Exploring Learner Engagement in an LLM-Supported Simulation Environment for Complex Systems Understanding

Amit Nair, Saarland University, amna00001@stud.uni-saarland.de
Man Su, Leibniz-Institut für Wissensmedien (IWM Tübingen), m.su@iwm-tuebingen.de
Tomohiro Nagashima, Saarland University, nagashima@cs.uni-saarland.de

Abstract: Understanding complex systems is challenging, as learners often rely on linear reasoning and struggle to recognize decentralized and emergent processes. Large language models (LLMs) offer new opportunities to support inquiry in simulation-based learning, yet how learners engage with these tools and how such engagement relates to learning outcomes remains underexplored. In this study, we investigate learner engagement within a theory-informed, LLM-supported simulation environment, based on an experiment with 69 postsecondary students that varied parameter control and questioning agency. Focusing on questioning agency, we examine how interactions with an LLM-based tutor and engagement with exploratory instructional regions relate to posttest performance. Results show that greater engagement with the tutor, particularly through content-focused interactions, is associated with higher learning gains. Time spent in exploratory regions also significantly predicts performance. These findings highlight the importance of fostering meaningful engagement with both conversational tutors and interactive features in LLM-supported environments.

Introduction

Understanding complex systems is challenging, as it requires reasoning across multiple levels (Jacobson & Wilensky, 2006). Learners often struggle to connect micro-level behaviors with macro-level patterns and to recognize structural commonalities across different phenomena (Chi, 2023). Frameworks such as Structure-Behavior-Function, Complex Systems Ontology (Hmelo-Silver & Pfeffer, 2004), and the more recent PAIR-C (Patterns, Agents, Interactions, Relations, and Causality), have been proposed to make these relationships explicit (Chi, 2023; Su et al., 2025). In practice, interactive technologies, particularly agent-based models (ABMs, i.e., NetLogo), are widely used to support complex systems learning by enabling learners to observe how local interactions give rise to emergent system-level behavior (Levy & Wilensky, 2009). Despite advances in both theory and technology, it is still challenging to integrate conceptual frameworks into the design of interactive learning environments (Yoon, 2021).

Recent advances in large language models (LLMs) have enabled context-aware conversational tutoring to be directly integrated alongside simulation-based learning environments, allowing learners to engage in empirical observation through simulations while receiving theory-informed guidance from the LLM-based tutor. NetLogo Chat, for instance, explores how an LLM-based tutor integrated within NetLogo modeling tasks can be supplied with documentation and reference examples to improve the quality of its responses to learner inquiries (Chen et al., 2024). When integrated as pedagogical supplements, LLMs have been shown effective at enhancing academic performance, promoting reflective thinking through Socratic inquiry (An et al., 2025; Xi et al., 2026).

To understand how LLM-based tutors, when combined with the PAIR-C framework, can contribute to student learning in complex systems learning environments, we developed an LLM-supported open-world simulation environment, focusing on the concept of animal flocking. Unlike traditional environments that rely on scripted progression, our environment emphasizes agency by utilizing an open-ended design where all instructional and interactive features are self-paced and accessible non-linearly (Dettori & Paiva, 2009). In this paper, our main research question is: *How does learner engagement within an open-world, self-paced LLM-supported learning environment relate to measurable learning outcomes?* Specifically, we formed two hypotheses: H1: Learner engagement with the LLM-based tutor will predict positive learning outcomes. H2: Learner engagement with the exploratory instructional regions will predict positive learning outcomes.

