Analyzing Middle School Students' Dialogue and Behaviors during Collaborative AI Chatbot Development Using Ordered Network Analysis

Shan Zhang[1 [0009−0003−3532−0661]], Andres Felipe Zambrano[2 [0000−0003−0692−1209]], Xiaoyi Tian[3 [0000−0002−5045−0136]], Yukyeong Song[4 [0000−0002−4084−2734]], Anthony F. Botelho[1 [0000−0002−7373−4959]], Kristy Elizabeth Boyer[1 [0000−0003−3434−3450]], Maya Israel[1 [0000−0003−0302−6559]], and Shiyan Jiang[2 [0000−0003−4781−846X]]

[1] University of Florida, Gainesville, FL, United States
{zhangshan, a.botelho, keboyer, misrael}@ufl.edu
[2] University of Pennsylvania, Philadelphia, PA, United States
{azamb13, jiang33}@upenn.edu
[3] North Carolina State University, Raleigh, NC, United States
xtian9@ncsu.edu
[4] The University of Tennessee, Knoxville, TN, United States
ysong51@utk.edu

Abstract. As Artificial Intelligence (AI) education has become a key component of K–12 curricula, activities such as designing and developing conversational agents are increasingly used as instructional practice. Prior work has primarily examined these activities by focusing on students' learning outcomes or the quality of final AI artifacts, offering limited insight into the collaborative processes through which learning unfolds during AI system development. Although the AIED community has a long history of studying collaborative learning in STEM and Computing education, the emergence of AI learning environments in which students build AI systems presents new opportunities to understand how collaboration unfolds in AI education contexts. Grounded in these foundational works, the current study examines collaborative interaction among middle school students engaged in the design and development of an AI chatbot. Using Ordered Network Analysis of students' dialogue and development actions, we characterize how collaboration is organized over time and how interaction patterns relate to chatbot quality and AI knowledge outcomes. Results reveal that higher-quality chatbots are associated with more integrated sequences linking explanation, testing, and refinement. Interaction patterns involving articulated reasoning and repeated testing and revision in response to chatbot output were also associated with stronger AI knowledge outcomes. These findings provide a process-oriented account of collaborative AI chatbot development and extend AIED research on collaborative learning processes to AI education contexts.

Keywords: Chatbot Development · Collaborative Learning · AI Literacy · Ordered Network Analysis · AI Education.

1 Introduction

Artificial intelligence (AI) has become increasingly embedded in everyday life, shaping how children and youth interact with technology [19]. Encompassing the understanding of core AI concepts, the use and creation of AI systems, and engagement with the ethical and societal implications of AI, the cultivation of AI literacy has emerged as a critical goal in K–12 education [9]. In response, frameworks for K–12 AI education have been developed to articulate what students should learn [28], how AI can be taught through developmentally appropriate pedagogical approaches [5], and why AI education matters for students' future civic participation and workforce readiness [19,27].

Building on these frameworks, prior work has implemented a wide range of AI learning experiences across formal and informal educational settings, reporting positive relationships with students' AI conceptual understanding [4], skills and performance [13], ethical awareness [1], and non-cognitive outcomes such as engagement and interest [32]. A common feature of many of these experiences is the emphasis on project-based, design-oriented activities in which students collaboratively design and refine AI artifacts such as classifiers or conversational agents [14,24]. Although this body of work has demonstrated what students can learn through collaborative AI design activities, it has offered more limited insight into how learning unfolds during the collaborative processes through which AI systems are designed and developed.

A large body of research in the Artificial Intelligence in Education (AIED) and Computer-Supported Collaborative Learning (CSCL) communities has examined how learning unfolds through collaboration across a range of domains, including programming [3,16], problem solving [6,15], and other open-ended, design-oriented activities [16,29]. Central to this work is a focus on understanding learning *as it unfolds* through moment-by-moment interaction, with prior studies modeling how collaborators coordinate dialogue, regulate joint activity, and construct shared understanding over time [6,15,16,17]. This literature provides strong foundations for studying collaborative learning as a temporally organized, interactional process, and offers guidance for analyzing how interaction patterns relate to learning and task outcomes.

The collaborative design and development of AI systems differs from many other domains in that AI artifacts often exhibit probabilistic, opaque, and emergent behaviors that are difficult for novices to predict or explain. Engaging productively with such systems therefore requires learners to interpret unexpected outputs, generate and test hypotheses, and negotiate shared explanations of system behavior through interaction [20,21]. In collaborative design settings, these sensemaking activities are distributed across dialogue and coordinated engagement with the AI system, making interaction a central locus of learning. However, much of the current AI education literature has primarily operationalized learning through post-tests or evaluations of final AI artifacts [31,33,34]. As a result, comparatively fewer studies have examined how students' collaborative dialogue and system-oriented development actions are jointly organized over time during

AI system design and development, or how such interactional patterns relate simultaneously to both design quality and AI knowledge outcomes.

