Explainable AI as an Interactional Resource: From Learning and Dialogue Patterns in Programming Education

Tianyi Chen[1]

[1]*The University of Hong Kong, Pokfulam, Hong Kong*

Abstract

Explainable artificial intelligence (XAI) is increasingly promoted as a means to enhance transparency and trust in AI-supported learning environments. However, existing research has largely evaluated explainability through performance outcomes or usability measures, with less attention to how learners actively engage with explanations during learning. This study reports preliminary findings from an ongoing classroom-based investigation examining learner interaction with AI-generated explanations in an introductory R programming context. A mixed-method approach was adopted. Learning engagement and conceptual understanding were measured across three instructional sessions, while learner-AI dialogue was analysed using an adapted version of the Toolkit for Systematic Educational Dialogue Analysis (T-SEDA). Emotional stress was also examined as a moderating factor influencing engagement behaviour during challenging tasks. Preliminary findings indicate that learners in the XAI condition demonstrated higher engagement and greater conceptual gains compared with control conditions. Dialogue analysis further revealed increased reasoning-oriented interaction patterns when stepwise explanations were available. Moderation analysis suggested that emotional stress may influence engagement stability during explanation-supported learning. Overall, these findings suggest that explainability functions not only as a technical feature but as an interactional resource shaping learner behaviour. The study contributes to emerging research on XAI in education by integrating outcome-based and interaction-based perspectives within authentic classroom settings.

Keywords

Explainable AI, Learning Engagement, Conceptual Understanding, Dialogic Learning, Emotional Stress, Programming Education

1. Introduction

Artificial intelligence (AI) is rapidly becoming embedded in everyday learning environments, supporting activities such as feedback generation, tutoring, and problem-solving assistance [1, 2]. As AI increasingly take on instructional roles, concerns about transparency, trust, and pedagogical reliability have become more prominent. Explainable artificial intelligence (XAI) has been widely proposed as a response to these concerns to make system outputs understandable and to support more informed user interaction [3, 4]. In educational settings, explanations are expected to help learners understand not only what an answer is, but also why it is correct, thereby supporting deeper forms of learning.

A central issue in this context is the development of conceptual understanding. In AI learning environments, learners may successfully complete tasks using automated guidance while failing to construct meaningful mental models of the underlying concepts. Such patterns risk reinforcing procedural performance without genuine comprehension. Transparent explanations have the potential to address this limitation by making reasoning processes visible, allowing learners to relate new information to prior knowledge and refine their conceptual structures [5]. However, empirical evidence on the XAI effectiveness in supporting conceptual understanding remains mixed. Some studies suggest well-designed explanations enhance reasoning and comprehension, while others warn that uncritical reliance on AI outputs may lead to superficial understanding or misplaced confidence [6]. These tensions highlight the need for more process-oriented research of how explanations are actually used during learning.

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.

u3011374@connect.hku.hk (T. Chen)

Despite growing interest in explainability, much of the existing literature evaluates AI explanations primarily through performance outcomes, usability metrics, or learner satisfaction [7]. While these indicators provide useful information, they offer only indirect evidence of how explanations influence learning processes. In authentic learning situations, explanations become meaningful only when learners actively engage with them. Without examining these interaction processes, it remains difficult to determine whether explanations genuinely support learning or simply provide surface-level reassurance.

Interaction-based perspectives on learning provide a useful lens for understanding how explanations function in educational contexts. Dialogic theories emphasise that learning emerges through meaning-making processes that occur during interaction, rather than from information delivery alone [8, 9]. From this perspective, explanations should be understood not as static outputs but as resources that learners interpret and negotiate during problem-solving. Similarly, the Active-Constructive-Interactive (ACI) framework suggests that learning becomes more effective when learners actively construct meaning or engage in interactive dialogue around instructional content [10]. These perspectives highlight the importance of analysing learner responses to explanations rather than focusing solely on final outcomes.

Emotional conditions also play an key role in shaping how learners interact with explanations. Confusion or difficulty are common in complex learning tasks, particularly in programming contexts where small errors can disrupt progress. Emotional responses such as stress or frustration may influence whether learners persist in engaging with explanations or disengage from cognitively demanding tasks [11, 12]. However, these emotional dynamics still remain underexplored in studies of XAI in higher education. Understanding how emotional stress influences engagement and conceptual understanding may therefore provide insights into when explainability supports learning and when it fails.