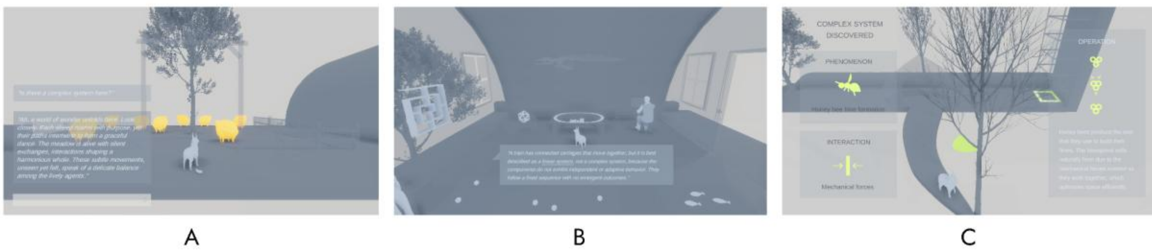
Learning environment

Simulator overview

The simulator was developed using a participatory design approach, involving collaboration with a biology teacher and a domain expert to ensure that the content was valid and pedagogically relevant (Qi & Yu, 2025). The simulator provides an immersive open-world learning environment where learners control a virtual character in the form of a dog to navigate the environment and interact with animal groups that exhibit collective behaviors (e.g., sheep forming a flock). These behaviors serve as analogies for emergent processes, allowing learners to observe how simple, local interactions can lead to complex system-level outcomes. The learning content is scaffolded based on the PAIR-C framework across three lessons: the Basic lesson, which focuses on foundational complex systems concepts (i.e., *Pattern:* Are the sheep currently showing emergent flocking behavior? *Agents:* Match each item in the environment with its most appropriate role); the Intermediate lesson, which covers the three flocking rules (i.e., separation, cohesion, and alignment) from the Boids model (Reynolds, 1987) and the *Interactions* and *Relations* dimensions of PAIR-C (i.e., What kinds of *Interactions* can be observed in this environment? Match each property of *Relations* with its corresponding description); and the Complex lesson, which emphasizes advanced ideas about *Causality* (i.e., Which of the following examples demonstrate net effects rather than chain effects?), misconceptions (i.e., Which of the following reflect misconceptions about the bat flocking phenomenon?), and transfer concepts (i.e., Could the interactions within the Earth's Climate be considered part of a complex system? Why or Why not?). Each lesson focuses on a different flock (sheep, ducks and bats), and is structured around two modes: Exploration and Evaluation.

In the Exploration mode, learners freely navigate the scene, where they can interact with non-player characters (NPCs) that respond with predefined static dialogue. Learners can also access a dynamic LLM-based Butterfly Tutor, powered by the GPT-4o model, which provides adaptive, context-aware support (see Figure 1A). The design of the butterfly agent follows three principles: (1) Theoretical alignment, grounding responses in the PAIR-C framework (Chi, 2023) to ensure conceptual consistency; (2) Environmental awareness, leveraging regional metadata to include context in its responses and avoid making definitive claims about surrounding elements it cannot verify, but intervene with certainty when it detects a misunderstanding or an attempt to deceive; and (3) Non-intrusive representation, a blinking butterfly icon appears in the top-right corner of the screen, accompanied by a brief pop-up prompt (e.g., "Alt: Call Butterfly") to remind learners of the feature without requiring immediate interaction. Additionally, it gently redirects off-topic conversations back to complex systems concepts, helping maintain focus without resorting to abrupt or explicit refusals (Tang, 2026).

Figure 1
Key features of the flocking simulator illustrating: (A) LLM-tutor interaction via a chat interface in front of a sheep flock, (B) An Exhibit Area where approaching with conceptual props (i.e., train carriages) triggers instructional text bubbles, (C) A Hidden System illustrating beehive formation



In the Evaluation mode, learners answer ten targeted questions designed based on the PAIR-C framework and receive instant feedback. The question types include multiple-choice, true-or-false, match-the-following, and open-ended questions. Responses to structured question types are automatically evaluated based on predefined correct answers, while open-ended responses are assessed using GPT-4o in combination with a predefined scoring rubric aligned with the PAIR-C framework, enabling consistent and theory-informed evaluation of learner explanations in real time. In conditions with higher parameter control agency, learners can additionally manipulate flock simulations using parameter sliders, adjusting the magnitude of key flocking rules (i.e., cohesion, alignment, and separation) alongside a physical constraining factor (i.e., speed), to observe system-level changes in real-time.

