

EdgeAIED: A Low-Cost Hardware–Software Stack for Local LLMs in Resource-Constrained Classrooms

Anonymous Author(s)

Anonymous Institution

Abstract. Large language models are increasingly used in Artificial Intelligence in Education, but most deployments assume reliable connectivity and cloud inference. We present EdgeAIED, a low-cost hardware–software stack for fully local LLM interaction in resource-constrained classrooms. The system combines an embedded Linux compute module, touchscreen, microphone, and speaker with local speech recognition, quantized LLM inference, and text-to-speech. EdgeAIED treats model size, runtime, and precision as deployment variables rather than fixed implementation choices. The stack therefore provides a common platform for comparing educational LLM configurations by latency, memory, energy use, cost, and response quality. The contribution is not new hardware or a new model, but an integrated reference architecture for studying when local educational LLMs become practical on inexpensive edge hardware. The design operationalizes infrastructure-aware principles from AIED Unplugged at the device level.

Keywords: Artificial Intelligence in Education · AIED Unplugged · Edge AI · Large Language Models · Offline AI

1 Introduction

Educational LLM applications commonly rely on cloud inference and persistent connectivity. This model works well where infrastructure is reliable, but it creates barriers in classrooms with intermittent Internet access, shared devices, recurring service costs, or restrictions on student data.

AIED Unplugged argues that educational AI should conform to existing infrastructure, operate without continuous connectivity, and support shared or mediated access [3]. Offline AI can broaden access to consistent educational support, although its effect on learning disparities requires empirical study. This work is also represented institutionally by the UNESCO UNITWIN Chair on Artificial Intelligence Unplugged in Education at the Federal University of Alagoas and the University of São Paulo [11].

Small quantized LLMs offer a device-level path to these goals. Pedagogical elasticity maps available educational functionality to deployment capacity [9, 10]. Prior systems show that offline execution is feasible, but latency, memory, and

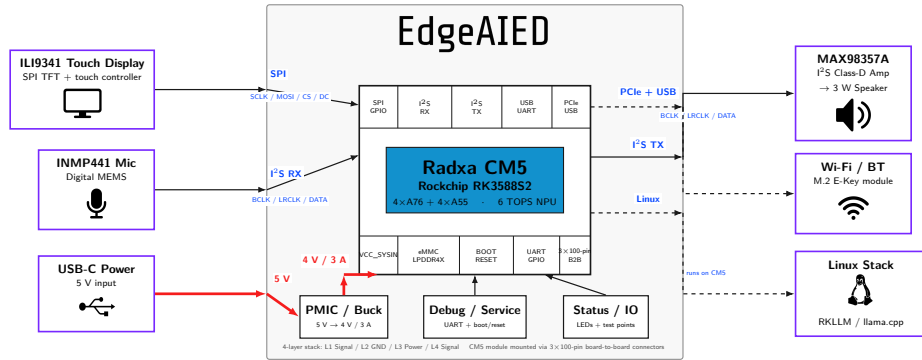


Fig. 1. EdgeAIED hardware architecture. The Radxa CM5 carrier integrates touch, audio, connectivity, power, and local Linux inference.

application footprint remain practical constraints [7, 6]. The engineering question is therefore which model, runtime, and hardware combination supports usable educational interaction.

EdgeAIED addresses this question with a dedicated shared endpoint built around embedded Linux, touch and voice input, and fully local inference. Its fixed interaction layer supports controlled comparison of models, runtimes, and compute paths.

This work makes three contributions. First, it presents a multimodal reference architecture for offline shared-device AIED. Second, it frames deployment as a joint model–runtime–hardware problem. Third, it evaluates systems performance together with educational response quality.

2 Related Work

We focus on prior work that implements and evaluates educational AI under constrained computation.

2.1 Elastic Local LLMs for AIED

Uema et al. define *pedagogical elasticity* as adapting educational functionality to available computation [9]. Their offline-first evaluation shows that lightweight models improve responsiveness and reliability on constrained devices [10]. This makes the model–hardware relationship an AIED design variable. EdgeAIED exposes model size, precision, runtime, and compute path as controlled configurations.

2.2 Efficient Inference on Low-Cost Hardware

Souza et al. optimize handwritten-equation detection for disconnected devices and compare TensorFlow Lite inference across constrained hardware [8]. Santos

et al. deploy TinyLlama, RedPajama, and Qwen2 on a disconnected smartphone, showing that successful execution can still entail high memory demand and long response times [7].