Programming education provides a particularly suitable context for investigating these processes. Introductory programming environments often involve structured reasoning, incremental problem-solving, and frequent debugging, all of which rely heavily on explanatory support [13]. At the same time, programming tasks commonly trigger uncertainty and cognitive overload, creating moments where learners must decide whether to engage with explanations or abandon them. These features make programming classrooms valuable environments for examining how explanations influence both behavioural engagement and conceptual development.

Building on these theoretical and contextual considerations, this study examines how learners interact with AI-generated explanations during short instructional sessions. Rather than treating explainability as a static system feature, the study conceptualises explanation use as a dynamic learning process involving behavioural engagement, conceptual understanding, and emotional regulation. Dialogue-based analysis methods provide a structured way to capture these processes. In particular, the Toolkit for Systematic Educational Dialogue Analysis (T-SEDA) offers a systematic framework for identifying patterns of questioning, clarification, and elaboration during instructional dialogue [9]. Applying such frameworks to AI-supported learning contexts enables more fine-grained examination of how explanations shape learner behaviour over time.

Based on these considerations, the study addresses the following research questions:

RQ1: To what extent does the use of AI-generated explanations influence learners' engagement and conceptual understanding in programming education context?

RQ2: How does learners' emotional stress relate to variation in engagement and conceptual understanding during explanation-supported tasks?

RQ3: What interaction patterns emerge when learners engage with AI-generated explanations, and how do these patterns evolve across successive learning sessions?

2. Methodology

2.1. Research Design

This study employed a mixed-methods randomised controlled design to examine how AI-generated explanations influence learning engagement, conceptual understanding, and interaction patterns. The

design integrated quantitative outcome measures with dialogue-based process analysis, allowing examination of both learning results and learner-AI interaction dynamics. Three instructional conditions were compared: (1) Control – learning without AI support; (2) Non-explainable AI – AI responses provided direct answers without visible reasoning; (3) Explainable AI (XAI) – AI responses included structured step-by-step reasoning.

2.2. Participants and Group Allocation

A power analysis using the `pwr` package indicated that a minimum sample of 66 participants would be required to detect a medium effect size ($f^2 = 0.15$) with $\alpha = .05$ and power = .80. A total of 72 undergraduate students aged 18–25 were recruited from multiple higher education institutions. Participants were randomly assigned to one of three conditions: Control (n = 24), Non-explainable AI (n = 24) , and Explainable AI (n = 24) . Participants represented mixed academic backgrounds but met baseline digital literacy requirements. All participants provided informed consent before participation. Data were anonymised and stored securely in accordance with institutional ethical guidelines and responsible AI principles, including transparency and accountability in automated reasoning systems. All data handling procedures followed standard data protection principles, including removal of identifiable metadata and restricted access to interaction logs. AI-generated explanations were derived from curated instructional materials to reduce risks of misleading or fabricated reasoning.

2.3. Learning Context and Instructional Environment

The intervention was conducted within a short introductory R programming module. The intervention consisted of three sessions, each lasting approximately 30 minutes, delivered across three consecutive weeks at the same scheduled time. All instructional sessions were delivered online using a video-conferencing platform as the shared learning platform. Teaching materials were provided as pre-recorded AI-avatar lectures developed specifically for this study, forming a progressive sequence of introductory R programming content across the three sessions. All participants attended the same recorded instruction before completing practice tasks under their assigned condition. Interaction data were logged automatically through the platform interface. In the control condition, learners communicated with instructors and peers using the conference platform chat function, and all text interactions were recorded to ensure comparability with AI-supported dialogue logs. Each session followed the same structure: Short instructional explanation, guided practice tasks, and problem-solving with optional AI interaction. All instructional materials were standardised across groups to ensure comparability. The only manipulated variable was the availability and format of AI support.

Learners in AI-supported conditions interacted with an agent embedded within the learning interface. During an orientation phase, students were shown how to ask questions such as: “Why does this code return an error and what does this function do?”, “How should I fix this output?”. Students were encouraged to formulate prompts on error interpretation, function usage, and output debugging, but were not provided with fixed prompt templates in order to preserve natural interaction behaviour.

In the non-XAI condition, responses provided concise answers without reasoning steps. In the XAI condition, responses included structured step-by-step explanations describing how outputs were generated and why corrections were required. Both AI systems were connected to curated instructional materials through retrieval-based mechanisms to reduce hallucination risks and improve reliability of responses. Explanations were generated through templates aligned with curated instructional logic. To illustrate the contrast between the two conditions, when a learner asked, “Why does `mean(x)` return `NA` when my vector contains numbers?”, the non-explainable AI returned the corrected expression `mean(x, na.rm = TRUE)` accompanied by a brief statement that this was the fix. In the explainable condition, the same query produced a response organised in three reasoning steps: (1) identifying that the input vector contained missing values, (2) explaining that R propagates `NA` through arithmetic operations by default, and (3) describing how the `na.rm` argument modifies this behaviour and why the corrected output is valid. To support pedagogical reliability, a sample of generated responses was

reviewed by two instructors for factual accuracy and instructional alignment prior to deployment.