To enrich the environment, we designed two types of instructional regions, namely *Exhibit Areas* and *Hidden Systems*, which learners encounter during the Exploration mode. These regions embed conceptually meaningful visual representations within the environment (e.g., a train with multiple carriages as a prop, see Figure 1B), where approaching or interacting with these elements triggers text bubbles that present instructional explanations (e.g., distinction between linear and complex systems). *Exhibit Areas* emphasize lesson-specific content aligned with each of the three flocking scenarios, whereas *Hidden Systems* are intentionally placed off the main path to encourage exploration and sustained engagement, introducing non-flocking concepts (i.e., beehive formation, see Figure 1C) that support conceptual transfer beyond the immediate context.

Methods

We conducted a between-subjects factorial design with repeated measures (pretest and posttest). Participants were randomly assigned to one of four conditions (Table 1) based on two independent variables: *parameter control agency* (high vs low) and *questioning agency* (high vs low). In the Low Parameter Control (LowC: C0 & C2) conditions, participants had no control over the simulation parameters. Conversely, the High Parameter Control (HighC: C1 & C3) conditions granted participants full parameter control of key flocking rules (cohesion, alignment, separation) and speed, via real-time sliders. In the Low Questioning (LowQ: C0 & C1) conditions, participants only received automated, GPT-4o-generated personalized feedback on their responses during the Evaluation mode with the pre-defined targeted questions (no access to the Butterfly Tutor). In the High Questioning (HighQ: C2 & C3) conditions, participants could also interact freely with the Butterfly Tutor via a chat window during the Exploration mode. The study was approved by the university's Ethics Review Board, and all participants provided informed consent before the experiment was conducted for their data to be used.

Table 1
Experimental design

Parameter Control Agency	Questioning Agency	
	LowQ	HighQ
LowC	C0: No parameter control, No LLM-based tutor (n = 17)	C2: No parameter control, With LLM-based tutor (n = 17)
HighC	C1: With Parameter control, No LLM-based tutor (n = 17)	C3: With Parameter control, With LLM-based tutor (n = 18)

A total of 75 university students participated, with 69 completing all phases. Participation was voluntary, with a small grade bonus offered to students enrolled in a lecture. Participants first completed an online demographic survey and a pretest on flocking, adapted from an existing assessment framework for complex systems understanding (Khodr et al., 2022). Then, they attended a scheduled on-site session, working independently for 40-60 minutes on a laptop with technical support available. The study concluded with two posttests, one mirroring the pretest with reordered items to reduce recall bias, and another assessing transfer to non-flocking domains, along with a user experience survey. The simulator was disabled during testing, though participants could refer to their notes.

We collected multimodal data, including system logs, screen recordings, learner notes, observations, and interviews. This paper focuses on system log data to examine how engagement with the LLM-based tutor and exploratory instructional regions (*Exhibit Areas* and *Hidden Systems*) relates to immediate posttest performance. Note that although the study employed a factorial design for two dimensions of learner agency, the present analysis centers on questioning agency and learner engagement.

Findings and discussion

Learner engagement with the LLM-based tutor (H1)

