SOVEREIGN: How does the inference throughput and memory efficiency of SMoES-based 7B VLMs compare against dense and hard-

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: How does the inference throughput and memory efficiency of SMoES-based 7B VLMs compare against dense and hard-routing MoE baselines on MMBench and SEED-Bench under varying batch sizes and sequence lengths?.

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

8 papers retrieved. 8 claims extracted, 0 verified. Tribunal: 3.0/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
ExpertFlow achieves up to 9.99× throughput improvement over Cache-MoE on Switch-64 model.	×	0.04
ExpertFlow achieves 2.04× throughput improvement over SE-MoE on Switch-32 model.	×	0.03
ExpertFlow achieves 2.16× throughput improvement over Pregated-MoE on Switch-32 model.	×	0.03
The Switch-128 model has 97.75% expert utilization rate.	×	0.05
Mixtral-8×7B model has 2/8 active experts per token.	×	0.06
Qwen1.5-MoE model has 4/60 active experts per token.	×	0.07
Experiments were conducted on a single NVIDIA A40 GPU with 48 GB of memory.	×	0.07
The Deepseek-MoE model has 27 layers and 16.40 billion total parameters.	×	0.03

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2410.21676v4
- http://arxiv.org/abs/2603.11114v1