Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Privacy-Focused AI Avatar for Educational Robotics in the Metaverse

Lovro JAKIC[a*], Mario VUKOJA[a], Antun DROBNJAK[a], Ivan TERZIC[a], & Ivica BOTICKI[a]
[a]*Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*
*lovro.jakic@fer.hr

Abstract: The convergence of immersive environments and educational robotics is transforming STEM learning by enabling students to design and deploy robots across both virtual and physical contexts. This paper presents a privacy-first conversational AI avatar that enables seamless voice interaction with robots in simulation and reality. All processing runs locally, safeguarding student data while ensuring real-time responsiveness. The system integrates Whisper for speech recognition, a lightweight Llama 3.2 model for intent understanding, and Glow-TTS for natural feedback. Robot commands are strictly validated against a JSON schema and executed via ROS 2, allowing seamless transfer between NVIDIA Isaac Sim and a Jetson Orin Nano–powered robot. A prototype workshop confirmed smooth interaction, safe command execution, and high engagement, highlighting potential for immersive, privacy-preserving robotics education.

Keywords: Conversational AI avatar, metaverse, educational robotics, privacy

1. Introduction

Immersive technologies such as the metaverse are opening new opportunities for STEM education by offering new safe and collaborative learning spaces (Kye et al., 2021). Within these environments, educational robotics provides a hands-on platform for teaching computational thinking and engineering principles (Benitti, 2012). Combining the two enables sim-to-real learning, where students design and test robots virtually before deploying them onto physical hardware.

However, integrating conversational AI into such systems introduces challenges. Concerns include student data privacy, reliance on cloud services, and the risk of AI-generated errors or unsafe instructions (Islam & Wang, 2025). Minimizing latency is also critical, since slow responses disrupt natural human–robot interaction (Pollini et al., 2025).

This paper introduces a privacy-first conversational AI avatar that operates fully on local hardware. The system transcribes speech, interprets intent with a lightweight language model, and produces validated JSON commands executed by simulated or physical robots through Robot Operating System 2 (ROS 2). By unifying speech recognition, safe intent parsing, and seamless sim-to-real execution, the approach supports accessible, near real-time interaction while protecting learner privacy.

2. System Design and Architecture

2.1 Overall architecture

The system connects a Unity WebGL avatar front end with a Flask-based back end that integrates speech recognition, language understanding, and command validation. Learners issue voice commands, which are transcribed, interpreted by a lightweight LLM, and converted into standardized JSON actions. These commands control a robot in either simulation or reality through ROS 2. Using NVIDIA Isaac Sim for virtual testing and a Jetson Orin Nano for physical deployment, the setup enables seamless switching between environments with only an endpoint change. The avatar thus provides a natural-language interface that translates student speech into safe, executable robot actions across both contexts.

2.2 Speech Recognition Module

Speech-to-text is handled by Whisper, chosen for its robustness to diverse accents and noisy classrooms (Radford et al., 2022). Running locally, it delivers accurate transcripts with low latency, ensuring reliable input for the reasoning module without dependence on external services.

2.3 Understanding via Lightweight LLM

After transcription, the natural-language command is processed by a local LLM to infer the appropriate robot action. Prior research in human–robot interaction emphasizes that minimizing command processing latency is critical for maintaining smooth, near real-time usability (Pollini et al., 2025). To balance capability with responsiveness, our system employs Llama 3.2 3B, a compact open-source model that runs efficiently on local hardware while delivering fast response times (Meta, n.d.). Running the model locally also preserves data privacy which is critical for classroom deployment.

Despite its smaller size and limited context window compared to larger models, Llama 3.2 3B provides sufficient language understanding to map spoken instructions into structured commands defined by a fixed JSON schema. Carefully designed prompts and few-shot examples guide the model in producing valid outputs. For example, when a learner says “Go forward a little and then turn left,” the model may output a JSON sequence such as: [{"action": “MoveForward”, “duration”: 1.0, “speed”: 50.0}, {"action": “TurnLeft”, “duration”: 1.0, “speed”: 50.0},].

2.4 JSON Command Schema

Robot actions are represented using a fixed JSON schema that constrains the LLM’s output to a predefined set of safe commands. This design simplifies execution and improves reliability by ensuring that only interpretable, executable instructions are passed to the robot. The schema currently supports common navigation and task-oriented actions, such as *MoveForward, MoveBackward, TurnLeft, TurnRight, FollowLine,* and *DetectObject* each defined with parameters (e.g., duration, speed or object name). Developed in consultation with robotics instructors, the command set reflects typical introductory activities like maze navigation and line following. Strict schema validation adds robustness, as invalid outputs are either rejected or automatically corrected by a lightweight parser before being executed.