We found that engagement with the Butterfly Tutor varied considerably among participants in the high questioning agency conditions (C2&C3). Some interacted extensively, while others engaged minimally or not at all. The number of queries ranged from 0 to 32, with a mean of 6.7 and a median of 3. Although no significant differences in posttest performance were found between tutor-present (C2&C3) and tutor-absent groups (C0&C1), higher interaction frequency within the tutor-present conditions predicted improved posttest scores (b = 0.14, p = .038, Adj. R^2 = .129). To evaluate H1, we analyzed inputs from 35 participants who generated a total of 235 inputs to the tutor. The first author manually coded all inputs, and the second author independently verified 50% of the coding, yielding 100% agreement. Five categories of learner inputs were identified: content-focused, greeting, meta, peripheral, and unrelated inputs (Table 2). The majority were content-focused (35.7%), followed by meta (28.5%), peripheral (26.8%), greeting (7.7%), and unrelated inputs (1.3%). Further regression analyses, controlling for pretest scores, revealed that content-focused inputs positively predicted posttest outcomes (b = 0.21, p = .048, Adj. R^2 = .062), as did greeting inputs (b = 1.94, p = .025, Adj. R^2 = .094). Additionally, total word count showed a positive relationship with learning gains (b = 0.023, p = .028, Adj. R^2 = .090). While greater engagement with the tutor was linked to improved learning, the models explained only a modest portion of variance, suggesting that other factors might also play a stronger role in posttest performance. Our findings indicate that it is not simply access to AI support that drives learning outcomes, but the quality and intentionality of learner-initiated interactions. Learners who proactively asked questions engaged in richer, elaborative dialogues and benefited from adaptive scaffolds that reinforced causal reasoning, consistent with prior evidence that active inquiry promotes productive learning (Su et al., 2023). Overall, this study extends current conceptions of agency in AI-supported learning (Vincoli et al., 2025), demonstrating that well-calibrated questioning agency, as opposed to unrestricted autonomy, can effectively balance guidance and exploration.

Table 2
Categories of learner inputs to the Butterfly Tutor

Category	Description	Examples
Content-focused (35.7%)	Inputs that represent direct engagement with the learning content	"What is flocking behavior?" "How do ants find food?"
Meta (28.5%)	Inputs about the simulator itself or its mechanics	"What is the purpose of this chatbot?" "Can I redo incorrect questions?"
Peripheral (26.8%)	Inputs that match the fictional setting but do not move the learner towards the learning goal	"You are indeed my true companion." "Do you see the ghost?"
Greeting (7.7%)	Inputs that represent basic salutations or farewells	"Hello" "Bye"
Unrelated (1.3%)	Inputs unrelated to the simulator or learning content	"Will I be able to pass my upcoming exams?"

Learner engagement with exploratory instructional regions (H2)

To evaluate H2, we analyzed participants' engagement time in the *Exhibit Areas* and *Hidden Systems*. A multiple regression model including time spent in each *Exhibit Area* (Lessons 1-3) and each Hidden System as predictors of posttest performance accounted for 32% of the variance. Time spent in the *Exhibit Area* associated with Lesson 1 was a significant positive predictor of posttest performance (b = 1.09, p = .004, Adj. R^2 = .323). In contrast, time spent in the *Exhibit Area* associated with Lesson 2 did not significantly predict posttest scores, while the *Exhibit Area* associated with Lesson 3 showed a small but non-significant positive trend. The lack of effect for the second *Exhibit Area* likely relates to the relatively high performance in answering the ten targeted questions in Lesson 2. It was noted that learners demonstrated higher accuracy in Lesson 2 (mean = 0.629) compared to Lesson 1 (mean = 0.445) and Lesson 3 (mean = 0.352), suggesting that the within-lesson, structured evaluation component may have provided sufficient conceptual cues for passing the lesson without reliance on additional instruction in the exploratory *Exhibit Areas*. Engagement with the *Hidden Systems* revealed an additional pattern. Time spent in the third *Hidden System* (i.e., beehive formation) showed a trend with posttest performance (b = 0.676, p = .067, Adj. R^2 = .323). This *Hidden System* was intentionally designed to be more accessible, ensuring that learners (82.6%, n = 57) would be introduced to at least one complex phenomenon distinct from flocking, thereby supporting performance on the transfer test. By comparison, the first and second *Hidden Systems* were found to be more difficult to locate, encountered by only 14.5% (n = 10) and 10.1% (n = 7) of participants respectively, and did not significantly predict posttest outcomes. The first two *Hidden Systems* were likely encountered less frequently, partly because participants prioritized more explicit, visible representations within the limited time available, rather than engaging with less explicit tasks that required long-term exploration and sustainable engagement. Nevertheless, these findings indicate that self-paced learner engagement with exploratory instructional regions can meaningfully support conceptual understanding of complex systems within the open-ended learning environment. Future research can examine how longer-term learning behaviors can allow a deeper exploration of system features, with LLM-based pedagogical tutors leveraged to sustain participation and deeper conceptual understanding.