To address this gap, the present study examines collaborative interaction during middle school students' AI chatbot design and development, focusing on how patterns of dialogue and development actions unfold over time and relate to both development and learning outcomes. Specifically, we address the following research questions:

- **RQ1**: How do students' collaborative dialogue and AI development actions co-occur and sequence during the chatbot design and development?
- **RQ2**: How do these interactional patterns differ between groups that produce higher- and lower-quality AI chatbots?
- **RQ3**: How are different patterns of collaborative interaction associated with students' AI knowledge outcomes?

2 Related Work

AIED and CSCL research has extensively examined how collaborative learning unfolds during open-ended tasks such as pair programming and coordinated problem-solving activities, where learners jointly plan, implement, and refine artifacts [3,15,16]. Prior work has shown that students' dialogue plays an important role in coordinating roles, negotiating meaning, and regulating joint activity [6]. Studies of small-group collaboration have identified dialogue acts such as suggestion, uptake, and elaboration that support coordination, alongside patterns of imbalanced participation that can hinder collaborative engagement [29,30].

Beyond static dialogue categorizations, AIED research has emphasized modeling collaboration as a temporally organized process. Sequence- and state-based approaches demonstrate that collaborative activity involves transitions among interactional states such as exploratory talk, confusion, and disagreement, with reasoning often interrupted by breakdowns before returning to exploration [6]. Modeling these temporal dynamics has allowed researchers to distinguish productive and unproductive collaboration trajectories and to identify patterns associated with learning outcomes [15,16]. Recent work using interpretable temporal clustering further shows that sequences linking explanation, testing, and refinement are associated with stronger learning outcomes [7]. Although these approaches have provided powerful tools for analyzing collaborative interaction, they have often focused on a single stream of activity (e.g., dialogue), offering limited insight into how multiple forms of activity (e.g., artifact development) are coordinated over time.

In AI education research, collaborative AI system design and development have been widely adopted in design-oriented project-based learning environments where students iteratively develop and test AI artifacts such as chatbots and classifiers [8,26]. Across these studies, learning has often been operationalized through outcome-oriented measures, including pre- and post-tests targeting AI concepts such as training data and classification [4,13,32], surveys capturing attitudinal or ethical outcomes [1,32], and rubric-based evaluations of final AI

artifacts [1,13]. Recent syntheses of the AI education literature similarly note a predominant reliance on post-hoc assessments and artifact evaluations to characterize learning [36]. Although these approaches have demonstrated learning gains, they provide limited insight into the interactional mechanisms through which learning occurs during the design and development of AI systems, such as how students collaboratively reason about system behavior, interpret system feedback, or coordinate design actions over time [31,33,34].

3 Methodology

3.1 Participants and Study Procedure

Data were collected from a public middle school in the southeastern United States during the Spring 2024 semester. Across six science classes ($N = 128$ students), parental consent and student assent were obtained for 100 students; all procedures were approved by the institutional IRB. Of the 97 students who reported demographic information, 49 identified as girls, 46 as boys, one as non-binary, and one preferred not to disclose. Students identified as 38 Asian, 34 White, 20 Black/African American, 6 Hispanic/Latinx, 5 self-described, 3 Native American, and 3 preferred not to disclose race/ethnicity (multiple selections allowed). The mean age was 11.7 years ($SD = 0.48$).

The study spanned ten 50-minute class sessions over four weeks. During initial sessions, students completed a pre-survey reporting prior programming experience (i.e., whether they had written a program before) and received an introduction to AI, conversational AI, and the chatbot-building environment used in the study, "AI Made By You" (AMBY). AMBY is a web-based learning environment that supports chatbot design by enabling users to define intents (main and follow-up), add training phrases and responses for intents, and iteratively test and refine chatbot behavior [25]. Figure 1 shows the interface.

Following the introductory sessions, students were randomly assigned to pairs and collaborated on scientific chatbot development in AMBY using a pair-programming approach, alternating between "driver" and "navigator" roles [6]. Student-created chatbot artifacts, including structured representations of intents, training phrases, and responses, were evaluated using a researcher-developed rubric assessing ten dimensions (e.g., intent structure, conversation design, training phrases, and response quality-scoring), with interrater reliability of $\kappa = 0.83$ across dimensions [26]. Details about the scoring rubric are available on our Open Science Framework (OSF) repository[5]. Rubric scores were aggregated into an overall **chatbot quality score** for each project. At the conclusion of the activity, students completed a paper-and-pencil **AI knowledge post-assessment** (available on OSF) consisting of 15 items (14 multiple-choice, one open-ended), each aligned with a specific learning objective related to AI and conversational agents.

