SOVEREIGN: What is the accuracy-throughput Pareto frontier of SMoES MoE-VLMs versus dense models on cross-modal reasoning

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Migrating computational intensive tasks from mobile devices to more resourceful cloud servers is a promising technique to increase the computational capacity of mobile devices while saving their battery energy. In this paper, we consider an MIMO multicell system where multiple mobile users (MUs) ask for computation offloading to a common cloud server. We formulate the offloading problem as the joint optimization of the radio resources-the transmit precoding matrices of the MUs-and the computational resources-the CPU cycles/second assigned by the cloud to each MU-in order to minimize the overal

1 Introduction

Analysis of: Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing. Research goal: What is the accuracy-throughput Pareto frontier of SMoES MoE-VLMs versus dense models on cross-modal reasoning tasks (e.g., ChartQA, DocVQA) when controlling for total parameter count at 7B and 34B scales, and how does expert specialization affect per-layer latency?.

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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 3.3/10 → REJECT (revision_round=0). 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.

References

- https://openalex.org/W2116040783
- https://doi.org/10.1109/les.2017.2774800
- https://doi.org/10.1109/tsipn.2015.2448520