LoXR: Performance Evaluation of LLMs Executed Locally on XR Devices

Xinyu Liu, Dawar Khan, Omar Mena, Donggan Jia, Alexandre Kouyoumdjian, Ivan Viola
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract: LoXR (LLMs on XR) evaluates 17 LLMs across four XR devices—Magic Leap 2 (ML2), Meta Quest 3 (MQ3), Vivo X100s Pro, and Apple Vision Pro (AVP)—assessing performance on key metrics: consistency, processing speed, & battery consumption. The study examines 68 model-device pairs under varying string lengths, batch sizes, and thread counts, providing insights into trade-offs for real-time XR applications. LoXR offers guidance for optimizing LLM deployment on XR devices and establish a foundation for future research in this rapidly evolving field.

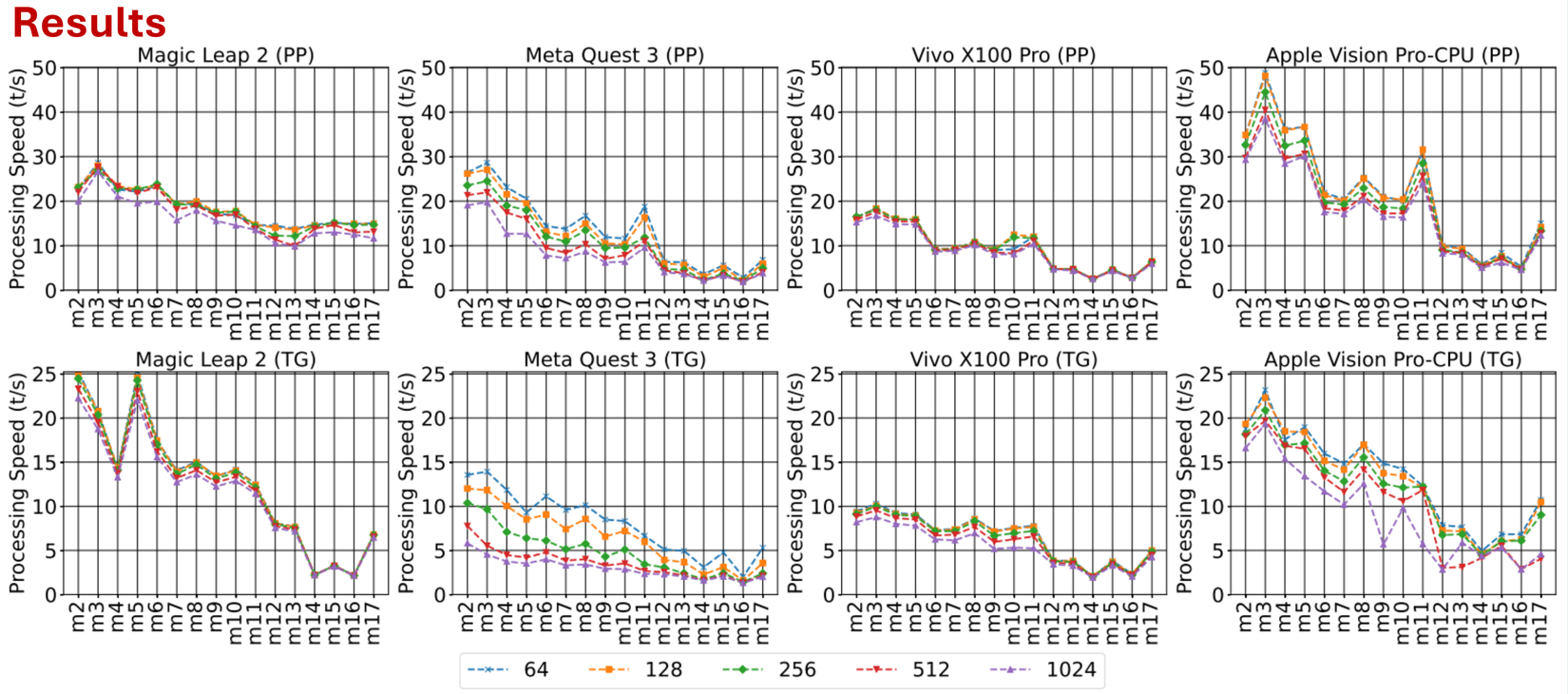
Methods and Metrics

- **Performance Consistency:** Each model-device pair is tested 20 times to assess stability over time.
- **Prompt Processing (PP):** Measures device speed in handling input prompts of lengths 64, 128, 256, 512, and 1024.
- **Token Generation (TG):** Evaluates the speed of generating output tokens for token set sizes 64, 128, 256, 512, and 1024.
- **Batch Test (BT):** with batch sizes 64,128,256,512, & 1024.
- **Thread Test:** with varying thread counts (1, 2, 4, 8, 16, 32).