2.4. Measures

Learning engagement was conceptualised as a multidimensional construct combining behavioural and self-reported indicators. Behavioural engagement was measured using: number of learner interactions, time spent on tasks, and persistence duration. For the control group, engagement indicators were derived from task completion time, written response activity, and recorded chat interactions within the learning platform to ensure comparability with AI-supported interaction measures. Self-reported engagement was measured using the Online Student Engagement Scale (OSE), which captures behavioural, cognitive, and emotional engagement using Likert-style responses. Conceptual understanding was assessed using a researcher-developed programming assessment focusing on reasoning rather than memorisation. Items required learners to: interpret code behaviour, predict program outputs, and explain logic behind solutions. Emotional stress was measured using the Perceived Stress Scale (PSS) administered at baseline. This measure assessed perceived stress levels during learning tasks and was used as a moderating variable. For dialog-based interaction analysis, learner–AI dialogue was analysed using an adapted version of the Toolkit for Systematic Educational Dialogue Analysis (T-SEDA). T-SEDA extends the original Scheme for Educational Dialogue Analysis (SEDA) by focusing specifically on reasoning- and explanation-oriented dialogue moves that support conceptual development. In this study, dialogue turns were coded into categories such as Clarification, Explanation, Reasoning, Connection, and Reflection. Two trained coders independently analysed sampled interaction transcripts. Inter-rater reliability was calculated using Cohen's κ, with values exceeding $.80$, indicating strong coding agreement.

2.5. Procedure

The study followed four sequential phases: (1) Pre-test phase: Participants completed baseline assessments measuring engagement, conceptual understanding, and emotional stress. (2) Intervention phase: Participants completed three weekly instructional sessions under assigned condition; (3) Interaction logging: All learner activities and AI interactions were recorded and anonymised. (4) Post-test phase: Participants completed post-intervention engagement and conceptual understanding assessments.

3. Results

3.1. Quantitiative Findings

The following findings are reported as preliminary results from an ongoing dataset.

For engagement, an ANCOVA was conducted to examine post-intervention engagement while controlling for baseline engagement levels. A significant effect of instructional condition was observed, $F(2, 65) = 3.52$, $p = .036$. Post-hoc comparisons indicated that learners in the XAI condition demonstrated significantly higher engagement than those in the control condition ($p < .05$). No significant difference was observed between the non-explainable AI and control groups. Across sessions, descriptive patterns suggested gradual increases in engagement in the XAI group, particularly during later tasks that required debugging and reasoning.

Table 1
Post-hoc Pairwise Comparisons for Engagement (ANCOVA)

Contrast	Estimate	SE	t	DF	p
Control - NormalAI	-4.76	3.08	-1.55	65	0.275
Control - XAI	-8.62	3.20	-2.69	65	0.024
NormalAI - XAI	-3.86	3.19	-1.21	65	0.453

For conceptual understanding, a second ANCOVA examined conceptual understanding while controlling for pre-test scores. Results indicated a significant effect of instructional condition, $F(2, 65) = 4.21$, $p = .019$. Learners in the XAI condition demonstrated greater gains in conceptual understanding compared with both the control and non-explainable AI groups.

Table 2
Post-hoc Pairwise Comparisons for Conceptual Understanding (ANCOVA)

Contrast	Estimate	SE	t	DF	p
Control - NormalAI	-0.16	1.28	-0.12	65	0.992
Control - XAI	-3.53	1.32	-2.67	65	0.025
NormalAI - XAI	-3.37	1.30	-2.59	65	0.031

For emotional stress, moderation analysis examined whether emotional stress influenced engagement outcomes across conditions. Emotional stress did not show a significant main effect ($p = .48$). However, the interaction between stress and the XAI condition approached significance ($\beta = 0.94$, $p = .10$).

