2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte

Paper ID #52750

Active Learning with an LLM Peer in K12 STEM

Mr. Chengzhang Zhu, Rowan University

Chengzhang Zhu is currently working toward his Ph.D. degree in Electrical Computer Engineering with the Department of ECE, Rowan University. His research interests include Serious Game and Generative AI.

Dr. Ying Tang, Rowan University

Ying Tang received the B.S. and M.S. degrees from the Northeastern University, P. R. China, in 1996 and 1998, respectively, and Ph.D degree from New Jersey Institute of Technology, Newark, NJ, in 2001. She is currently a Professor of Electrical and Comput

LuoBin Cui, Rowan University

Active Learning with an LLM Peer in K12 STEM

Abstract

This paper reports on the design, classroom implementation, and evaluation of a game-based learning module aligned with relevant science and engineering curriculum standards for K12 students. Leveraging active learning principles, the module combines engaging game mechanics with a conversational LLM peer agent to provide on-demand hints and prompts that scaffold each student's inquiry process during the play. To evaluate the module's effectiveness, we conducted a mixed-methods study with three classroom conditions, employing multiple assessments: pre- and post-module concept tests measured learning gains; in-game behavioral analytics captured student engagement and inquiry behaviors; dialogue transcripts from student–LLM interactions were coded to analyze inquiry strategies; and student perceptions were gathered through surveys and focus groups.

Results suggest that students in the LLM-supported group showed higher engagement and more productive inquiry patterns than the comparison conditions, including a game-only condition and a passive-media classroom baseline, highlighting the added value of interactive AI support over less interactive instructional formats. Dialogue transcript analysis further showed that CAP often elicited explanation, justification, clarification requests, and self-correction, suggesting that the agent functioned as a conversational peer that supported reflective reasoning during play. Student surveys and focus group feedback reflected a positive reception of the AI peer, with many reporting increased engagement and confidence during the learning process. Because the pre/post assessments were collected anonymously and the posttest completion rate was lower than the pretest completion rate, the quantitative findings are best interpreted as group-level trends rather than matched individual growth trajectories. Implementation insights highlight device and network constraints encountered when deploying a local LLM in the classroom, which informed iterative UI refinements and technical optimizations for smoother integration of the peer agent. These findings underscore the promise of AI-driven, active learning interventions for K–12 STEM education and illustrate the feasibility of integrating advanced AI tools in real classrooms under typical resource constraints.

Introduction

K–12 STEM education often struggles to keep students engaged and help them truly understand complex science. Traditional teaching forces students to be passive[1], which limits their motivation and problem-solving skills. This issue is even worse in under-resourced schools.

To address these gaps, SPARC (Systematic Problem-Solving & Algorithmic Reasoning for Children) is designed as a story-based, multiplayer mobile game to teach science. By working together to solve problems inside a virtual human body, students learn computational thinking and feel real and engaged. Our goals are simple: help students understand science more deeply, solve problems better, and get excited about future STEM careers.

Early pilots showed that SPARC successfully sparked student interest[2]. However, we also noticed that students often get stuck with complex problems and need more support. To address this, we introduced CAP ('Capsule') in the third year of the project. CAP is an AI-powered in-game companion designed to act like a helpful teammate. Instead of giving answers directly, CAP guides students with hints, questions, and prompts to help them think deeper. This study analyzes how CAP influences student engagement and learning compared to playing without AI support.

Background

Game-Based Learning and Peer Interaction

Research shows that gamification is great for motivation because it taps into curiosity and provides instant feedback[3]. SPARC uses this approach with a story inspired by The Magic School Bus, where students explore the blood circulation 'journey' of the human body.

In addition, we go beyond just fun gameplay. Social learning theory suggests that students learn best when they explain their ideas to others. That's why we designed CAP as a 'peer' rather than a teacher. By chatting with a virtual classmate, students are encouraged to explain their reasoning in their own words, which helps them understand the material more deeply.

AI-Powered Conversational Agents

New AI tools, especially Large Language Models (LLMs), allow us to support students in ways that weren't possible before. In SPARC, we use an AI agent to combine the fun of gaming with personalized help.

Unlike old-school video game characters with scripted lines, CAP talks naturally to guide the student. From our setting, CAP is designed to ask 'why?' and 'how?' like a curious classmate. This triggers the 'protégé effect'[4]: by explaining concepts to CAP, the student becomes the teacher. This role reversal forces them to clarify their own thinking. Over time, CAP acts like a supportive friend, keeping students engaged through conversation and encouragement.

