SOVEREIGN: What is the inference throughput (tokens per second) and memory footprint trade-off for MoE-LLaVA versus dense

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Sparse Mixture-of-Experts (MoE) models can outperform dense large language models at similar computation by activating only a small set of experts per token. However, stacking many expert modules introduces substantial parameter memory, which makes MoE models difficult to deploy in memory-constrained environments such as single-GPU devices. Offloading alleviates this issue by storing inactive experts in CPU memory and loading them on demand, but existing methods remain limited: static caches disregard input-dependent routing, and methods that train separate models to predict expert usage ahead

1 Introduction

Analysis of: ExpertFlow: Efficient Mixture-of-Experts Inference via Predictive Expert Caching and Token Scheduling. Research goal: What is the inference throughput (tokens per second) and memory footprint trade-off for MoE-LLaVA versus dense models of equivalent total parameter count on the MMMU benchmark at 7B and 13B scales?.

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

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 2.2/10 → 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

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2401.15947v5