3.2. Interaction Pattern Findings

Analysis of interaction pattern across sessions remains ongoing. Early observations suggest that some learners shifted from clarification-focused interaction toward reasoning-oriented moves over successive sessions, particularly within the XAI condition. Table 3 presents representative learner utterances drawn from the coded transcripts, illustrating the kinds of moves observed across the adapted T-SEDA categories. A recurring pattern within the XAI condition involved learners issuing an initial clarification move (e.g., asking what an error message meant), followed, after engaging with the stepwise explanation, by a reasoning move that drew on the structure of the explanation when proposing the next attempted solution. Such shifts were less frequently observed in the non-explainable condition, where interaction more often remained at the clarification level. The proportional distribution of coded dialogue moves across both AI conditions is presented in Appendix A. However, full sequential analysis has not yet been completed. Future coding will examine how interaction trajectories evolve across extended learning periods and whether early reasoning patterns predict later conceptual gains.

Table 3
Representative Learner Utterances Across Adapted T-SEDA Categories

Category	Example learner utterance
Clarification	"What does the error `object not found` actually mean here?"
Explanation	"So the function needs a numeric vector, not a list of strings."
Reasoning	"If I convert the column to numeric first, then `mean()` should work, because the input type will match what the function expects."
Connection	"This is similar to the indexing problem we ran into in the first session."
Reflection	"I think I was confused earlier about how NA propagates through calculations."

4. Discussion

4.1. RQ1: Effects of Explainability on Engagement and Conceptual Understanding

The first research question examined whether the availability of AI explanations influenced learner engagement and conceptual understanding. Preliminary quantitative patterns indicated that learners in the explainable AI condition demonstrated higher levels of engagement and greater conceptual gains compared with those in the control condition. In contrast, the non-explainable AI condition showed patterns similar to the control group, suggesting that access to AI responses alone may not be sufficient

to enhance learning outcomes. These findings support the interpretation that explanation visibility plays a meaningful role in supporting reasoning-oriented learning processes. In structured domains such as programming, conceptual understanding depends not only on obtaining correct answers but also on understanding how outputs are generated. Step-by-step explanations may therefore function as cognitive scaffolds that help learners connect procedural steps with underlying conceptual relationships.

4.2. RQ2: Emotional Conditions and Engagement Stability

The second research question explored how emotional conditions, particularly stress, related to learner engagement during AI-supported learning. Preliminary moderation patterns suggested that learners experiencing higher stress levels tended to maintain engagement when interacting with explainable AI, although this pattern did not reach conventional statistical thresholds. One possible explanation is that visible reasoning reduces uncertainty during problem-solving. Programming tasks often involve unexpected errors that can disrupt progress and increase cognitive load. When explanations explicitly show how outputs are derived, learners may experience greater clarity and reduced ambiguity, allowing them to continue working through challenges rather than disengaging.

4.3. RQ3: Interaction Patterns and Dialogic Processes

Preliminary dialogue analysis suggested differences in interaction patterns between instructional conditions. Learners in the explainable AI condition appeared more likely to engage in reasoning-oriented dialogue moves, while interactions in the non-explainable AI condition remained primarily clarification-focused. These early patterns indicate that explanation visibility may influence not only outcomes but also the nature of learner interaction. From a dialogic perspective, explanations may function as resources that learners interpret and explore rather than passively accept. However, as interaction coding remains ongoing, these interpretations should be considered provisional.

4.4. Implications for Explainable AI Design in Education

The findings highlight the importance of treating explainability as a pedagogical feature rather than solely a technical property. Explanations that explicitly represent reasoning steps appear to support conceptual engagement, particularly in domains requiring structured logic. Designing explanations that invite questioning and reflection may therefore be more beneficial than simply providing concise answers. In addition, these results emphasise the need to evaluate not only the availability of explanations but also their quality and usability. Future systems should therefore prioritise clarity, faithfulness, and alignment with learner knowledge levels to ensure meaningful reasoning explanations.

4.5. Limitations

Several limitations constrain the interpretation of the present findings. Most importantly, the sample size (N = 72) remains modest, limiting statistical sensitivity, particularly for exploratory moderation analyses. Although determined through power analysis, the sample size restricts the strength of generalisable claims and highlights the need for replication with larger cohorts. In addition, the intervention consisted of three short instructional sessions, which limits conclusions about long-term learning effects. The study was also conducted exclusively in an R programming context, a highly structured domain that may particularly benefit from stepwise reasoning. Whether similar effects would occur in less structured disciplines remains uncertain. Finally, although a sample of explanations was reviewed by instructors for accuracy prior to deployment, no systematic evaluation of explanation quality, including dimensions such as faithfulness, clarity, and pedagogical alignment across the full response set, was conducted.