Nguyen and Nguyen benchmark 25 quantized LLMs on Raspberry Pi 4, Raspberry Pi 5, and Orange Pi 5 Pro systems. Processor architecture and runtime materially affect performance and resource use [5]. The LLMPi study further shows that numerical precision changes throughput, energy use, and model quality on Raspberry Pi hardware [1]. These results motivate treating hardware, runtime, and numerical precision as joint experimental variables rather than benchmarking model identity alone.

2.3 Educational LLMs on Edge Devices

Lamouille et al. benchmark local language models for a social-educational robot and compare systems performance with teaching effectiveness [4]. Generic model capability does not necessarily track teaching quality. Gupta and Bhatia evaluate offline educational generation on Raspberry Pi 4B and Jetson Nano platforms, including multilingual output quality [2].

Existing work establishes offline educational inference on low-cost hardware and shows that systems performance and pedagogical quality both matter. Less explored is a shared multimodal endpoint that exposes model, runtime, and hardware choices under one fixed interaction stack. EdgeAIED targets this systems-level gap.

3 System / Approach

EdgeAIED separates the learner-facing interaction from the model, runtime, and compute platform. Figure 1 shows the hardware, while Fig. 2 shows the local inference pipeline. This separation allows model and runtime configurations to change without redesigning the device.

3.1 Hardware Architecture and Cost

The Radxa CM5 SoM is the reference compute module. Its RK3588S2 combines four Cortex-A76 and four Cortex-A55 cores. The module provides a 6-TOPS NPU, 8 GB of RAM, and 64 GB of eMMC storage.

The CM5 mounts to a purpose-built four-layer carrier through three 100-pin board-to-board connectors (Fig. 1). An ILI9341 touchscreen connects over SPI. An INMP441 microphone and MAX98357A amplifier with a 3-W speaker use I²S. USB-C supplies power, optional Wi-Fi/Bluetooth uses M.2 E-key, and USB/UART provide service access.

Using a SoM avoids the unused interfaces of a general-purpose development board and allows the interaction hardware to remain fixed as compute modules change. The carrier acts as a stable deployment interface rather than part of

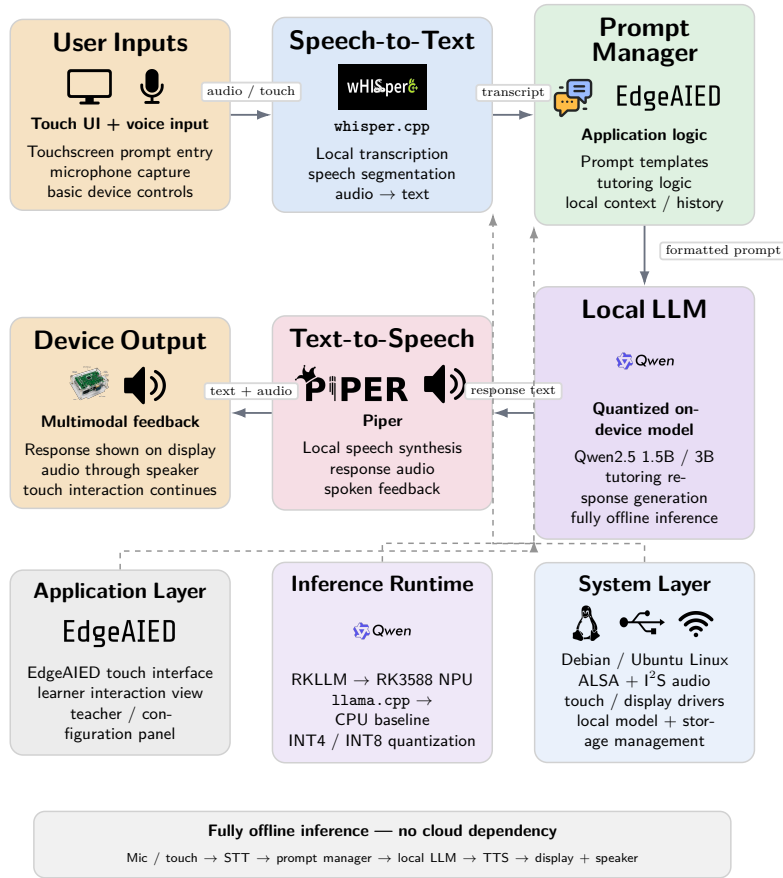


Fig. 2. EdgeAIED software stack. Touch and speech inputs pass through a fully local pipeline for transcription, prompt management, LLM inference, and audio/visual feedback.

a model-specific implementation. Table 1 summarizes the estimated component cost.

The reference BOM is approximately \$149 at low-volume pricing. This is a design estimate rather than a production quote. Because one endpoint can be shared, the hardware cost is not incurred per learner.

