Design and Feasibility of an LLM-Powered Humanoid Robot as a Reading Companion to Support Children's AI Literacy

Feiwen Xiao[1(✉)], Zhaohui Li[2], Shiyan Jiang[1], and Jinjun Xiong[3]

[1] University of Pennsylvania, Philadelphia, PA 19104, USA
{feiwenx,jiang33}@upenn.edu
[2] University of Buffalo, Buffalo, NY 14068, USA
zli253@buffalo.edu
[3] The University of Texas at San Antonio, San Antonio, TX 78249, USA
jinjun@utsa.edu

Abstract. Artificial intelligence (AI) is increasingly prevalent in children's everyday experiences, creating a growing need for developmentally appropriate AI literacy learning opportunities. Recent advances in large language models and social robotics enable new possibilities for embodied, interactive learning. We present an LLM-powered humanoid social robot as a reading companion for AI literacy books for children aged 58, integrating speech, vision, language generation, and gesture to support multimodal interaction. We implement and compare three interaction policies (i.e., heuristic, dynamic, and guided) representing different levels of control in LLM-mediated dialogue. Through controlled interaction evaluation, we demonstrate the feasibility of deploying this system on real hardware and identify a key insight: a structured interaction framework with bounded LLM support yields more robust and pedagogically coherent interactions than fully scripted or open-ended approaches. These findings highlight design directions for safe and effective embodied AI learning companions for young children.

Keywords: Social Robot · Reading Companion · AI Literacy

1 Introduction

With AI technologies becoming commonplace in children's daily lives, there is a need for developmentally appropriate AI education [1]. However, foundational concepts such as how AI systems perceive the world remain difficult for young learners to grasp through existing approaches, which often rely on simplified explanations or screen-based interactions. Recent advances in large language models (LLMs) and social robotics enable new forms of embodied, interactive learning. While robots can support engagement and LLMs enable flexible dialogue [2], integrating the two raises challenges for pedagogical coherence: scripted

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
E. G. Blanchard et al. (Eds.): AIED 2026, CCIS 3031, pp. 498–504, 2026.
https://doi.org/10.1007/978-3-032-29788-4_71

systems lack adaptability, whereas open-ended systems may lead to inconsistent and off-topic interactions.

To address these challenges, we present an embodied, LLM-powered humanoid robot as a co-reading companion for children's AI literacy books. The system integrates speech, vision, language generation, and gesture, and is structured through a four-stage interaction framework that coordinates dialogue and motion to scaffold children's understanding from observation to explanation. We further examine how different levels of control in LLM-mediated interaction affect learning design through comparing three dialogue policies, including heuristic, dynamic, and guided. This work demonstrates the feasibility of an LLM-powered embodied learning system for children's AI literacy and provides insights into balancing flexibility and pedagogical coherence in robot-assisted learning.

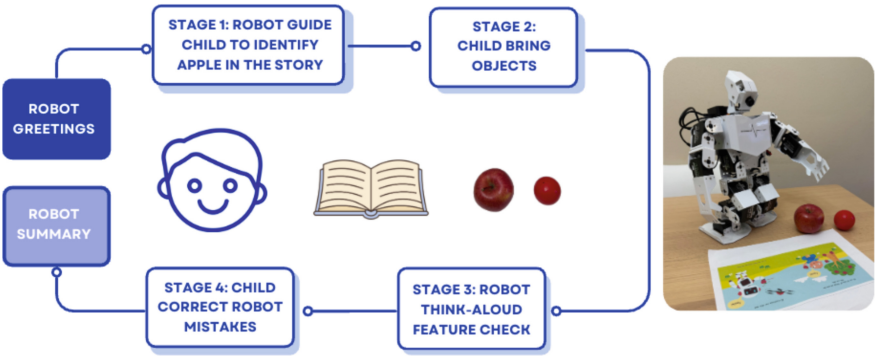


Fig. 1. Overview of the robot-assisted AI literacy activity.

2 Related Work

Children's AI literacy. AI literacy is increasingly important, encompassing not only core concepts [9] but also critical engagement with AI systems [8]. Children's storybooks are well suited for young children's learning of AI concepts, as they introduce abstract ideas through familiar narrative contexts [1]. However, storybooks alone are often insufficient for deep learning; prior work emphasizes the need for guided interaction to support conceptual understanding [5]. In practice, caregivers may lack the knowledge or strategies to effectively scaffold children's reasoning about AI, limiting the potential of these materials.

Social Robots as Children's Reading Companions. Social robots have been explored as reading companions that support children's engagement during shared reading [2,6,7]. Most prior systems rely on short text-bound questionanswer exchanges grounded in dialogic reading practices [5,10], which support basic comprehension but limit extended discussion beyond the story. A key

advantage of social robots is their physical embodiment, enabling communicative channels such as gaze, gesture, and movement. These modalities are especially important in early learning, where embodied action supports comprehension [4]. However, existing robotic reading companions have focused primarily on spoken dialogue, with little use of semantically meaningful gestures and limited support for sustained, exploratory conversation connected to children's lived or embodied experiences [6]. These gaps motivate the design of AI-driven embodied reading companions that integrate verbal and nonverbal scaffolding.