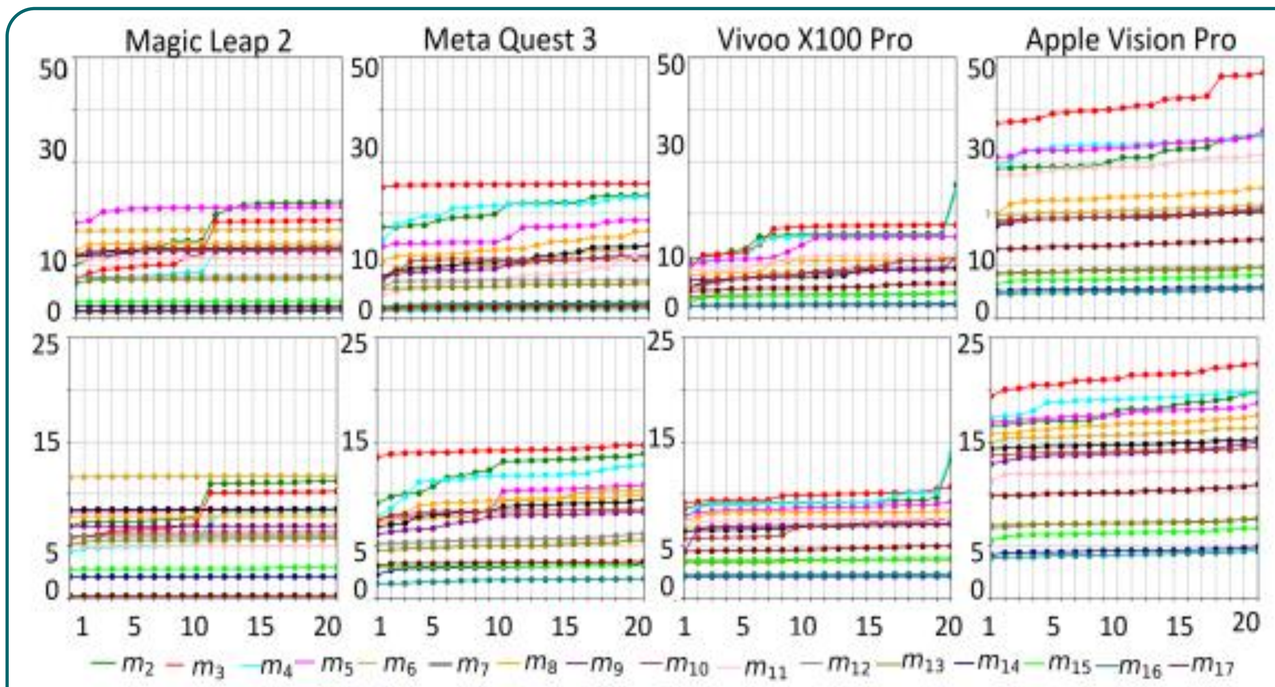
Large Language Models

ID (Size)	Model Name
m_1 (0.942 GB)	qwen2-0_5b-instruct-fp16
m_2 (1.36 GB)	Vikhr-Gemma-2B-instruct-Q3_K_M
m_3 (1.51 GB)	Vikhr-Gemma-2B-instruct-Q4_0
m_4 (1.75 GB)	Vikhr-Gemma-2B-instruct-Q5_0
m_5 (2.00 GB)	Vikhr-Gemma-2B-instruct-Q6_K
m_6 (1.32 GB)	Phi-3.1-mini-4k-instruct-Q2_K
m_7 (1.94 GB)	Phi-3.1-mini-4k-instruct-Q3_K_L
m_8 (2.30 GB)	Phi-3.1-mini-4k-instruct-Q4_K_L
m_9 (2.68 GB)	Phi-3.1-mini-4k-instruct-Q5_K_L
m_{10} (2.92 GB)	Phi-3.1-mini-4k-instruct-Q6_K