2.5 Text-to-Speech Module

For spoken feedback, the system uses Glow-TTS, a flow-based neural model that generates natural speech with low latency (Kim et al., 2020). Fast synthesis supports smooth interaction and accommodates longer sentences. In practice, the avatar uses Glow-TTS to confirm actions (e.g., *“Turning left”*) or request clarification when input is unclear (e.g., *“Sorry, can you repeat the command?”*).

2.6 Integration with Simulation and Physical Robots

A central feature of the system is seamless switching between simulation and real robots. Both are accessed through a unified API in the Flask back end, which forwards JSON commands to ROS 2. In virtual mode, commands are executed in NVIDIA Isaac Sim, where the robot is modeled with its actual 3D design. In physical mode, the same commands are sent via Flask API to a ROS 2 network on a Jetson Orin Nano–powered robot. The Jetson provides sufficient edge AI performance for responsive execution. This architecture allows students to prototype tasks safely in simulation and transfer them directly to physical robots without reconfiguration (Boras et al., 2025).

3. Prototype Demonstration and Discussion

A prototype was demonstrated at a workshop on educational technology, where professors controlled a robot through the Unity WebGL avatar using voice commands (Zunic, 2025). The interaction proved relatively smooth, as the local LLM and optimized processing pipeline enabled low-latency responses, making communication with the avatar feel natural and engaging. The system handled phrasing variations (e.g., “go straight” vs. “move forward”) and proved adequate for basic command understanding.

Most errors stemmed from uncommon vocabulary or long, multi-step instructions. Due to its limited context window, the model also struggled with dialogue memory (e.g., “do the last step again”). Still, the lightweight design ensured responsiveness without relying on external services, a trade-off judged acceptable at this stage. Future work may include domain-specific fine-tuning or hybrid setups that combine local speed with larger models for complex reasoning.

4. Conclusion

In summary, the proposed conversational AI avatar system demonstrates a practical, privacy-preserving approach to integrating voice-based interaction with educational robotics in the metaverse. By combining local speech recognition, lightweight language understanding, and a strict JSON schema for safe execution, the system delivers responsive and reliable robot control in both simulated and physical environments. While current limitations include handling extended dialogue and complex instructions, the prototype highlights the potential of this architecture to enhance STEM learning through seamless, immersive, and safe human–robot interaction.

Acknowledgements

This research has been funded by the European Union – NextGenerationEU, under the project “MetaRoboLearn - Seamless learning with educational robots through the metaverse” (NPOO.C3.2.R3-I1.04.0194). We would like to thank Microline company for providing us with the physical robot components.

References

Kye, B., Han, N., Kim, E., Park, Y., & Jo, S. (2021). Educational applications of metaverse: Possibilities and limitations. Journal of Educational Evaluation for Health Professions, 18, 32. https://doi.org/10.3352/jeehp.2021.18.32

Benitti, F. B. V. (2012). Exploring the educational potential of robotics in schools: A systematic review. Computers & Education, 58(3), 978–988. https://doi.org/10.1016/j.compedu.2011.10.006

Islam, M. Z., & Wang, G. (2025). Avatars in the educational metaverse. Visual Computing for Industry, Biomedicine, and Art, 8(1), 15. https://doi.org/10.1186/s42492-025-00196-9

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Pollini, D., Guterres, B. V., Guerra, R. S., & Grando, R. B. (2025). Reducing Latency in LLM-Based Natural Language Commands Processing for Robot Navigation (No. arXiv:2506.00075). arXiv. https://doi.org/10.48550/arXiv.2506.00075

Meta (n.d.). Llama 3.2 | Model Cards and Prompt formats. Retrieved August 24, 2025, from https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search (No. arXiv:2005.11129). arXiv. https://doi.org/10.48550/arXiv.2005.11129

Boras, B., Drobnjak, A., Jakic, L., Terzic, I., & Boticki, I. (2025). Designing Architecture and Application Interfaces for Educational Robotics Based on Advanced Hardware Components. *Metaverse and Artificial Companions in Education and Society*, 14.

Zunic, M. (2025, June 18). Održana radionica Učenje računalnog razmišljanja i programiranja kroz obrazovnu robotiku. https://degames.uniri.hr/?p=3035