3 System Design

3.1 Learning Context and Design Principle

In this feasibility study, we designed the social robot reading companion around a two-page spread from *Artificial Intelligence for Kids*[1]. We selected this spread because it introduces the core AI concept of *features* and uses visually rich illustrations that are not easily understood without explanation, making it a strong context for robotic scaffolding. The spread contrasts human and AI perception: a child immediately recognizes an apple, while an AI robot identifies it through labeled features such as color, shape, leaf, and stem. This contrast makes visible a key idea in AI learning: humans recognize holistically, whereas AI relies on feature checking. The interaction unfolds in four stages shown in Fig. 1. The robot first guide the child to identify the apple in the book and retrieve a real one. It then performs a think-aloud feature check on the apple, followed by the same process on a red ball, incorrectly concluding, "Apple—am I right?" This prompt the child to explain the mismatch. Finally, the robot highlight that both children and AI recognize objects, but in different ways.

To make the interaction engaging and pedagogically meaningful, we coordinated robot motion with dialogue, drawing on prior work showing that legible robot behavior supports intent inference [3], and that embodied action is important in early learning [4]. Three principles guided the motion design. First, motion was used to direct and redirect attention, waving and open-arm gestures during greeting, pointing and nodding when drawing the child's focus to a visible feature. Second, gesture types were mapped to dialogue functions: nodding for confirmation, leaning or lowering the head to signal thinking, and pointing to highlight specific object features such as color, shape, or stem. Third, motion was constrained to upper-body and head-based actions throughout the lesson, avoiding larger full-body movements that introduced instability or visual noise.

3.2 Hardware Setup

We selected the TonyPi humanoid robot for its affordability, programmability, and compact form factor. It supports articulated motion, a controllable head, an onboard camera, and a Python-based integration of speech and action, making

[1] https://www.amazon.com/dp/B07MJZQ56X.

it suitable for a co-reading prototype. Although other humanoid robots such as the Unitree G1 and NAO 6 offer more advanced hardware, their cost and size make them impractical for a home reading context. TonyPi provided the most cost-effective platform for an embodied AI literacy companion. The hardware setup comprised the robot body, camera, microphone, and speaker. The built-in microphone was replaced with a wireless *DJI Mic Mini* to improve speech capture, as the onboard mic was affected by the CPU fan noise.

3.3 Design LLM-Powered Robot

The robot was connected to the LLM through a structured multimodal pipeline that combines speech recognition, camera-based visual understanding, language generation, text-to-speech, and robot motion. For speech input, child speech was captured through an external microphone and transcribed with OpenAI's real-time speech recognition API, using the gpt-4o-transcribe model. For visual input, the robot camera captured the book page and physical objects placed in front of the robot. These images were sent to a GPT-4o visual model API for visual analysis, including page understanding and object checking, such as determining whether an apple was present and which visible features (e.g., red, round, stem) could be detected. For dialogue generation, the system used an LLM API to produce child-facing responses from three sources of context: the current reading material state, the child's transcribed utterance, and the latest visual analysis. The lesson controller managed the overall reading sequence, including greeting, page discussion, apple request, apple checking, red-ball trick case, explanation, and closing summary. This ensured that the educational objective remained stable even when language generation was model-based. For output, the generated text is given to a text-to-speech API, and the spoken response was coordinated with robot gesture and head movement. We use the `gpt-4o-transcribe` for ASR and `gpt-4o-mini-tts` with the `nova` voice for TTS.

Table 1. Comparison of three dialogue policies. Latency is reported as the average time-to-first-byte across runs.

Method	Dialogue Control	Response Type	Latency (s)
Heuristic	Rule-based	Predefined	1.17
Guided	Rule-based	LLM-generated	1.07
Dynamic	Open-ended	LLM-generated	1.00

4 Evaluation

4.1 Evaluation Setup

The experiment compared three policies implemented on the same robot method: *Heuristic*, *Guided*, and *Dynamic* as shown in Table 1. The heuristic method relied

on a scripted structure and responses; the dynamic method used a free structure and generated responses with the LLM; and the guided method followed a scripted structure while generating responses using LLM outputs. This design enabled us to compare different levels of conversational flexibility.

To make the comparison reproducible, we evaluated all three tutors under the same three response profiles: *Correct*, *Off-topic*, and *Incorrect*, with all responses simulated by a researcher acting as the child.

4.2 Evaluation Metrics

An expert in AI education evaluated the system based on the following dimensions on a 1–5 scale. *Pedagogical coherence* assessed whether the four-stage interaction logically progresses toward the learning objective. *Dialogue compliance* examined whether robot speech followed design constraints (stepwise prompts, no self-answering or over-praise, within a 12-word limit). *Gesture–speech coherence* evaluated whether robot motion supported and reinforced spoken content. *Scaffolding logic* assessed whether responses appropriately adjusted to engagement levels, progressing from open questions to yes/no prompts to physical comparison without skipping or over-scaffolding.