3.2 Offline Multimodal Software Pipeline

The INMP441 microphone supplies spoken input to `whisper.cpp`. Touch input bypasses transcription. Both paths enter a prompt manager that applies the tutoring prompt and local dialogue context.

The LLM backend supports a portable `llama.cpp` CPU baseline and an RKLLM NPU path where the model format is supported. Both use the same

Table 1. Estimated low-volume EdgeAIED BOM. Prices are design targets in USD and exclude labor and external power supply.

Component	Role	Est. USD
Radxa CM5, 8 GB / 64 GB	Compute and storage	\$89.00
4-inch ILI9341 touchscreen	Touch display	\$18.00
4-layer carrier PCB + B2B connectors	Integration	\$20.00
M.2 Wi-Fi/Bluetooth module	Connectivity	\$8.00
INMP441 MEMS microphone	Audio input	\$1.50
MAX98357A amplifier + 3-W speaker	Audio output	\$3.50
Enclosure, mounting, and connectors	Packaging and service I/O	\$9.00
Estimated total		\$149.00

application interface, so model and runtime can change independently of the interaction.

The initial family is Qwen2.5 at 1.5B and 3B parameters, but the architecture is model-agnostic. Piper converts generated text to speech. Models and dialogue state remain on eMMC, and the interaction does not require a network.

3.3 Shared Interaction and Deployment Model

EdgeAIED is intended as a shared classroom endpoint. Learners interact through touch or speech, while a teacher-facing configuration layer selects prompts, models, or interaction modes. Local inference avoids installing multi-gigabyte models on each learner device.

The same interface is retained while model size, precision, runtime, or compute path changes. This turns the stack into a controlled testbed for pedagogical elasticity. The individual components are existing technologies. The technical contribution is their integration into an evaluation-ready AIED endpoint with stable interfaces between interaction, inference, and output.

4 Evaluation

We evaluate EdgeAIED as a coupled model–runtime–hardware system. Configurations are compared by systems performance and educational response quality.

RQ1: Systems feasibility. How do model size, precision, and runtime affect latency, throughput, memory, and energy use?

RQ2: Educational quality. How much educational response quality is retained as model size and precision are reduced?

RQ3: Deployment trade-offs. Which configurations provide the best trade-off between response quality, latency, power, and endpoint cost?

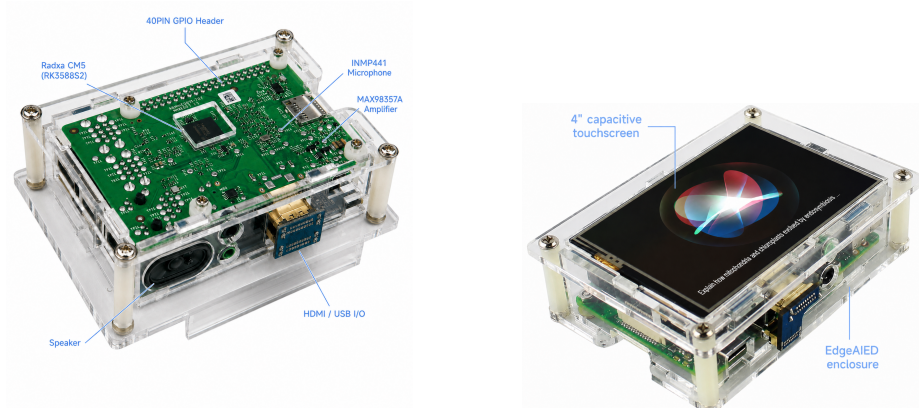


Fig. 3. AI-generated design renders of the proposed EdgeAIED enclosure: exposed hardware (left) and touchscreen configuration (right). These are conceptual renders, not photographs of a fabricated prototype.

4.1 Experimental Configurations

The EdgeAIED hardware and learner-facing pipeline remain fixed across conditions. We vary lightweight instruction-tuned models in the 1–3B range, initially centered on Qwen2.5. Where supported, models are tested at 4-bit and 8-bit precision through `llama.cpp` and RKLLM. This isolates model, precision, and inference backend while holding the interaction constant.

4.2 Systems Metrics

We record TTFT, end-to-end latency, generation throughput, peak memory, storage footprint, and device-input power. Repeated runs use the same prompt set for each configuration.

4.3 Educational Response Quality

All model configurations receive the same educational prompts and system instructions. Tasks cover explanation, misconception correction, worked guidance, and Socratic questioning.

Responses are assessed for factual correctness and relevance, then for clarity and pedagogical appropriateness. This study does not measure learning gains. The goal is to determine whether cheaper configurations retain enough educational quality to justify later classroom evaluation.

Table 2. Evaluation dimensions varied or measured in the EdgeAIED study.

