.com/conradborchers/collaboration-edm25

results, for *training phrase count*, fusion improved AUC from 0.8053 (log-only) to 0.8301.[3] However, for *lexical density* of training phrases, the combined model (AUC = 0.5700) did *not* outperform the dialogue-only model (AUC = 0.6551), suggesting that log-based features provide limited insight into the content richness of learner input. For *lexical variation* of training phrases, fusion yielded a slight AUC increase (0.6089 vs. 0.6016 log-only), but confidence intervals suggest the gain is likely not significant. Overall, multimodal fusion yields the most substantial gain for predicting *training phrase count*, slightly improves the prediction of *lexical variation* of training phrases, but does not help predict *lexical density* of training phrases. Hence, the utility of combining modalities is outcome-dependent.

Table 1
Holdout AUC and 95% Confidence Intervals for Three Outcomes Across Training Modalities

Modality	Training phrase count	Lexical Density	Lexical Variation
Log Only	0.8053 [0.7470, 0.8604]	0.5112 [0.4556, 0.5655]	0.6016 [0.5418, 0.6615]
Dialogue Only	0.5971 [0.5250, 0.6671]	0.6551 [0.5920, 0.7168]	0.5260 [0.4579, 0.5933]
Combined	0.8301 [0.7732, 0.8822]	0.5700 [0.5042, 0.6352]	0.6089 [0.5438, 0.6727]

6. Discussion

6.1. RQ1: Unimodal Performance

Our findings reveal the differential effectiveness of dialogue and log data in predicting project quality on the process level. The superior performance of log-derived features in predicting *training phrase count* aligns with past work on using clickstream data to model learning trajectories [25], indicating behavioral engagement metrics provide critical insights into productivity and task completion. The stronger performance of dialogue features in predicting *lexical density* indicates the spoken discourse appears to translate into more content-rich chatbot training phrases. This builds on research demonstrating the value of transcribed speech in detecting self-regulated learning strategies [20] and suggests that students' verbal articulation during collaboration may influence the quality of their computational artifacts [54].

Overall, log-data-based features were more predictive of engagement, while dialogue-based features were more indicative of the linguistic characteristics of chatbot training phrases. Future research should explore how these models learn associations between collaborative actions and outcomes, potentially incorporating interpretability methods to examine feature importance more precisely [55].

6.2. RQ2: Multimodal Performance

Multimodal fusion demonstrated mixed results in enhancing predictive accuracy. The improved prediction of *training phrase count* through combined features suggests that integrating behavioral and linguistic cues provides a more comprehensive understanding of student engagement. However, for the *lexical density* of training phrases, the combined model underperformed compared to the dialogue model, suggesting that adding log features may interfere with or dilute the signal from dialogue features for this task. The cross-modal interaction has also been noted in past work [56], emphasizing that fusion requires careful consideration of interference. The modest gains in predicting *lexical variation* align with past work [25, 31]: the effectiveness of multimodal fusion depends on the prediction task.

The results reinforce the novelty of our study's focus on open-ended project-based learning in K-12 AI education, highlighting that predicting linguistic quality in this context presents unique challenges and opportunities compared to traditional educational data mining tasks. It is evident from our results

[3]As one attentive reviewer pointed out, counting 'add-training-phrase' logs might straightforwardly reveal the number of training phrases. We performed an ablation study, recomputing these accuracies without features dependent on 'add-training-phrase', 'delete-phrase', and 'add-response'. For completeness outcome, AUCs changed to 0.6865 (log only) and 0.6759 in the combined modalities, respectively.

that straightforward associations between collaborative dialog and process outcomes during learning are challenging to isolate in our context (as one might conceive that predicted log inputs may be clearly visible and explicitly mentioned in dialog, and hence straightforward to predict; from our experience, these cases rarely occur). This contributes to the ongoing discourse on the role of multimodal data for learning process prediction [26, 19, 20].