m_{11} (3.78 GB)	Phi-3.1-mini-4k-instruct-Q8_0
m_{12} (2.63 GB)	llama-2-7b-chat.Q2_K
m_{13} (2.75 GB)	llama-2-7b-chat.Q3_K_S
m_{14} (1.64 GB)	Mistral-7B-Instruct-v0.3.IQ1_M
m_{15} (2.05 GB)	Mistral-7B-Instruct-v0.3.IQ2_XS
m_{16} (2.81 GB)	Mistral-7B-Instruct-v0.3.IQ3_XS
m_{17} (3.64 GB)	Mistral-7B-Instruct-v0.3.IQ4_XS

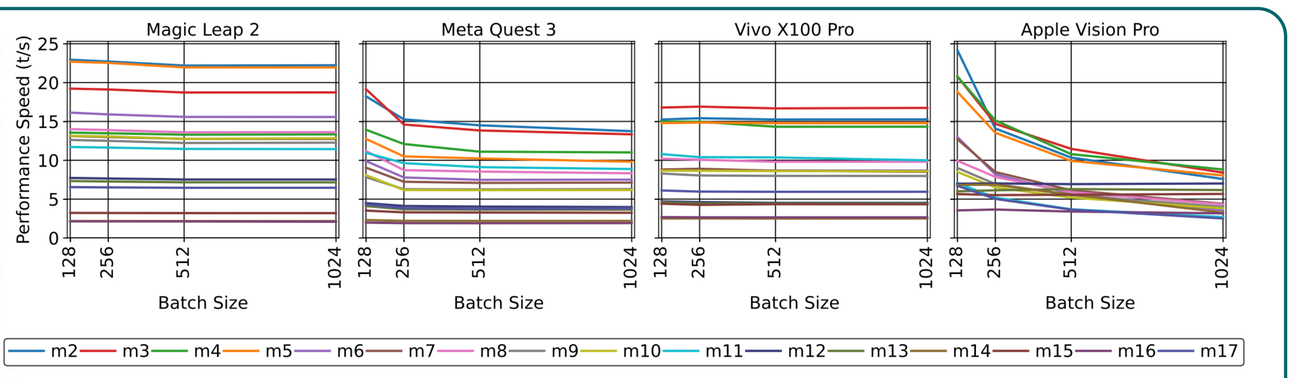
Results



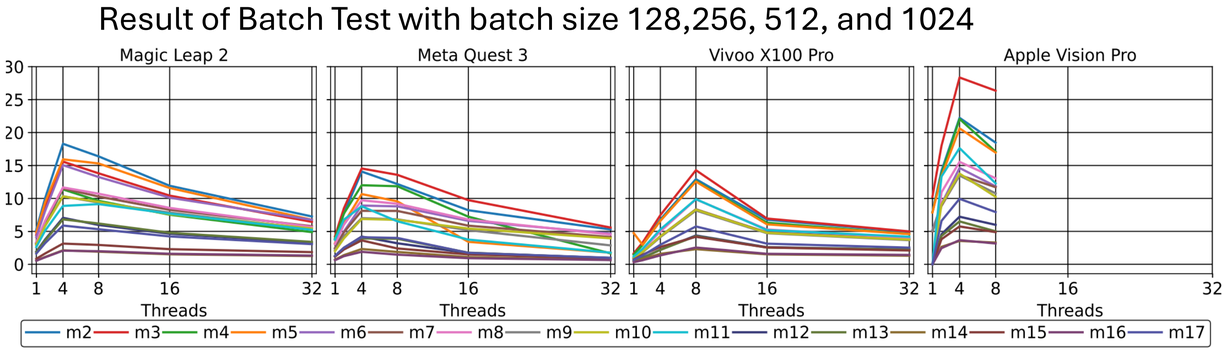
Processing speed with varying string lengths: 64, 128, 256, 512, and 1024. Top: PP, Bottom: TG .