References

An, Y., Liu, J., Acharya, N., & Hashmi, R. (2025). Enhancing Student Learning with LLM-Generated Retrieval Practice Questions: An Empirical Study in Data Science Courses (arXiv:2507.05629). arXiv. https://doi.org/10.48550/arXiv.2507.05629

Chen, J., Lu, X., Du, Y., Rejtig, M., Bagley, R., Horn, M., & Wilensky, U. (2024). Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. *Proceedings of the CHI Conference on Human Factors in Computing Systems, 13*, 1–18. https://doi.org/10.1145/3613904.3642377

Chi, M. T. H. (2023). Individualistic and collective causal knowledge structures for understanding sequential and emergent processes. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1198362

Dettori, G., & Paiva, A. (2009). Narrative Learning in Technology-Enhanced Environments An Introduction to Narrative Learning Environments.

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science, 28*(1), 127–138. https://doi.org/10.1207/s15516709cog2801_7

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *Journal of the Learning Sciences, 15*(1), 11–34. https://doi.org/10.1207/s15327809jls1501_4

Khodr, H., Kothiyal, A., Bruno, B., & Dillenbourg, P. (2022). An Assessment Framework for Complex Systems Understanding. Proceedings of the 16th International Conference on Learning Sciences. https://infoscience.epfl.ch/entities/publication/558a3ccc-a3af-4199-92e7-940cdd2944c2/files

Levy, S. T., & Wilensky, U. (2009). Crossing levels and representations: The connected chemistry (CC1) curriculum. *Journal of Science Education and Technology, 18*(3), 224–242. https://doi.org/10.1007/s10956-009-9152-8

Qi, X., & Yu, J. (2025). Participatory Design in Human-Computer Interaction: Cases, Characteristics, and Lessons: 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025. *CHI 2025 - Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–26. https://doi.org/10.1145/3706598.3713436

Reynolds, C. (1987). Flocks, Herds and Schools: A Distributed Behavioral Model. Computer Graphics, 21, 25–34. https://doi.org/10.1145/37402.37406

Su, M., Ha, J., Cortés, L. E. P., Bernier, J., Yan, L., Nelson, B. C., Bowman, J. D., & Bowman, C. D. D. (2023). Understanding museum visitors' question-asking through a mobile app. *Educational Technology Research and Development, 71*(6), 2483–2506. https://doi.org/10.1007/s11423-023-10265-6

Su, M., Chi, M. T. H., & Nagashima, T. (2025). Applying the PAIR-C Framework to foster deep understanding and address misconceptions in science education. *Journal of the Learning Sciences*, 1–46. https://doi.org/10.1080/10508406.2025.2546471

Tang, N., Qian, A., Wang, Q., Howe, E., Bullwinkel, B., Pedrelli, P., Suh, J., Heidari, H., & Shen, H. (2026). Beyond the Single Turn: Reframing Refusals as Dynamic Experiences Embedded in the Context of Mental Health Support Interactions with LLMs (arXiv:2602.01694). arXiv. https://doi.org/10.48550/arXiv.2602.01694

Vincoli, M., Scholz, N., & Nagashima, T. (2025). Multidimensional Student Agency in Learning with AI: A Conceptual Framework and Design Implications. *Artificial Intelligence in Education: 26th International Conference, AIED 2025, Palermo, Italy, July 22–26, 2025, Proceedings, Part IV*, 218–232. https://doi.org/10.1007/978-3-031-98459-4_16

Xi, L., Zhang, Y., & Wang, Q. (2026). Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. Computers & Education, 241, 105494. https://doi.org/10.1016/j.compedu.2025.105494

Yoon, S. A. (2021). Complex Systems Research in K12 Science Education: A Focus on What Works for Whom and under Which Conditions. *Systems*, 9(2), 29. https://doi.org/10.3390/systems9020029

Acknowledgements

We thank the educators, complex systems experts, and students who participated in this research.