6.3. Implications

The findings highlight the importance of selecting appropriate data modalities for specific outcomes, contributing to multimodal learning analytics [26, 24]. In K-12, AI learning involving iterative chatbot design, log data, and dialogue provides complementary insights into productivity and AI understanding.

For educators assessing collaborative learning, log features offer insights into engagement and productivity [25], while dialogue features reveal conceptual understanding and self-regulation [19, 20]. However, our mixed multimodal fusion results suggest that careful feature selection and tuning are essential to maximize predictive accuracy. These insights inform future research on collaborative AI literacy tools. With sufficient accuracy, our models could monitor engagement and identify when students need scaffolding [34, 35]. Dialogue analysis could help educators understand students' AI concept mastery, enabling targeted feedback and aligning with growing interest in natural language-informed interventions [20, 27, 22]. Future work should evaluate automated transcription feasibility in real classrooms, balancing effectiveness with privacy considerations and teacher preferences [57].

6.4. Limitations and Future Work

First, our data collection was limited to a single middle school science class with a particular collaborative task, limiting generalizability. Second, our predictive models showed only modest performance in predicting *lexical variation* of training phrases, suggesting that the features and modeling techniques used might not fully capture the complexity of this outcome. These models are not deployment-ready; their current utility might lie in offline teacher analytics to inform instruction [22]. Future studies could explore additional features, such as discourse-level linguistic properties or turn-taking dynamics [8], other data fusion techniques [58, 59] and feature importance methods (e.g., SHAP) to identify key collaborative characteristics. Third, our analysis did not separate individual student contributions within the transcripts, potentially overlooking nuances in student interactions. However, we note that, in the present study, project-based outcomes are measured and graded as group contributions. Therefore, such analyses are out of scope for this study and may require finer-grain grading and log data where interface actions are attributable to specific students, as is common in collaborative learning systems [19]. Finally, the outcome measurement we selected may not capture certain students' unexpected behaviors (e.g., inputting a complex but irrelevant training phrase) [60]. While this case is rare in our dataset, future work could explore relevance-based outcome measures.

7. Conclusion

This study examines the accuracy of dialogue and log data in predicting process-level collaboration outcomes in middle school chatbot development. Log features best predict student productivity, while dialogue-based features are better suited to capturing training phrase richness. Multimodal fusion improves predictions for *training phrase count* and, to a lesser extent, *lexical variation*, highlighting the outcome-dependent nature of multimodal models. Our research contributes to the field of multimodal learning analytics by demonstrating the potential and limitations of integrating dialogue and log data to predict collaborative learning processes and deploying related models in real-time educational settings. We also contribute insights into the interplay between student interactions and linguistic outputs in natural language programming tasks. By leveraging both dialogue and log data, educators and researchers may gain a more holistic understanding of student learning processes, informing the development of AI-powered collaborative learning tools in K-12 settings.

Acknowledgments

This research was supported by the National Science Foundation through grant DRL-2048480. We thank Christine Fry Wise for her copy-editing of the manuscript.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Pucher, M. Lehner, Project based learning in computer science–a review of more than 500 projects, Procedia-Social and Behavioral Sciences 29 (2011) 1561–1566.

[2] D. Kokotsaki, V. Menzies, A. Wiggins, Project-based learning: A review of the literature, Improving schools 19 (2016) 267–277.

[3] J. H. L. Koh, S. C. Herring, K. F. Hew, Project-based learning and student knowledge construction during asynchronous online discussion, The Internet and Higher Education 13 (2010) 284–291.

[4] P. Guo, N. Saab, L. S. Post, W. Admiraal, A review of project-based learning in higher education: Student outcomes and measures, International journal of educational research 102 (2020) 101586.

[5] S. W. Widyaningsih, I. Yusuf, Implementation of project-based learning (pjbl) assisted by e-learning through lesson study activities to improve the quality of learning in physics learning planning courses., International Journal of Higher Education 9 (2020) 60–68.

[6] M. H. Wilkerson-Jerde, Construction, categorization, and consensus: Student generated computational artifacts as a context for disciplinary reflection, Educational Technology Research and Development 62 (2014) 99–121.