Relevance to Engineering Education

SPARC not only aligns with NJ science curriculum standards, but we also design our stories as engineering challenges. Students investigate scenarios like disease and cell behavior using an engineering mindset, as shown in Fig.1.

The AI agent can also connect subjects like biology to computer science. It might explain basic AI concepts by comparing how the human brain processes signals to how computers classify data. This approach helps students learn science while getting an early look at modern tech. Our goal is

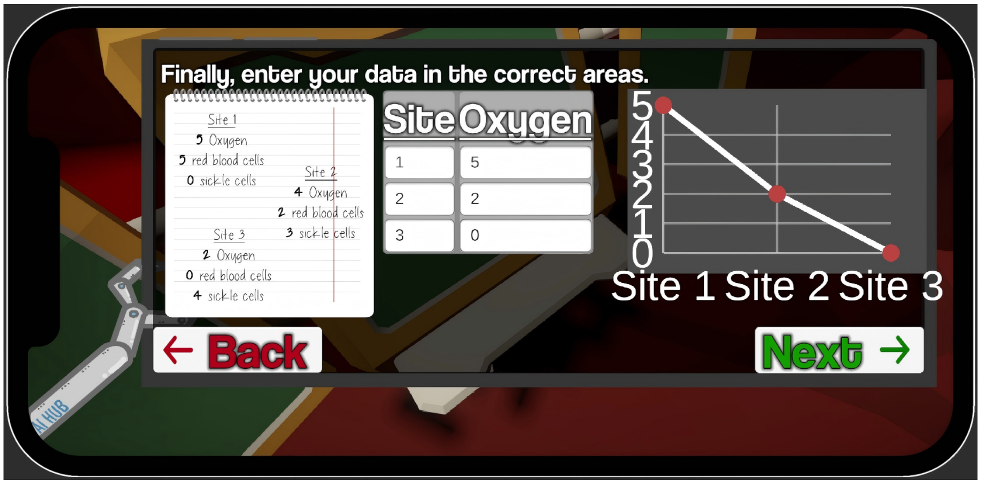

Figure 1: Player solves problems about sickle cells.

to show how AI tools can make real-world STEM problems exciting and accessible for K12 students.

Methodology

Learning Platform and AI Peer Agent

SPARC is a game-based platform build with Unity, where allows students explore the human body to solve science and engineering problems.

The study features CAP, an AI agent designed to act like a helpful teammate rather than a strict teacher. As shown in Fig.2, it appears in levels like 'Meeting Cells' and 'Sickle Cell Detective' to offer help naturally during gameplay. CAP has three main jobs

- Ask questions: It uses prompts like 'Why do you think so?' to encourage students to explain their reasoning.
- Give hints: It offers guidance when students get stuck, preventing them from guessing blindly.
- Keep motivation high: It provides friendly support or triggers an interesting story to keep students engaged with the task.

Pilot Setting and Participants

This year's pilot study took place in Spring 2025, involving 60 K12 students from an urban district. Teachers integrated the activities into standard science lessons using a shared guide. Students used mobile devices, occasionally working in pairs due to limited hardware.

Research Design and Conditions

We used a three-group design to measure the specific effect of the AI peer:

Figure 2: Cap discusses with the player.

- CAP Condition: Students played SPARC with the AI agent, receiving real-time questions and hints.
- Game-only Condition: Students played the same game with the AI disabled. They experienced the full story and gameplay, but without the conversational support.
- Passive-media Condition: Students learned through video-based modules, serving as a baseline for passive-game learning.

All conditions shared the same learning goals and assessments, and the sessions lasted for the same amount of time.

Data Sources and Measures

We used a mixed-methods approach to evaluate both learning outcomes and the learning process. Data sources included:

1. Concept Assessment (Pre/Post): We designed a content test covering key concepts, such as blood cell functions and disease mechanisms. Students took a pre-test before the module and a post-test after completion. The pre-test was administered anonymously at the classroom level, which supported broad participation but prevented individual pre-post matching. In contrast, the post-test was embedded in the gameplay sequence; as a result, some students did not fully complete the game, while others submitted blank or otherwise invalid responses. We therefore retained only valid post-test responses for analysis and interpreted the quantitative findings as condition-level patterns rather than student-level growth trajectories or matched effect sizes.
2. In-game Behavioral Logs: The system automatically tracked gameplay metrics, including task completion, time-on-task, interaction frequency, error rates, and hint usage. We used this data to analyze engagement levels and compare problem-solving strategies between the CAP and non-CAP conditions.