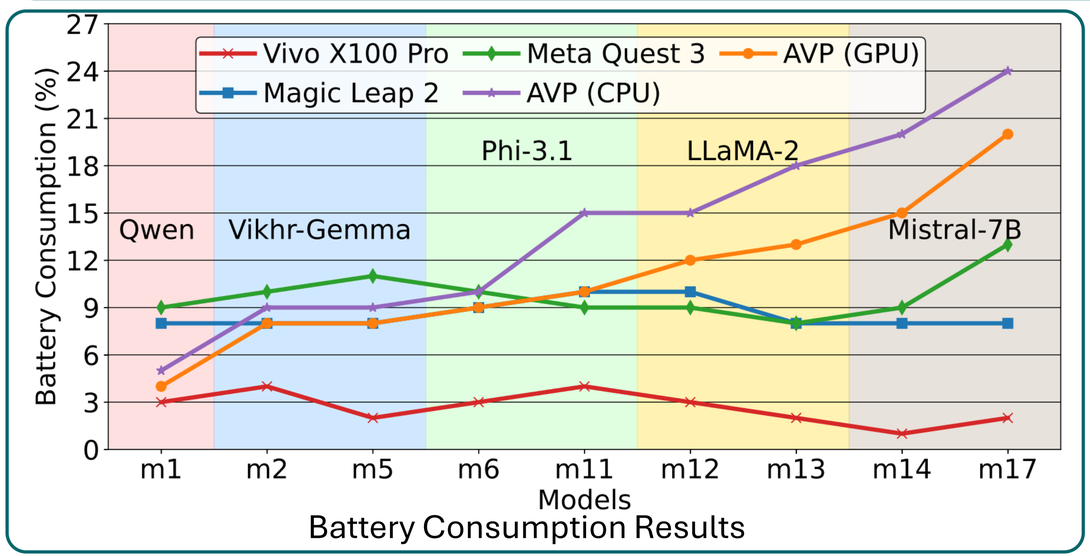
Consistency results of the four devices over time: PP (top) and TG (bottom) speeds in tokens per second across 20 sorted runs (X-axis: run no., Y-axis: speed in t/s).



Result of Batch Test with batch size 128,256, 512, and 1024



Result of Thread Test with thread counts of 1,2, 4, 8, 16 and 32.



Conclusion: This study deploys LLMs on four XR devices and conducts a comprehensive evaluation using a fair experimental setup. Our work establishes a baseline for on-device AI in XR, enabling researchers to benchmark models, devices, and algorithms effectively.
For the models (m2 to m16), the average inference speeds were as follows. AVP achieved the highest inference speed (16.91 t/s PP, 11.04 t/s TG), followed by ML2 (12.68 t/s PP, 8.51 t/s TG), MQ3 (9.52 t/s PP, 5.77 t/s TG), and Vivo (8.62 t/s PP, 6.12 t/s TG), reflecting performance variations across XR devices. For more details and best performing model-device pairs as identified by Pareto analysis please refer to the full paper on arXiv.

Full Paper: 1. Dawar Khan, Xinyu Liu, Omar Mena, Donggang Jia, Alexandre Kouyoumdjian, and Ivan Viola. 2024."LoXR: Performance Evaluation of Locally Executing LLMs on XR Devices." arXiv preprint arXiv:2502.15761. https://arxiv.org/abs/2502.15761