[7] S. Oviatt, A. Cohen, Written and multimodal representations as predictors of expertise and problem-solving success in mathematics, in: Proceedings of the 15th ACM on International conference on multimodal interaction, 2013, pp. 599–606.

[8] J. Yoo, J. Kim, Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns, International Journal of Artificial Intelligence in Education 24 (2014) 8–32.

[9] X. Tian, A. Mannekote, C. E. Solomon, Y. Song, C. F. Wise, T. Mcklin, J. Barrett, K. E. Boyer, M. Israel, Examining llm prompting strategies for automatic evaluation of learner-created computational artifacts, in: Proceedings of the International Conference on Educational Data Mining (EDM), 2024, pp. 698–706.

[10] M. A. Samadi, N. Nixon, Cultural diversity in team conversations: A deep dive into its effects on cohesion and team performance, in: Proceedings of the International Conference on Educational Data Mining (EDM), 2024, pp. 821–827.

[11] A. E. Stewart, Z. Keirn, S. K. D'Mello, Multimodal modeling of collaborative problem-solving facets in triads, User Modeling and User-Adapted Interaction 31 (2021) 713–751.

[12] H. Acosta, S. Lee, B. Mott, H. Bae, K. Glazewski, C. Hmelo-Silver, J. Lester, Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning (2024).

[13] X. Tian, A. E. Griffith, Z. Price, K. E. Boyer, K. Tang, Investigating linguistic alignment in collaborative dialogue: A study of syntactic and lexical patterns in middle school students, Language and Speech (2024) 00238309241234565.

[14] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, J. Lester, Automatically recognizing facial expression: Predicting engagement and frustration, in: Proceedings of International Conference on Educational Data Mining (EDM), 2013.

[15] F. Morshed Fahid, X. Tian, A. Emerson, J. B. Wiggins, D. Bounajim, A. Smith, E. Wiebe, B. Mott, K. Elizabeth Boyer, J. Lester, Progression trajectory-based student modeling for novice block-based programming, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 189–200.
[16] S. Marwan, J. Jay Williams, T. Price, An evaluation of the impact of automated programming hints on performance and learning, in: Proceedings of the 2019 ACM Conference on International Computing Education Research, 2019, pp. 61–70.
[17] A. Emerson, F. J. Rodríguez, B. Mott, A. Smith, W. Min, K. E. Boyer, C. Smith, E. Wiebe, J. Lester, Predicting early and often: Predictive student modeling for block-based programming environments., Proceedings of International Conference on Educational Data Mining (EDM) (2019).
[18] C. Borchers, J. Zhang, R. S. Baker, V. Aleven, Using think-aloud data to understand relations between self-regulation cycle characteristics and student performance in intelligent tutoring systems, in: Proceedings of the 14th Learning Analytics and Knowledge Conference, 2024, pp. 529–539.
[19] C. Borchers, K. Yang, J. Lin, N. Rummel, K. R. Koedinger, V. Aleven, Combining dialog acts and skill modeling: What chat interactions enhance learning rates during ai-supported peer tutoring?, in: Proceedings of the International Conference on Educational Data Mining (EDM), 2024.
[20] J. Zhang, C. Borchers, V. Aleven, R. S. Baker, Using large language models to detect self-regulated learning in think-aloud protocols, in: Proceedings of the International Conference on Educational Data Mining (EDM), 2024.
[21] D. Spikol, E. Ruffaldi, M. Cukurova, Using multimodal learning analytics to identify aspects of collaboration in project-based learning, in: CSCL'17: The 12th International Conference on Computer Supported Collaborative Learning, volume 1, International Society of the Learning Sciences., 2017, pp. 263–270.
[22] L. Yan, V. Echeverria, Y. Jin, G. Fernandez-Nieto, L. Zhao, X. Li, R. Alfredo, Z. Swiecki, D. Gašević, R. Martinez-Maldonado, Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning, British Journal of Educational Technology 55 (2024) 1900–1925.
[23] J. K. Olsen, K. Sharma, N. Rummel, V. Aleven, Temporal analysis of multimodal data to predict collaborative learning outcomes, British Journal of Educational Technology 51 (2020) 1527–1547.
[24] K. Sharma, M. Giannakos, Multimodal data capabilities for learning: What can multimodal data tell us about learning?, British Journal of Educational Technology 51 (2020) 1450–1484.
[25] S. Karumbaiah, C. Borchers, T. Shou, A.-C. Falhs, P. Liu, T. Nagashima, N. Rummel, V. Aleven, A spatiotemporal analysis of teacher practices in supporting student learning and engagement in an ai-enabled classroom, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 450–462.
[26] W. Chango, J. A. Lara, R. Cerezo, C. Romero, A review on data fusion in multimodal learning analytics and educational data mining, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12 (2022) e1458.
[27] A. Scarlatos, A. Lan, Exploring knowledge tracing in tutor-student dialogues, arXiv preprint arXiv:2409.16490 (2024).
[28] E. Walker, N. Rummel, K. R. Koedinger, Adaptive intelligent support to improve peer tutoring in algebra, International Journal of Artificial Intelligence in Education 24 (2014) 33–61.
[29] J. Costley, How system functionality improves the effectiveness of collaborative learning, Interactive Learning Environments 30 (2022) 971–983.
[30] J. Roschelle, S. D. Teasley, The construction of shared knowledge in collaborative problem solving, in: Computer supported collaborative learning, Springer, 1995, pp. 69–97.
[31] V. Echeverria, L. Yan, L. Zhao, S. Abel, R. Alfredo, S. Dix, H. Jaggard, R. Wotherspoon, A. Osborne, S. Buckingham Shum, et al., Teamslides: a multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space, in: Proceedings of the 14th Learning Analytics and Knowledge Conference, 2024, pp. 112–122.
[32] Q. Zhou, W. Suraworachet, M. Cukurova, Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions,