3. Student–CAP Dialogue Transcripts: For the CAP condition, we recorded and coded all student–agent interactions. The coding analysis focused on conceptual accuracy, depth of explanation, use of scientific vocabulary, and reasoning coherence (linking evidence to claims). To align the qualitative analysis with our theoretical framing, we interpreted explanation-rich and teaching-oriented turns as evidence consistent with the protégé effect, because students were positioned to explain ideas to a peer-like partner. We also interpreted clarification, uptake, and revision during the dialogue as evidence consistent with social learning, because meaning was negotiated through interaction rather than delivered as one-way instruction. In addition, we tracked process indicators, such as when students revealed misconceptions or corrected their own mistakes, to understand how CAP's prompts influenced their inquiry.
4. Surveys and Focus Groups: After the module, students completed a survey on their motivation, perceived learning, confidence, and the platform's usability. To add context to these quantitative results, we conducted small focus groups to gather detailed feedback on the gameplay and AI interactions. We also collected teacher observations to assess classroom feasibility.

Technical Deployment and Constraints

To ensure data privacy and reduce latency, we hosted a lightweight LLM on a local school server rather than using a cloud-based API[5]. Student devices accessed the agent (CAP) directly through the local network. This setup helped us navigate typical classroom challenges, such as older hardware and unstable Wi-Fi. To minimize disruptions, we optimized the platform for web access and faster loading. We also added visual ”thinking” indicators to manage student expectations during delays and built a fallback mode that allowed gameplay to continue smoothly even if the AI connection dropped or lagged.

Results

Learning Outcomes – Conceptual Understanding

As shown in Table 1, the descriptive pre/post patterns suggest more favorable learning outcomes in the AI-supported condition than in the game-only and passive-media comparison conditions. Because the assessments were anonymous and unmatched, however, these results are interpreted cautiously as group-level trends rather than student-level gains.

Baseline (Pre-test): Across module types, pre-test accuracy ranged from 39.2% to 50.0% (overall: 46.6%). Specifically, both agent-based module categories started at 39.2%, while passive video instruction modules started at 50.0% and interactive gameplay modules started at 47.3%.

Post-test Results: Post-test accuracy showed divergence across module types (overall: 49.7%). The agent-based module with LLM inquiry increased from 39.2% to 45.7% (+6.5 percentage points). In contrast, the agent-based module without LLM inquiry decreased from 39.2% to 24.4% (-14.8 points), and the passive video instruction modules decreased from 50.0% to 28.3% (-21.7 points). Interactive gameplay modules increased from 47.3% to 59.2% (+11.9 points).

Table 1: Pre- and Post-test Accuracy (%) by Module Type

Module Type	Pre-test (%)	Post-test (%)
Agent-based Module(LLM Inquiry)	39.2	45.7
Agent-based Module(No LLM Inquiry)	39.2	24.4
Passive Video Instruction Modules	50.0	28.3
Interactive Gameplay Modules	47.3	59.2
Overall	46.6	49.7

Note: Sixty students participated in the pilot overall. Valid concept-test responses were N = 60 for the pre-test and N = 44 for the post-test. The smaller post-test sample reflects the fact that the post-test was embedded in gameplay, so some students did not finish the game or submitted blank/incomplete responses that were excluded from analysis.

Data Note: Pre- and post-test sample sizes differed (N = 60 pre-test; N = 44 post-test; Table 1), so these comparisons are descriptive of the aggregated results by module type. They should not be interpreted as matched student-level growth trajectories.

In-Game Inquiry Behavior and Engagement

The gameplay logs revealed two very different playing styles.

- The ”Guessing” Group (No AI): Without help, students often resorted to trial-and-error. In the Sickle Cell module, when they didn’t know the answer, they would often click through options rapidly just to move forward. We call this ”spam-clicking” or blind guessing.
- The ”Thinking” Group (With CAP): Students with CAP played differently. Instead of rushing, they paused to chat with the agent. CAP’s questions forced them to stop and think before acting.