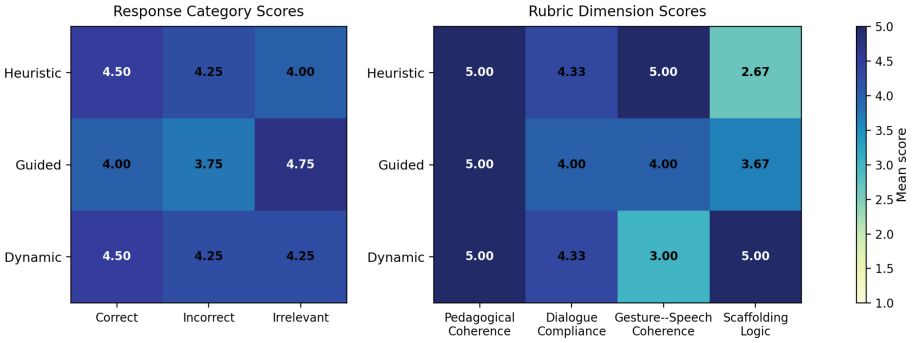


Fig. 2. Heatmaps of average expert ratings for different policies by response category (left) and rubric dimension (right). Darker cells indicate higher scores.

4.3 Results and Findings

Figure 2 shows that across response categories, heuristic and dynamic policies performed similarly in *Correct* (4.50) and *Incorrect* (4.25) conditions, both outperforming the guided policy. In contrast, the guided policy achieved the highest score in the *Irrelevant* condition (4.75), indicating better handling of off-task inputs. Across rubric dimensions, all policies achieved perfect *Pedagogical Coherence* (5.00). Heuristic and dynamic policies scored slightly higher in *Dialogue Compliance* (4.33) than guided (4.00). The heuristic policy achieved the

highest *GestureSpeech Coherence* (5.00), while the dynamic policy performed best in *Scaffolding Logic* (5.00), outperforming guided and heuristic. Overall, the results suggest trade-offs: heuristic supports consistency, dynamic supports adaptive scaffolding, and guided improves handling of irrelevant responses.

Table 1 reports latency as *time to first byte*, which measures the interval from the end of the child's last spoken input to the moment the robot begins receiving the audio stream for its next utterance. The observed delay (1 s) is acceptable for child-robot interaction and does not disrupt conversational flow.

The results highlight a trade-off between consistency and adaptability across interaction policies, suggesting that future work should explore hybrid approaches that balance structured guidance with flexible LLM-driven dialogue while improving gesturespeech alignment and validating effectiveness with real children.

5 Conclusion and Future Work

The work demonstrates the feasibility of an LLM-powered humanoid robot as a co-reading companion for early AI literacy, highlighting the importance of balancing structure and flexibility in embodied, multimodal interaction design. Our findings suggest that structured frameworks with LLM support can enable more pedagogically coherent interactions than scripted or open-ended approaches.

Future work will extend this feasibility study in three directions. First, we will conduct in situ studies with K2 learners to examine engagement, interaction patterns, and learning outcomes when using the robot as a co-reading companion for AI literacy books. Second, we will systematically investigate how different combinations of gestures, embodiment cues, and conversational strategies influence children's understanding of AI and other complex computing concepts, including how to balance guidance and open-ended exploration. Third, we will refine the system architecture to improve robustness, safety, and personalization, including adaptive scaffolding based on children's responses and more reliable perception in real-world classroom and home environments. These efforts aim to advance the design of effective, trustworthy, and developmentally appropriate embodied AI learning companions for young children.

References

1. Arn, L., Huang, E.M.: "robots can do disgusting things, but also good things": fostering children's understanding of ai through storytelling. ACM Trans. Comput. Educ. **24**(3), 1–55 (2024)
2. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: a review. Sci. Robot. **3**(21), eaat5954 (2018)
3. Dragan, A., Lee, K., Srinivasa, S.: Legibility and predictability of robot motion. In: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 301–308 (2013)
4. Goldin-Meadow, S.: How gesture works to change our minds. Trends Neurosci. Educ. **3**(1), 4–6 (2014)

5. Hargrave, A.C., Sénéchal, M.: A book reading intervention with preschool children who have limited vocabularies: the benefits of regular reading and dialogic reading. Early Childhood Res. Q. **15**(1), 75–90 (2000)
6. Kory-Westlund, J.M., Breazeal, C.: Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. Front. Robot. AI **6**, 54 (2019)
7. Leite, I., Martinho, C., Paiva, A.: Social robots for long-term interaction: a survey. Int. J. Soc. Robot. **5**(2), 291–308 (2013)
8. Long, D., Magerko, B.: What is ai literacy? competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2020)
9. Touretzky, D., Gardner-McCune, C., Martin, F., Seehorn, D.: Envisioning AI for k-12: What should every child know about ai? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9795–9799 (2019)
10. Whitehurst, G.J., et al.: Accelerating language development through picture book reading. Dev. Psychol. **24**(4), 552 (1988)