SOVEREIGN: What is the generalization capability of SMoES on out-of-domain VQA benchmarks like VQA-CP v2 and A-OKVQA, and

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Mixture-of-Experts architectures have become the standard for scaling large language models due to their superior parameter efficiency. To accommodate the growing number of experts in practice, modern inference systems commonly adopt expert parallelism to distribute experts across devices. However, the absence of explicit load balancing constraints during inference allows adversarial inputs to trigger severe routing concentration. We demonstrate that out-of-distribution prompts can manipulate the routing strategy such that all tokens are consistently routed to the same set of top-k experts,

1 Introduction

Analysis of: RepetitionCurse: Measuring and Understanding Router Imbalance in Mixture-of-Experts LLMs under DoS Stress. Research goal: What is the generalization capability of SMoES on out-of-domain VQA benchmarks like VQA-CP v2 and A-OKVQA, and how does this compare to fixed routing strategies in terms of accuracy drop and expert utilization metrics?.

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

13 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 0.5/10 → 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
Over 50% of surveyed MoE models from Huggingface with over 1,000 downloads are Mixtral-like models with E $\leq$ 8.	×	0.04
Remaining models with architectures like DeepSeekV3 and Qwen3Moe have typically only 2 to 12 experts activated out of E	×	0.03
The study focuses on 13 popular MoE models including 4 Mixtral-like low-sparse models and 9 high-sparse models.	×	0.05
Selected models cover both base and post-trained variants, different attention mechanisms (standard and linear), and exp	×	0.03
For EP size $\leq$ 8, a single NVIDIA A800 GPU is utilized.	×	0.05
For EP size = 16 and 32, the setup uses 2 and 4 nodes of NVIDIA A800 GPUs respectively.	×	0.02

References

- https://arxiv.org/abs/2512.23995
- http://arxiv.org/abs/2408.15664v1
- http://arxiv.org/abs/2506.07366v1