Education and Information Technologies 29 (2024) 1071–1098.
[33] M. Abdelshiheed, J. K. Jacobs, S. K. D'Mello, Aligning tutor discourse supporting rigorous thinking with tutee content mastery for predicting math achievement, in: Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED'24), Recife, Brazil, 2024.
[34] A. M. Latham, K. A. Crockett, D. A. McLean, B. Edmonds, K. O'shea, Oscar: An intelligent conversational agent tutor to estimate learning styles, in: International conference on fuzzy systems, IEEE, 2010, pp. 1–8.
[35] N. Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, G. J. Stylianides, W. W. Cohen, K. R. Koedinger, Learning by teaching simstudent–an initial classroom baseline study comparing with cognitive tutor, in: International Conference on Artificial Intelligence in Education (AIED 2011), Springer, 2011, pp. 213–221.
[36] X. Tian, N. Lubold, L. Friedman, E. Walker, Understanding rapport over multiple sessions with a social, teachable robot, in: International Conference on Artificial Intelligence in Education (AIED 2020), Springer, 2020, pp. 318–323.
[37] A. Weerasinghe, A. Mitrovic, M. Van Zijl, B. Martin, Evaluating the effectiveness of adaptive tutorial dialogues in database design, in: Proceedings of the 18th International Conference on Computers in Education, 2010, pp. 33–40.
[38] S. D'Mello, A. Olney, C. Williams, P. Hays, Gaze tutor: A gaze-reactive intelligent tutoring system, International Journal of Human-Computer Studies 70 (2012) 377–398.
[39] Y. Song, G. A. Katuka, J. Barrett, X. Tian, A. Kumar, T. McKlin, M. Celepkolu, K. E. Boyer, M. Israel, Ai made by youth: A conversational ai curriculum for middle school summer camps, in: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Innovative Applications of Artificial Intelligence Conference and Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence, 2023.
[40] D. Touretzky, C. Gardner-McCune, F. Martin, D. Seehorn, Envisioning ai for k-12: What should every child know about ai?, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 9795–9799.
[41] X. Tian, A. Kumar, C. E. Solomon, K. D. Calder, G. A. Katuka, Y. Song, M. Celepkolu, L. Pezzullo, J. Barrett, K. E. Boyer, et al., Amby: A development environment for youth to create conversational agents, International Journal of Child-Computer Interaction 38 (2023) 100618.
[42] B. Hanks, S. Fitzgerald, R. McCauley, L. Murphy, C. Zander, Pair programming in education: A literature review, Computer Science Education 21 (2011) 135–173.
[43] J.-y. Kim, Predicting l2 writing proficiency using linguistic complexity measures: A corpus-based study., English teaching 69 (2014).
[44] X. Lu, The relationship of lexical richness to the quality of esl learners' oral narratives, The Modern Language Journal 96 (2012) 190–208.
[45] W. Park, H. Kwon, Implementing artificial intelligence education for middle school technology education in republic of korea, International Journal of Technology and Design Education 34 (2023) 109–135. doi:10.1007/s10798-023-09812-2.
[46] K. Lakkaraju, T. Hassan, V. Khandelwal, P. Singh, C. Bradley, R. Shah, F. Agostinelli, B. Srivastava, D. Wu, Allure: a multi-modal guided environment for helping children learn to solve a rubik's cube with automatic solving and interactive explanations, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 13185–13187. doi:10.1609/aaai.v36i11.21722.
[47] G. Yu, Lexical diversity in writing and speaking task performances, Applied linguistics 31 (2010) 236–259.
[48] B. Kondal, Effects of lexical density and lexical variety in language performance and proficiency, International Journal of IT, Engineering and Applied Sciences Research (IJIEASR) 4 (2015) 25–29.
[49] K. Rivers, E. Harpstead, K. R. Koedinger, Learning curve analysis for programming: Which concepts do students struggle with?, in: Proceedings of 2016 ACM Conference on International Computing Education Research, volume 16, ACM, 2016, pp. 143–151.
[50] M. Chi, K. Koedinger, G. Gordon, P. Jordan, Instructional factors analysis: A cognitive model for multiple instructional interventions, in: Proceedings of the 4th International Conference on