[5] `https://osf.io/ga2bk/overview?view_only=5f076ab2f04144738bc3815ee49310df`

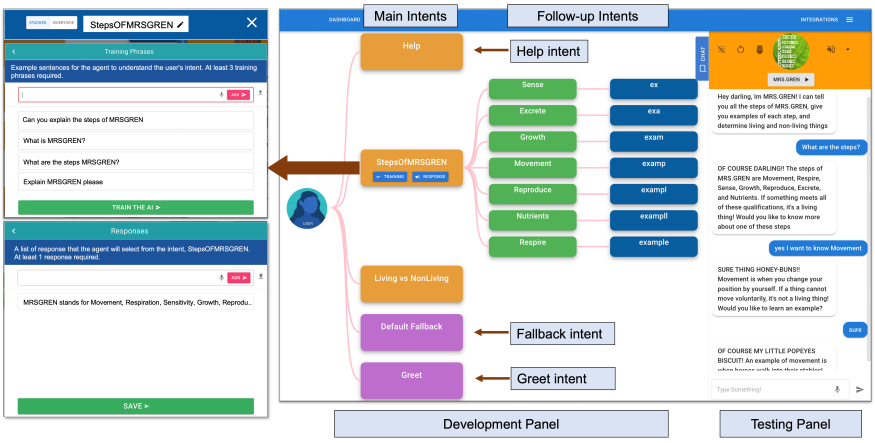

Fig. 1: Learning Environment with an example chatbot

3.2 Dialogue Tagging

We analyzed collaborative dialogue and chatbot artifacts from 47 student pairs participating in pair programming sessions. Audio and video recordings were captured using laptop-based recording software, and human transcriptions were obtained through a third-party service (https://www.rev.com/). The resulting corpus includes 32,976 utterances across 47 dyads, with an average of 701.62 utterances per dyad (SD = 293.61, min = 174, max = 1,514).

To examine how students collaborated while co-constructing chatbot artifacts, we adapted an established dialogue coding scheme from prior AIED research on collaborative learning in pair programming [6]. The scheme draws on Mercer's exploratory talk framework [11] and a dialogue act taxonomy developed by Zakaria et al. [35] to characterize collaborative discourse during programming activities. Grounded in learning sciences perspectives, the scheme captures fine-grained dialogue acts reflecting key collaborative processes, including idea generation, sense-making and reasoning, coordination and regulation, and affective or socio-relational actions. These dialogue acts characterize how learners coordinate ideas, negotiate decisions, and regulate joint activity during project-based collaboration. Because differentiating fine-grained dialogue acts required contextual information beyond transcripts alone, dialogue tagging was conducted while simultaneously viewing classroom video recordings.

To establish reliability, two researchers independently coded 20% of the data, achieving a Cohen's kappa of 0.84, indicating strong inter-rater agreement [22]. Discrepancies were resolved through discussion, after which one researcher completed the remaining coding. Table 1 presents the full set of dialogue categories used in this study; detailed definitions and examples are available on OSF.

In addition to the collaborative dialogue coding scheme, we developed an artifact development coding scheme to capture students' behaviors as they designed their chatbots. This scheme records what students were doing in the AI

Table 1: Collaborative dialogue coding scheme with relative frequencies (N = 32,976)

Dialogue	**Freq.**	**Dialogue**	**Freq.**
Explanation/justification	15.22%	Checking	1.62%
Other	12.72%	Disagreement	1.51%
Directive	12.48%	Brainstorming	1.40%
Question–other	9.13%	Acknowledge	0.97%
Repeat	7.77%	Help-seeking	0.89%
Agreement/confirmation	6.96%	Confusion	0.73%
Suggestion	6.75%	Read aloud	0.68%
Chatbot response	6.64%	Voice command	0.68%
Off-task	5.48%	Error correction	0.34%
Facilitator guidance	5.26%	Antagonistic action	0.18%
Justified Disagreement	2.35%	Frustration	0.09%
Reference	0.08%	High-order question	0.06%

chatbot-building process as they spoke. The same coding procedure was applied: two researchers jointly reviewed the videos and independently coded 20% of the data, achieving a Cohen's kappa of 0.89, indicating strong agreement. After resolving discrepancies through discussion, one researcher coded the remaining data. Table 2 presents the AI chatbot development coding scheme (further explanations and examples are available on our OSF repository).

3.3 Data Analysis

To examine differences in interaction patterns associated with chatbot performance, we conducted an Ordered Network Analysis (ONA; [23]) on both collaborative dialogue acts and AI chatbot development behaviors. ONA uses a moving window to identify connections among constructs—dialogue acts in the collaboration dialogue coding scheme and student behaviors in the AI chatbot coding scheme—based on their co-occurrence within the window, while explicitly accounting for their sequential order. Specifically, ONA distinguishes between the strength of a connection when dialogue act A is followed by dialogue act B (e.g., the group provides an explanation or justification and then explicitly states what to do next) versus when dialogue act B is followed by dialogue act A (e.g., the group explicitly states what to do next before providing an explanation). ONA also captures self-transitions, which reflect repeated occurrences of the same construct (e.g., providing multiple explanations consecutively and recurrently).