Students with CAP were generally more engaged and worked more effectively. They sometimes spent a bit more time on a single step, but that time was usually spent reading prompts and explaining their thinking, not being stuck. Because CAP helped them avoid wrong paths, many students completed the full mission faster overall. They also explored more optional materials and used the chat voluntarily (often 2–3 short back-and-forth exchanges per challenge) to explain their reasoning, which seemed to encourage “thinking out loud” and deeper processing. By contrast, students in the game-only condition enjoyed the game but had less support when they got stuck; they were more likely to pause, give up briefly, or show frustration. The passive video condition showed the lowest engagement, since there was no interactive element to keep students involved after the video ended.

Dialogue Quality Analysis

We analyzed over 150 student-agent conversations to examine how CAP shaped student reasoning during play and whether the dialogue patterns aligned with our theoretical framing. Overall, the transcripts suggest that CAP did more than provide hints: it prompted students to explain ideas, externalize uncertainty, revise partial understandings, and occasionally teach the idea back to the agent.

Explanation and Teaching Moves: In 78% of the exchanges, students did not stop at one-word answers; they provided a justification when CAP asked ”Why?” or ”How?”. Many of these turns resembled teaching moves rather than simple responses. For example, students explained that red blood cells ”need more room to carry oxygen,” that the nucleus was ”an extra burden,” or that the cell had ”traded” its nucleus for oxygen-carrying space. We interpret these explanation-rich turns as evidence consistent with the protégé effect, because CAP positioned students to articulate knowledge to a peer-like listener.

Mechanistic Reasoning in Their Own Words: Students frequently used newly encountered vocabulary in causal explanations rather than isolated recall. For example, one student explained: ”Because its shape makes it get stuck and not carry oxygen well, so the body’s tissues don’t get enough oxygen.” Other students described red blood cells as needing more ”space” for oxygen, being more ”flexible,” or carrying more hemoglobin after giving up the nucleus. These responses suggest that students were not simply repeating definitions; they were linking structure, function, and consequence in their own language.

Visible Uncertainty, Misconceptions, and Repair: The transcripts also made partial understandings visible. Some students openly stated, ”I dont know i just think,” ”I dont understand,” or asked CAP, ”Do you know more?” Others voiced misconceptions, such as believing that red blood cells must keep a nucleus in order to carry oxygen. These moments were analytically valuable because they surfaced reasoning that would otherwise remain hidden in a multiple-choice format. In several exchanges, CAP’s follow-up questions then prompted students to refine or correct their ideas. For example, when a student claimed ”white blood cells carry oxygen,” CAP’s response led the student to revise the explanation: ”Oh wait, no, red blood cells do that; white cells fight infection.” We interpret these clarification and repair sequences as evidence consistent with social learning, because understanding was negotiated through dialogue rather than delivered as one-way instruction.

Taken together, the transcript findings help explain why the CAP condition differed from the other conditions. The value of the AI peer was not only that it increased interaction frequency, but that it shifted the quality of interaction toward explanation, clarification, and self-repair.

Student-Reported Engagement and Perceptions

As shown in table 2, the feedback confirmed what we saw in the logs: students loved the AI-supported game.

As shown in Table 2, the post-module survey (N = 54; 29 boys and 25 girls) indicated generally positive student perceptions of the 2025 SPARC experience. On a 5-point Likert scale, students reported moderate-to-high overall learning from the game (M = 3.81). For perceived improvements in problem-solving skills, “somewhat” was the most common response (51.9%), followed by “a lot” (20.4%). Students also reported strong enjoyment of gameplay: 42.3% rated the experience as enjoyable (4) and 23.1% as highly enjoyable (5), while no students selected “not enjoyable” (1). When comparing SPARC to traditional assignments, most responses fell on the positive side of the scale (44.4% selected 4 and 16.7% selected 5), whereas only 5.6%

Table 2: Summary of Student Surveys of SPARC Game Experience

Survey Item	1	2	3	4	5
Overall learning from SPARC	0.0%	1.9%	33.3%	46.3%	18.5%
Improved problem-solving skills	5.6%	14.8%	51.9%	20.4%	7.4%
Overall enjoyment of the gameplay experience	0.0%	3.8%	30.8%	42.3%	23.1%
SPARC compared to traditional assignments	1.9%	3.7%	33.3%	44.4%	16.7%

Note: Response options on a 5-point Likert scale. For Items 1–3, 1 = Not at all, 5 = A significant amount. For Item 4, 1 = Much worse than traditional assignments, 5 = Much better. N = 54.

selected the negative side of the scale (1 or 2 combined). Although this item is perceptual rather than a direct achievement comparison, it provides a useful classroom-relevant baseline showing that students generally experienced SPARC as more beneficial than typical assignments. Taken together, these results align with our behavioral logs by indicating sustained engagement and favorable attitudes toward learning through the SPARC module.