Educational Data Mining (EDM), 2011.
[51] J. Lin, S. Singh, L. Sha, W. Tan, D. Lang, D. Gašević, G. Chen, Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues, Future Generation Computer Systems 127 (2022) 194–207.
[52] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[53] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, arXiv preprint arXiv:2401.00368 (2023).
[54] S. Grover, R. Pea, Computational thinking in k–12: A review of the state of the field, Educational researcher 42 (2013) 38–43.
[55] A. Condor, Z. Pardos, Explainable automatic grading with neural additive models, in: International Conference on Artificial Intelligence in Education, Springer, 2024, pp. 18–31.
[56] J. Hessel, L. Lee, Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think!, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020). doi:10.18653/v1/2020.emnlp-main.62.
[57] K. B. Yang, C. Borchers, A.-C. Falhs, V. Echeverria, S. Karumbaiah, N. Rummel, V. Aleven, Leveraging multimodal classroom data for teacher reflection: Teachers' preferences, practices, and privacy considerations, in: European Conference on Technology Enhanced Learning, Springer, 2024, pp. 498–511.
[58] Y. Ma, Multimodal Modeling of Collaborative Learning with Adaptive Data Fusion, Ph.D. thesis, 2023.
[59] C. Cohn, E. Davalos, C. Vatral, J. H. Fonteles, H. D. Wang, M. Ma, G. Biswas, Multimodal methods for analyzing learning and training environments: A systematic literature review, arXiv preprint arXiv:2408.14491 (2024).
[60] M. Wixon, R. S. d. Baker, J. D. Gobert, J. Ocumpaugh, M. Bachmann, Wtf? detecting students who are conducting inquiry without thinking fastidiously, in: User Modeling, Adaptation, and Personalization: 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings 20, Springer, 2012, pp. 286–296.