After identifying these ordered co-occurrences, ONA normalizes the transition counts using cosine normalization and constructs a network that represents the strength of connections among constructs (dialogue acts or behaviors) for each unit of analysis (e.g., individual students or collaboration pairs). In this network, nodes represent constructs, node size reflects the frequency of self-transitions (i.e., construct repetition), and edges represent the strength of connections between different constructs. Both self-transition strengths and tran-

Table 2: Chatbot development codes and relative frequencies ($N = 32{,}976$)

Code	**Freq.**	**Code**	**Freq.**
Other	21.89%	Debugging	2.29%
Testing AI Chatbot	13.74%	Add train phrases (main)	2.18%
Search online	7.77%	Add responses (greet)	1.26%
Dev responses (follow-up)	7.06%	Dev responses (fallback)	1.25%
Dev responses (main)	5.59%	Dev train phrases (greet)	1.08%
Add responses (follow-up)	5.19%	Add responses (fallback)	0.89%
Chatbot identity setup	3.97%	Explore interface	0.85%
Dev train phrases (follow-up)	3.92%	Develop entity	0.75%
Add responses (main)	3.41%	Understanding AI logic	0.50%
Setup follow-up intent	3.29%	Adding entity	0.49%
Dev train phrases (main)	2.94%	Dev train phrases (fallback)	0.46%
Add train phrases (follow-up)	2.92%	Add train phrases (greet)	0.33%
Setup main intent	2.45%	Implement entity	0.25%
Dev responses (greet)	2.44%	Setup entity	0.25%
Greet intent	0.02%	Explore entity	0.18%
Add responses (help)	0.02%	Issue resolved	0.15%
Dev responses (help)	0.02%	Info gathering	0.05%
Add train phrases (help)	0.05%	Add train phrases (fallback)	0.05%
Dev train phrases (help)	0.04%		

sitions between different constructs are quantified on the same scale, and their values are referred to as connection weights (CW).

Each unit's network is embedded in a two-dimensional space, where the relative proximity of nodes and the centroid of each unit's network indicate which constructs and connections are more frequent among particular students or groups. These networks can also be aggregated into higher-level groups (e.g., high-performing versus low-performing students), enabling statistical comparisons between groups. In addition, the connection weights for each unit can be used in further statistical analyses alongside external measures.

The ONA models were created using WebENA [10]. Groups were divided according to their chatbot performance scores using a median split. Groups with scores at or below the median (3.19) were categorized as the low chatbot performance group ($N = 25$), whereas groups with scores above the median were categorized as the high chatbot performance group ($N = 22$). For both coding schemes (collaborative dialogue acts and AI chatbot development behaviors), we used a moving window of size four to identify code co-occurrences. This window size is commonly used in similar analysis [18] and was further validated through qualitative inspection of our data. We also tested multiple window sizes ranging from 2 to 10 without observing substantive changes in the overall patterns or group differences. Each group's data were segmented by day, reflecting the assumption that discourses or behaviors occurring on the same day are more strongly connected than those separated by multiple days. Finally, for visualization purposes, we excluded codes that were infrequent in the data and did not exhibit meaningful differences between the groups under study (CW < 0.01).

To examine whether chatbot performance was associated with students' prior programming experience, we analyzed the distribution of students' responses to the pre-survey item, whether students had prior experience with written programming, across the high and low chatbot score groups. Responses to this item were cross-tabulated with chatbot performance group (high vs. low) to assess potential differences in prior programming experience between groups. This analysis allowed us to evaluate whether observed differences in chatbot performance were potentially related to students' prior programming exposure rather than differences in collaborative interaction processes.

We examined associations between students' AI knowledge and interaction patterns using Spearman's rank-order correlations. Because the AI knowledge assessment was designed around the chatbot system and students had no prior experience with this system, individual scores were aggregated at the group level ($M = 25.8$, $SD = 5.71$) and correlated with ONA-derived connection weights capturing the strength of ordered transitions among collaborative dialogue acts and AI chatbot development behaviors. Spearman's correlation was used to accommodate non-normally distributed network weights and assess monotonic relationships between AI knowledge and interaction dynamics. To control for false discoveries, we applied a Benjamini–Hochberg correction [2] to individual correlations and conducted a Monte Carlo analysis [12] with 10,000 iterations to evaluate whether the overall pattern of observed correlations was unlikely to have occurred by chance.

4 Results

Figure 2 presents the ordered network analysis (ONA) comparing collaborative dialogue patterns between groups with high and low chatbot performance (CWs for both coding schemes are available on our OSF repository). Across both groups, interaction networks are primarily organized around the dialogue acts of *explanation* and *directive*, which exhibit the strongest self-loop connection weights (CW > 0.2). These dialogue acts capture students' articulation of reasoning and clarification of ideas, as well as utterances used to guide or coordinate partner actions during collaborative AI chatbot construction. A Mann–Whitney U test comparing the two ordered network models reveals significant differences in the dialogue structure between the groups ($p < 0.001$).