Fig.3 summarizes what students reported learning while playing the game (N=54; multiple selections allowed). The most frequently selected learning outcome was understanding how red blood cells work (92.6%, 50/54), followed by how the heart works (79.6%, 43/54) and how the body works in cold conditions (59.3%, 32/54). Over half of students also reported learning how movement affects the heart (53.7%, 29/54) and how mass can affect speed (51.9%, 28/54), and many selected friction affecting speed (38.9%, 21/54) and sickle cell anemia (33.3%, 18/54). Notably, no students selected “Nothing” (0%), and only 1.9% (1/54) indicated they learned nothing new, suggesting that most students perceived meaningful content learning from the module.

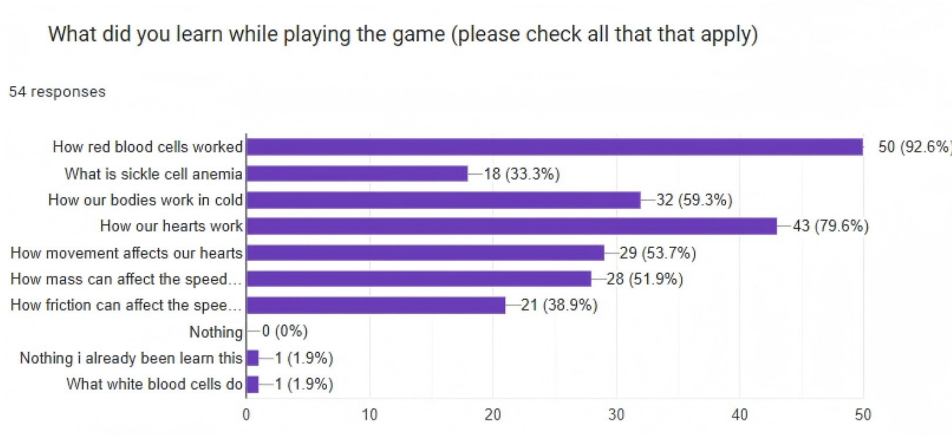

Figure 3: Students’ survey outcome after gameplay.

Conclusion

Our findings highlight a key opportunity: well-designed AI companions can provide meaningful, scalable support in K–12 settings, even under typical school constraints. Across the present comparison conditions, the CAP-supported module showed more favorable engagement and inquiry patterns than the game-only and passive-media baselines, while the dialogue evidence suggests that these benefits may arise from peer-like prompting that elicits explanation, clarification, and self-correction. At the same time, the anonymous and incomplete pre/post dataset limits claims about matched individual growth. Moving forward, we aim to expand SPARC's reach, refine the technology, and incorporate privacy-preserving identifiers so that future studies can connect learning trajectories more directly with gameplay and dialogue evidence. As we showed here, bringing AI into real classrooms is a promising way to support the problem-solvers of tomorrow.

References

[1] M. Fitzmaurice, "Considering teaching in higher education as a practice," *Teaching in Higher Education*, vol. 15, no. 1, pp. 45–55, 2010. [Online]. Available: https://doi.org/10.1080/13562510903487941

[2] C. Zhu and Y. Tang, "Gamified learning: Integrating the human body with systematic and computational thinking for k-12 education," in *2023 Fall Mid Atlantic Conference: Meeting our students where they are and getting them where they need to be*, no. 10.18260/1-2–45121. Ewing, New Jersey: ASEE Conferences, October 2023, https://peer.asee.org/45121.

[3] T. Laine and R. Lindberg, "Designing engaging games for education: A systematic literature review on game motivators and design principles," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 804–821, Oct. 2020.

[4] C. C. Chase, D. B. Chin, M. A. Oppezzo, and D. L. Schwartz, "Teachable agents and the protege effect: Increasing the effort towards learning," *Journal of Science Education and Technology*, vol. 18, no. 4, pp. 334–352, Aug. 2009.

[5] C. Zhu, L. Cui, Y. Tang, and J. Wang, "The evolution and future perspectives of artificial intelligence-generated content," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 56, no. 1, pp. 546–564, 2026.