Near Feasibility Distant Practicality: Empirical Analysis of Deploying and Using LLMs on Resource-Constrained Smartphones

Mateus Monteiro Santos[1], **Diego Dermeval** [1], **Luiz Rodrigues**[1,2]

[1] Center of Excellence in Social Technologies (NEES) –
Federal University of Alagoas, Brazil
Maceió – AL – Brazil

[2]Technological Federal University of Paraná (UTFPR) – Apucarana, Brazil
Apucarana – PR – Brazil

mateus.monteiro@nees.ufal.br, diego.matos@nees.ufal.br

luiz.rodrigues@utfpr.edu.br

Artificial Intelligence in Education (AIED) has reshaped teaching and learning practices, but its integration also widened inequalities, especially in underserved contexts lacking infrastructure and resources [Chen et al. 2022, Crompton et al. 2024]. Large Language Models (LLMs) offer personalized feedback and support for teaching tasks [Wang et al. 2024, Veloso et al. 2023, Jeon and Lee 2023], yet their benefits are not equally distributed, as cloud dependence excludes students and teachers in low-resource environments [Gottschalk and Weise 2023]. The AIED Unplugged paradigm emerged as a pragmatic response [Isotani et al. 2023], leveraging lightweight and offline solutions through devices already present in schools, such as smartphones [Rodrigues et al. 2024a]. In this context, we address the gap of whether LLMs can operate under unplugged conditions by empirically evaluating TinyLlama, RedPajama, and Qwen2 running locally on smartphones, considering their feasibility to support teaching and learning where connectivity is limited.

Our approach deployed pre-quantized models optimized with the Machine Learning Compiler (MLC) on an Android application, with performance monitored through Sentry focusing on inference time, memory, and storage. The dataset included 180 Biology questions labeled by Bloom's taxonomy [Rodrigues et al. 2024b]. TinyLlama achieved the lowest average response time (1.35 min) and memory use (3.2 GB RAM, 622 MB storage), RedPajama was slower but more consistent (1.8 min, 3.9 GB RAM, 1.57 GB storage), and Qwen2, though robust, required more resources (2.3 min, 4.1 GB RAM, 902 MB storage). Response times increased with task complexity, especially in "Creating" and "Evaluating," posing challenges for timely classroom feedback. Thus, TinyLlama appears most promising for AIED Unplugged, but trade-offs in speed, stability, and resource demand highlight the need for optimization. Key aspects such as energy consumption, multilingual support, and broader datasets remain crucial, yet enabling LLMs to run offline on smartphones marks a step toward democratizing AIED in underserved contexts.

The entire research work Near Feasibility, Distant Practicality: Empirical Analysis of Deploying and Using LLMs on Resource-Constrained Smartphones [Monteiro Santos et al. 2025], which includes comprehensive methodology, datasets, and

DOI: 10.5753/cbie_estendido.2025.13052

analyses, provides the foundation for this summary article.

References

[Chen et al. 2022] Chen, X., Zou, D., Xie, H., Cheng, G., and Liu, C. (2022). Two decades of artificial intelligence in education. *Educational Technology & Society*, 25(1):28–47.

[Crompton et al. 2024] Crompton, H., Jones, M. V., and Burke, D. (2024). Affordances and challenges of artificial intelligence in k-12 education: A systematic review. *Journal of Research on Technology in Education*, 56(3):248–268.

[Gottschalk and Weise 2023] Gottschalk, F. and Weise, C. (2023). Digital equity and inclusion in education: An overview of practice and policy in oecd countries.

[Isotani et al. 2023] Isotani, S., Bittencourt, I. I., Challco, G. C., Dermeval, D., and Mello, R. F. (2023). Aied unplugged: Leapfrogging the digital divide to reach the underserved. In *International Conference on Artificial Intelligence in Education*, pages 772–779. Springer.

[Jeon and Lee 2023] Jeon, J. and Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892.

[Monteiro Santos et al. 2025] Monteiro Santos, M., Barros, A., Rodrigues, L., Dermeval, D., Primo, T., Ibert, I., and Isotani, S. (2025). 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*, ICTD '24, page 224–235, New York, NY, USA. Association for Computing Machinery.

[Rodrigues et al. 2024a] Rodrigues, L., Guerino, G., Silva, T. E., Challco, G. C., Oliveira, L., da Penha, R. S., Melo, R. F., Vieira, T., Marinho, M., Macario, V., et al. (2024a). Mathaide: A qualitative study of teachers' perceptions of an its unplugged for underserved regions. *International Journal of Artificial Intelligence in Education*, pages 1–29.

[Rodrigues et al. 2024b] Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., and Mello, R. F. (2024b). Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, page 100248.

[Veloso et al. 2023] Veloso, T., Chalco Challco, G., Rogrigues, L., Versuti, F., Sena da Penha, R., Silva Oliveira, L., and Isotani, S. (2023). Its unplugged: Leapfrogging the digital divide for teaching numeracy skills in underserved populations. In *Towards the Future of AI-augmented Human Tutoring in Math Learning 2023-Proceedings of the Workshop on International Conference of Artificial Intelligence in Education co-located with The 24th International Conference on Artificial Intelligence in Education*.

[Wang et al. 2024] Wang, B., Liu, J., Karimnazarov, J., and Thompson, N. (2024). Task supportive and personalized human-large language model interaction: A user study. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 370–375.