As shown in Figure 2, groups with high chatbot scores exhibit higher connection weights for *explanation* (CW = 0.345 vs. 0.282), *directive* (CW = 0.257 vs. 0.201), *repeat* (CW = 0.119 vs. 0.100), and *off-task* (CW = 0.249 vs. 0.225), along with stronger ordered transitions among these dialogue acts. Within the coding scheme, *repeat* reflects the verbal repetition of typed input or previously stated content to confirm or reinforce actions, while *off-task* captures brief departures from the task that do not directly advance chatbot development. Bidirectional transitions between *directive* and *explanation* are more pronounced in high-performing groups (directive $\rightarrow$ explanation: CW = 0.175 vs. 0.139; expla-

nation → directive: CW = 0.181 vs. 0.152), indicating recurrent coordination between action guidance and articulated reasoning during chatbot construction.

Groups with high chatbot scores exhibited stronger transitions involving *justified disagreement*, including self-loops (CW = 0.051 vs. 0.034) and transitions to *explanation* (CW = 0.051 vs. 0.034), reflecting disagreement accompanied by articulated reasoning. By contrast, groups with low chatbot scores showed higher connection weights for *question–other* (CW = 0.134 vs. 0.093), *chatbot response* (CW = 0.181 vs. 0.146), and *suggestion* (CW = 0.084 vs. 0.050), indicating greater emphasis on procedural questioning, proposal generation, and orientation toward system output. Overall, high-performing groups demonstrated more densely connected transitions among explanation, directive, repetition, and disagreement acts, whereas low-performing groups exhibited interaction patterns centered on procedural and chatbot-response-focused sequences.

Figure 3 presents the ordered network comparison of AI chatbot development behaviors between groups with high and low chatbot performance. In both groups, the strongest self-loop connections occur for *testing* and *developing responses* (CW > 0.15), reflecting repeated engagement in evaluating chatbot output and refining responses. A Mann–Whitney U test indicates significant differences in the structure of development behaviors between the groups ($p = 0.015$).

Groups producing higher-quality chatbots exhibited stronger self-loops for *testing* (CW = 0.564 vs. 0.511), *developing responses for the main intent* (CW = 0.229 vs. 0.187), and *adding responses for both main and follow-up intents*

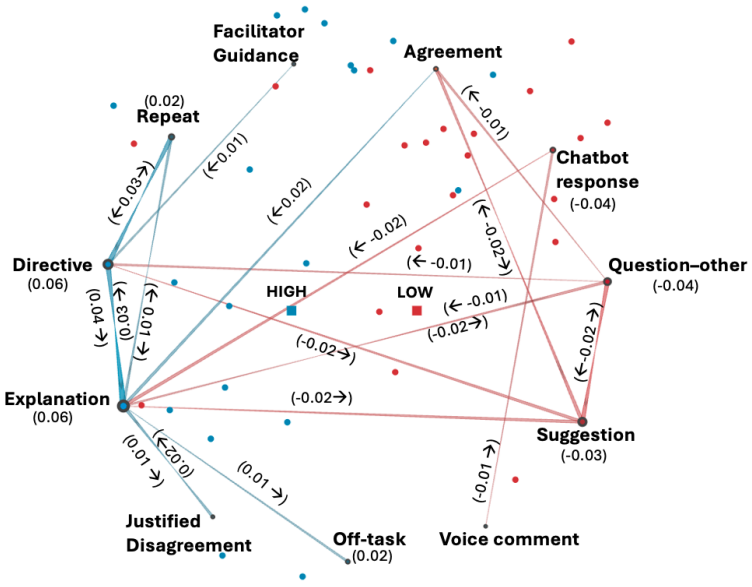

Fig. 2: Ordered Network Analysis comparing collaborative dialogue acts between groups with high (blue) and low (red) chatbot performance.

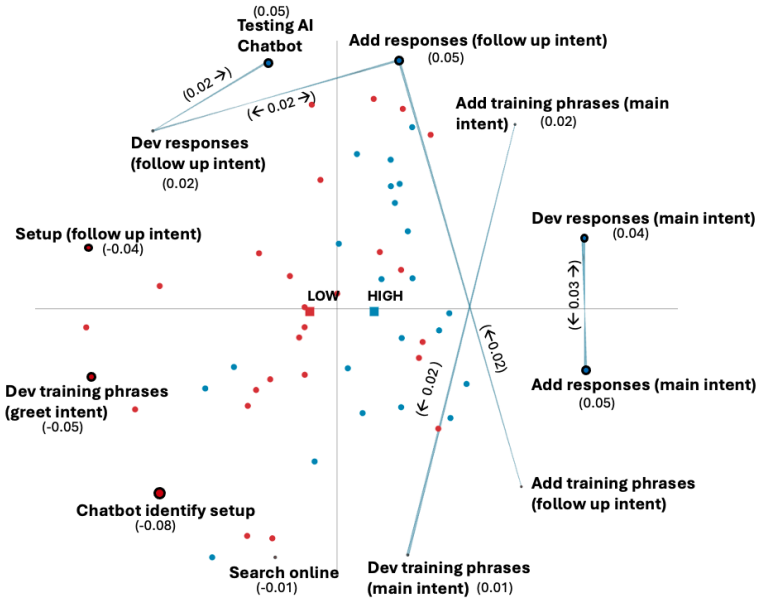

Fig. 3: Ordered Network Analysis comparing AI chatbot development behaviors between groups with high (blue) and low (red) chatbot performance.