Dimension	Experimental variable / measurement
Model	Lightweight instruction-tuned LLM, approximately 1–3B
Precision	INT4 / INT8 where supported
Runtime	<code>llama.cpp</code> and RKLLM where supported
Compute path	RK3588S2 CPU / NPU-supported execution
Latency	TTFT and end-to-end response time
Throughput	Generated tokens per second
Memory	Peak resident memory
Energy	Input power and energy per interaction
Storage	Local model/runtime footprint
Educational quality	Correctness, clarity, relevance, tutoring behavior

5 Preliminary Results

6 Discussion

7 Limitations and Future Work

8 Conclusion

References

1. Ardakani, M., Malekar, J., Zand, R.: Llmpi: Optimizing llms for high-throughput on raspberry pi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 6433–6442 (2025)
2. Gupta, R., Bhatia, G.: Agentic educational content generation for african languages on edge devices. In: 2025 IEEE Global Humanitarian Technology Conference (GHTC). IEEE (2025). <https://doi.org/10.1109/GHTC66843.2025.11265979>
3. Isotani, S., Bittencourt, I.I., Chalco, G.C., Dermeval, D., Mello, R.F.: Aied unplugged: Leapfrogging the digital divide to reach the underserved. In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. Communications in Computer and Information Science, vol. 1831, pp. 772–779. Springer (2023). https://doi.org/10.1007/978-3-031-36336-8_118
4. Lamouille, D., Zorec, M.B., Baksh, F., Kruusamäe, K.: Benchmarking local language models for social robots using edge devices. In: 2026 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO). pp. 237–244. IEEE (2026). <https://doi.org/10.1109/ARSO68304.2026.11536131>
5. Nguyen, T., Nguyen, T.: An evaluation of llms inference on popular single-board computers. arXiv preprint arXiv:2511.07425 (2025). <https://doi.org/10.48550/arXiv.2511.07425>
6. Rodrigues, L.A.L., Guerino, G., Silva, T.E.V., Chalco, G.C., Oliveira, L., Penha, R.S.d., Mello, R.F., Vieira, T., Marinho, M.L.M., Macario, V., Bittencourt, I.I., Dermeval, D., Isotani, S.: Mathaide: A qualitative study of teachers’ perceptions of an its unplugged for underserved regions. International Journal of Artificial Intelligence in Education **35**, 2–30 (2025). <https://doi.org/10.1007/s40593-024-00397-y>

7. Santos, M.M., Barros, A.P., Rodrigues, L., Dermeval, D., Primo, T.T., Bittencourt, I.I., Isotani, S.: Near feasibility, distant practicality: Empirical analysis of deploying and using llms on resource-constrained smartphones. In: Proceedings of the 13th International Conference on Information & Communication Technologies and Development. pp. 224–235. ACM (2024). <https://doi.org/10.1145/3700794.3700817>
8. Souza, E., dos Santos, E.L., Rodrigues, L., Rosa, D.C., Cordeiro, F.R., Pereira, C., Chevtchenko, S.F., Carvalho, R., Vieira, T., Marinho, M., Dermeval, D., Bittencourt, I.I., Isotani, S., Macario, V.: Handwritten equation detection in disconnected, low-cost mobile devices. In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. Communications in Computer and Information Science, vol. 2151, pp. 132–139. Springer (2024). https://doi.org/10.1007/978-3-031-64312-5_16
9. Uema, M.A., de Souza, T.d.P.C., Rodrigues, L., Dermeval, D., Bittencourt, I.I., Isotani, S.: Designing pedagogical elasticity for aided unplugged with local llms. In: Artificial Intelligence in Education. Late Breaking Results, WideAIED, Practitioners, Industry and Policies, Blue Sky, Doctoral Consortium, FoL Workshops and Tutorials, FoL Invited Papers. Communications in Computer and Information Science, vol. 3032, pp. 138–152. Springer (2026). https://doi.org/10.1007/978-3-032-29791-4_10
10. Uema, M.A., de Souza, T.d.P.C., Rodrigues, L., Dermeval, D., Bittencourt, I.I., Isotani, S.: Towards elastic offline-first applications for aided unplugged. In: Artificial Intelligence in Education. Lecture Notes in Computer Science, vol. 16586, pp. 415–423. Springer (2026). https://doi.org/10.1007/978-3-032-29773-0_33
11. UNESCO UNITWIN Chair on AIED Unplugged: Unesco unitwin chair on artificial intelligence unplugged in education. <https://iaedu.org.br/en/unescochair/> (2026), hosted by the Federal University of Alagoas and the University of São Paulo; accessed August 29, 2026