SOVEREIGN: What is the trade-off between expert count (k) and throughput (tokens/sec) on edge CPU devices for sparse MoE

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Rapid advancements in large language models (LLMs) have increased interest in deploying them on mobile devices for on-device AI applications. Mobile users interact differently with LLMs compared to desktop users, creating unique expectations and data biases. Current benchmark datasets primarily target at server and desktop environments, and there is a notable lack of extensive datasets specifically designed for mobile contexts. Additionally, mobile devices face strict limitations in storage and computing resources, constraining model size and capabilities, thus requiring optimized efficiency a

1 Introduction

Analysis of: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research goal: What is the trade-off between expert count (k) and throughput (tokens/sec) on edge CPU devices for sparse MoE VLMs evaluated on the MMLU and ChartQA benchmarks, under latency constraints similar to those in LFM2?.

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

11 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 5.2/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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
The Mobile-MMLU benchmark includes models ranging from 1B to 9B parameters.	×	0.08
Qwen2.5-3B-instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
Llama-3.2-3B-instruct scores 50.2% accuracy on Mobile-MMLU.	×	0.05
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.07
The performance spread on MMLU ranges from 45.9% to 71.8%.	×	0.03
Phi-3.5-mini-instruct scored 63.7% on Mobile-MMLU.	×	0.05

References

- http://arxiv.org/abs/2503.20786v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2603.11114v1