(CW = 0.121 vs. 0.074; CW = 0.180 vs. 0.134), alongside more frequent transitions among development and testing activities. These patterns reflect iterative movement across response conceptualization, implementation, and evaluation. In contrast, groups with lower chatbot scores showed higher self-loop weights for activities such as *chatbot identity setup* (CW = 0.207 vs. 0.123), *developing training phrases for the greet intent* (CW = 0.077 vs. 0.028), and *setting up follow-up intents* (CW = 0.124 vs. 0.083), indicating extended engagement within individual development stages and fewer transitions into testing and response refinement.

Importantly, differences in chatbot performance were not explained by prior programming experience. Although the low-performing group included more students reporting prior programming experience on both survey items (*Yes, Yes*: $n = 4$), the high-performing group included students with mixed, partial, or no prior experience (e.g., *Yes, No*: $n = 6$; *Yes, Don't Know*: $n = 3$; *No, No*: $n = 3$), suggesting that observed performance differences are more closely related to interactional patterns than to prior experience alone.

Consistent with this interpretation, associations between interaction patterns and AI knowledge outcomes revealed a non-random structure. Across 250 Spearman correlations examining collaborative dialogue and development behaviors, 31 yielded p-values below 0.05 (all results available on the OSF repository). Although none remained significant after Benjamini–Hochberg correction, the

Monte Carlo analysis indicated that observing this number of correlations by chance is unlikely ($p < 0.001$; 95% CI: 0 to 20 significant results due to chance), suggesting that interaction patterns during collaborative AI chatbot development are meaningfully related to students' AI knowledge outcomes.

For collaborative dialogue, AI knowledge outcomes were positively associated with interaction sequences centered on *explanation* and *repeat*, including bidirectional connections between *explanation* and *repeat* ($\rho = 0.417$; $\rho = 0.337$) and between *explanation* and *chatbot response* ($\rho = 0.357$; $\rho = 0.343$). In contrast, AI knowledge outcomes were negatively associated with sequences dominated by *directive* and *suggestion* acts, including *directive–suggestion* ($\rho = -0.467$; $\rho = -0.418$), *suggestion* self-loops ($\rho = -0.333$), *suggestion–question–other* ($\rho = -0.319$; $\rho = -0.314$), as well as *directive* self-loops ($\rho = -0.288$) and *directive–question–other* transitions ($\rho = -0.291$). Thus, these patterns indicate that collaborative dialogue supporting explanation, repetition, and mutual grounding is more strongly associated with AI knowledge gains, whereas interaction sequences emphasizing procedural direction may limit opportunities for conceptual understanding.

For chatbot development behaviors, higher AI knowledge outcomes were associated with stronger *testing* self-loops ($\rho = 0.372$) and refinement-oriented transitions, including *develop responses for main intent–testing* ($\rho = 0.308$) and *testing–search online* ($\rho = 0.293$). In contrast, several of the strongest negative associations involved connections with *adding training phrases for follow-up intent*, including its coupling with *adding responses for follow-up intent* ($\rho = -0.496$), *develop responses for follow-up intent* ($\rho = -0.454$), *setup follow-up intent* ($\rho = -0.377$), and *adding responses for main intent* ($\rho = -0.344$), characterizing how development activity patterns align with AI knowledge outcomes. Overall, iterative testing and refinement of core intents align with stronger AI knowledge outcomes, whereas extensive focus on follow-up intent configuration may be less conceptually productive.

5 Discussion

This study investigated how collaborative interaction unfolds during AI chatbot design and development and how these interactional patterns relate to design and learning outcomes. Our findings characterize collaborative AI chatbot design and development as a temporally organized process in which dialogue and development actions are closely intertwined (RQ1). Across groups, students' activity was structured around recurring interactional sequences that linked explanation, directive coordination, and engagement with chatbot behavior. These patterns indicate that collaborative AI chatbot design and development involve ongoing cycles of articulating ideas, testing responses, and revising designs, rather than a more linear progression. This aligns with AIED work showing that collaborative learning can be productively represented as transitions among problem-solving modes over time, rather than as isolated frequencies of talk moves [6,15].

Examining differences between groups (RQ2), the analyses reveal that groups whose chatbots demonstrated higher quality exhibited more interconnected dialogue and development patterns, with frequent transitions among explanation, testing, and response refinement. In contrast, lower-quality chatbots were associated with interaction patterns that emphasized more segmented or procedural trajectories of activity. These differences were not explained by students' prior programming experience, suggesting that how students coordinated their interaction during design played an important role in shaping design outcomes.

