SOVEREIGN: LFM2 Technical Report

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

We present LFM2, a family of Liquid Foundation Models designed for efficient on-device deployment and strong task capabilities. Using hardware-in-the-loop architecture search under edge latency and memory constraints, we obtain a compact hybrid backbone that combines gated short convolutions with a small number of grouped query attention blocks, delivering up to 2x faster prefill and decode on CPUs compared to similarly sized models. The LFM2 family covers 350M-8.3B parameters, including dense models (350M, 700M, 1.2B, 2.6B) and a mixture-of-experts variant (8.3B total, 1.5B active), all with

1 Introduction

Analysis of: LFM2 Technical Report. Research goal: What is the impact of increasing the number of active experts (k) on inference latency and VQA accuracy in sparse MoE vision-language models, and does the optimal k vary with visual complexity of the input?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

14 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
LFM2 models achieve up to 2x faster prefill and decode on CPUs compared to similarly sized models.	✓	0.23
The LFM2 family covers 350M-8.3B parameters, including dense models (350M, 700M, 1.2B, 2.6B) and a mixture-of-experts va	✓	0.36
LFM2's training pipeline includes a tempered, decoupled Top-K knowledge distillation objective that avoids support misma	✓	0.26
LFM2 models are pre-trained on 10-12T tokens.	✓	0.17
LFM2-2.6B reaches 79.56% on IFEval and 82.41% on GSM8K.	✓	0.23
LFM2-Audio enables real-time speech-to-speech interaction competitive with models 3x larger.	✓	0.25

References

- https://www.semanticscholar.org/paper/dcedb2b0f21d5731144d6475363b37deaf634ce0
- https://www.semanticscholar.org/paper/e6c3e40973c9f51ebbd36d13dbb6b2470ae5c9b7
- https://www.semanticscholar.org/paper/e8cb9fafeb45cd64230548084c3420d72deb2217