4.6. Future Directions

Future research should prioritise larger-scale studies across multiple institutions to strengthen statistical power and improve generalisability. Extending this work to disciplinary contexts beyond programming will be particularly important for determining whether the observed benefits of explainable AI generalise to domains involving less structured reasoning. Longitudinal studies are also needed to examine how interaction patterns evolve over time and whether early engagement gains translate into sustained conceptual development. A more complete sequential analysis of learner–AI dialogue trajectories within and across sessions is a particularly important next step for understanding the mechanisms behind explanation-supported learning. In addition, future research should examine potential risks associated with explanation use, including overtrust and the illusion of understanding, to support responsible and pedagogically sound AI design.

Declaration on Generative AI

The author have not employed any Generative AI tools.

References

[1] W. Holmes, M. Bialik, C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning, Center for Curriculum Redesign, Boston, 2019.

[2] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educational Psychologist 46 (2011) 197–221. doi:10.1080/00461520.2011.611369.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[5] M. T. H. Chi, Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes, in: S. Vosniadou (Ed.), International Handbook of Research on Conceptual Change, Routledge, New York, 2013, pp. 61–82.

[6] L. Rozenblit, F. Keil, The misunderstood limits of folk science: An illusion of explanatory depth, Cognitive Science 26 (2002) 521–562. doi:10.1207/s15516709cog2605_1.

[7] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI), IEEE, New York, 2015, pp. 160–169. doi:10.1109/ICHI.2015.26.

[8] R. Wegerif, Dialogic Education and Technology: Expanding the Space of Learning, Springer, New York, 2007.

[9] S. Hennessy, C. Howe, N. Mercer, M. Vrikki, A dialogic inquiry approach to working with teachers in developing classroom dialogue, Research Papers in Education 35 (2020) 1–25. doi:10.1080/02671522.2019.1677752.

[10] M. T. H. Chi, R. Wylie, The icap framework: Linking cognitive engagement to active learning outcomes, Educational Psychologist 49 (2014) 219–243. doi:10.1080/00461520.2014.965823.

[11] R. Pekrun, Emotions as drivers of learning and cognitive development, in: R. Pekrun, L. Linnenbrink-Garcia (Eds.), International Handbook of Emotions in Education, Routledge, New York, 2014, pp. 23–39.

[12] J. Sweller, P. Ayres, S. Kalyuga, Cognitive Load Theory, Springer, New York, 2011. doi:10.1007/978-1-4419-8126-4.

[13] A. Robins, J. Rountree, N. Rountree, Learning and teaching programming: A review and discussion, Computer Science Education 13 (2003) 137–172. doi:10.1076/csed.13.2.137.14200.

A. Dialogue Code Distribution Across Conditions

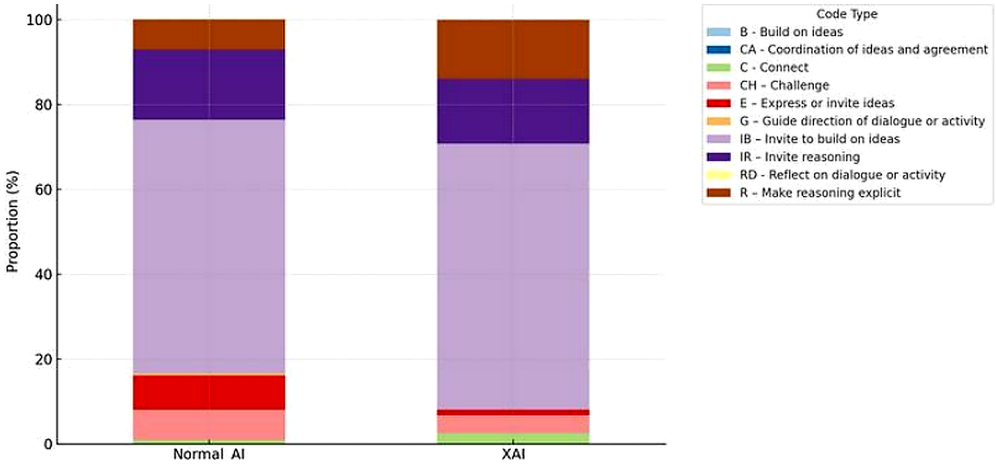

Figure 1: Proportional distribution of T-SEDA dialogue codes in the Normal AI and XAI conditions. The XAI condition shows a higher proportion of reasoning-oriented moves (R, IB, IR) relative to the Normal AI condition, in which interaction was more dominated by expressive and clarification moves (E).