By analyzing students' AI knowledge outcomes (RQ3), we observe that interaction sequences involving articulated reasoning, repetition of prior contributions, and engagement with chatbot responses were positively associated with AI knowledge scores, whereas patterns dominated by directive or suggestion-oriented exchanges showed negative associations. Although these relationships do not establish causality, they point to alignment between how collaborative interaction is organized and variation in students' learning outcomes.

6 Limitations and Future Work

Several limitations should be considered when interpreting these findings. First, the analyses focus on middle school students engaged in a single AI chatbot design and development, which may limit generalizability to other age groups, learning contexts, or AI tools, where collaborative interaction may take different forms. Second, although the analytic approach captures fine-grained temporal coordination between dialogue and development actions, it does not account for all aspects of interaction (e.g., gesture, affect). Finally, the observed associations remain correlational and therefore do not establish causal relationships. Future work could extend this approach by incorporating additional data modalities, examining longitudinal trajectories across activities, exploring how instructional supports impact interactional organization, and further developing scalable and interpretable analytic methods for studying collaborative AI learning processes.

7 Conclusion

This study examined collaborative interaction during middle school students' AI chatbot design and development by analyzing the temporal coordination of students' dialogue and development actions. Interpreted through a collaborative learning perspective, the findings highlight the role of explanation, coordination, and iterative engagement with system behavior in shaping both design and learning outcomes. From a design perspective, supporting interactional cycles that integrate explanation, testing, and revision may foster more coordinated collaborative AI system design and development. From a research perspective, the results underscore the value of integrating multiple process data streams to better capture how learning unfolds during AI system design and development. Overall, this work builds on previous research on collaboration by modeling the temporal coordination of dialogue and development actions, extending

process-oriented analyses to the context of collaborative AI system design and development.

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. Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., Breazeal, C.: Children as creators, thinkers and citizens in an ai-driven future. Computers and Education: Artificial Intelligence **2**, 100040 (2021)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
3. Carpenter, D., Emerson, A., Mott, B.W., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Lester, J.C.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In: Int'l conference on artificial intelligence in education. pp. 55–66. Springer (2020)
4. Chiu, T.K., Meng, H., Chai, C.S., King, I., Wong, S., Yam, Y.: Creation and evaluation of a pretertiary artificial intelligence (ai) curriculum. IEEE Transactions on Education **65**(1), 30–39 (2021)
5. Dogan, S., Nalbantoglu, U.Y., Celik, I., Agacli Dogan, N.: Artificial intelligence professional development: a systematic review of tpack, designs, and effects for teacher learning. Professional Development in Education **51**(3), 519–546 (2025)
6. Earle-Randell, T.V., Wiggins, J.B., Ruiz, J.M., Celepkolu, M., Boyer, K.E., Lynch, C.F., Israel, M., Wiebe, E.: Confusion, conflict, consensus: Modeling dialogue processes during collaborative learning with hidden markov models. In: Int'l Conference on Artificial Intelligence in Education. pp. 615–626. Springer (2023)
7. Kim, Y.J., Hong, D., Min, W., Chaturvedi, S., Hmelo-Silver, C.E., Lester, J.: Collaborative problem-solving dialogue analysis with interpretable temporal clustering. In: Artificial Intelligence in Education: 26th Int'l Conference, AIED 2025, Palermo, Italy, July 22–26, 2025, Proc., Part III. p. 30–44. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-3-031-98420-4_3
8. Kokotsaki, D., Menzies, V., Wiggins, A.: Project-based learning: A review of the literature. Improving schools **19**(3), 267–277 (2016)
9. Long, D., Magerko, B.: What is ai literacy? competencies and design considerations. In: Proc. of the 2020 CHI conference on human factors in computing systems. pp. 1–16 (2020)
10. Marquart, C.L., Hinojosa, C., Swiecki, Z., Eagan, B., Shaffer, D.W.: Epistemic network analysis (version 1.7. 0)[software]. Available from app. epistemicnetwork. org (2018)

11. Mercer, N.: Words and minds: How we use language to think together. Routledge (2002)
12. Metropolis, N., Ulam, S.: The monte carlo method. Journal of the American statistical association **44**(247), 335–341 (1949)
13. Park, W., Kwon, H.: Implementing artificial intelligence education for middle school technology education in republic of korea. Int'l journal of technology and design education **34**(1), 109–135 (2024)
14. Pearce, K., Alghowinem, S., Breazeal, C.: Build-a-bot: teaching conversational ai using a transformer-based intent recognition and question answering architecture. In: Proc. of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 16025–16032 (2023)
15. Rodríguez, F.J., Boyer, K.E.: Discovering individual and collaborative problem-solving modes with hidden markov models. In: Int'l conference on artificial intelligence in education. pp. 408–418. Springer (2015)
16. Rodríguez, F.J., Kerby, N.D., Boyer, K.E.: Repairing disengagement in collaborative dialogue for game-based learning. In: Int'l Conference on Artificial Intelligence in Education. pp. 807–810. Springer (2013)
17. Roll, I., Wylie, R.: Evolution and revolution in artificial intelligence in education. Int'l journal of artificial intelligence in education **26**(2), 582–599 (2016)
18. Siebert-Evenstone, A.L., Irgens, G.A., Collier, W., Swiecki, Z., Ruis, A.R., Shaffer, D.W.: In search of conversational grain size: Modeling semantic structure using moving stanza windows. Journal of Learning Analytics **4**(3), 123–139 (2017)
19. Song, Y., Weisberg, L.R., Zhang, S., Tian, X., Boyer, K.E., Israel, M.: A framework for inclusive ai learning design for diverse learners. Computers and Education: Artificial Intelligence **6**, 100212 (2024)
20. Stahl, G.: Contributions to a theoretical framework for cscl. In: Computer support for collaborative learning. pp. 62–71. Routledge (2002)
21. Stahl, G.: Guiding group cognition in cscl. Int'l Journal of Computer-Supported Collaborative Learning **5**(3), 255–258 (2010)
22. Sun, S.: Meta-analysis of cohen's kappa. Health Services and Outcomes Research Methodology **11**(3), 145–163 (2011)
23. Tan, Y., Ruis, A.R., Marquart, C., Cai, Z., Knowles, M.A., Shaffer, D.W.: Ordered network analysis. In: Int'l Conference on Quantitative Ethnography. pp. 101–116. Springer (2022)
24. Tian, X.: Designing for Children to Build Conversational Agents and Learn About Artificial Intelligence. Ph.d. dissertation, University of Florida (2024)
25. Tian, X., Kumar, A., Solomon, C.E., Calder, K.D., Katuka, G.A., Song, Y., Celepkolu, M., Pezzullo, L., Barrett, J., Boyer, K.E., et al.: Amby: A development environment for youth to create conversational agents. International Journal of Child-Computer Interaction **38**, 100618 (2023)
26. Tian, X., Mannekote, A., Solomon, C.E., Song, Y., Wise, C.F., Mcklin, T., Barrett, J., Boyer, K.E., Israel, M.: Examining llm prompting strategies for automatic evaluation of learner-created computational artifacts. In: Proceedings of the 17th international conference on educational data mining. pp. 698–706 (2024)
27. Touretzky, D., Gardner-McCune, C., Martin, F., Seehorn, D.: Envisioning ai for k-12: What should every child know about ai? In: Proc. of the AAAI conference on artificial intelligence. vol. 33, pp. 9795–9799 (2019)
28. Touretzky, D., Gardner-McCune, C., Seehorn, D.: Machine learning and the five big ideas in ai. Int'l journal of artificial intelligence in education **33**(2), 233–266 (2023)

29. Tsan, J., Lynch, C.F., Boyer, K.E.: "alright, what do we need?": A study of young coders' collaborative dialogue. Int'l Journal of Child-Computer Interaction **17**, 61–71 (2018)
30. Tsan, J., Vandenberg, J., Zakaria, Z., Boulden, D.C., Lynch, C., Wiebe, E., Boyer, K.E.: Collaborative dialogue and types of conflict: An analysis of pair programming interactions between upper elementary students. In: Proc. of the 52nd ACM technical symposium on computer science education. pp. 1184–1190 (2021)
31. Weng, X., Ye, H., Dai, Y., Ng, O.l.: Integrating artificial intelligence and computational thinking in educational contexts: A systematic review of instructional design and student learning outcomes. Journal of Educational Computing Research **62**(6), 1420–1450 (2024)
32. Xia, Q., Chiu, T.K., Lee, M., Sanusi, I.T., Dai, Y., Chai, C.S.: A self-determination theory (sdt) design approach for inclusive and diverse artificial intelligence (ai) education. Computers & education **189**, 104582 (2022)
33. Yim, I.H.Y., Su, J.: Artificial intelligence (ai) learning tools in k-12 education: A scoping review. Journal of Computers in Education **12**(1), 93–131 (2025)
34. Yoder, S., Tatar, C., Aderemi, I., Boorugu, S., Jiang, S., Akram, B.: Gaining insight into effective teaching of ai problem-solving through csedm: A case study. In: 5th Workshop on CS EDM (2020)
35. Zakaria, Z., Vandenberg, J., Tsan, J., Boulden, D.C., Lynch, C.F., Boyer, K.E., Wiebe, E.N.: Two-computer pair programming: Exploring a feedback intervention to improve collaborative talk in elementary students. Computer Science Education **32**(1), 3–29 (2022)
36. Zhang, S., Ganapathy Prasad, P., Schroeder, N.L.: Learning about ai: A systematic review of reviews on ai literacy. Journal of Educational Computing Research p. 07356331251342081 (2025)