SOVEREIGN: For sparse MoE vision-language models, how does the optimal number of active experts (k) change when evaluated

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Sparse Mixture-of-Experts (MoE) architectures enable efficient scaling of large language models through conditional computation, yet the routing mechanisms responsible for expert selection remain poorly understood. In this work, we introduce routing signatures, a vector representation summarizing expert activation patterns across layers for a given prompt, and use them to study whether MoE routing exhibits task-conditioned structure. Using OLMoE-1B-7B-0125-Instruct as an empirical testbed, we show that prompts from the same task category induce highly similar routing signatures, while prompts

1 Introduction

Analysis of: Task-Conditioned Routing Signatures in Sparse Mixture-of-Experts Transformers. Research goal: For sparse MoE vision-language models, how does the optimal number of active experts (k) change when evaluated on visually complex inputs (e.g., high-resolution or cluttered scenes from COCO or VizWiz) versus simple ones, and what is the resulting accuracy gap compared to dense baselines?.

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

12 papers retrieved. 5 claims extracted, 0 verified. Tribunal: 4.5/10 → 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
The routing signature similarity matrix shows strong diagonal dominance with within-category similarities between 0.83 a	×	0.05
Within-category routing similarities are higher than cross-category similarities, with within-category values around 0.8	×	0.07
The ordering of routing similarity follows the sequence: Across < Load-Balance < Within.	×	0.06
Task separation in routing similarity is weakest in early layers and strongest in deeper layers, peaking around layer 13	×	0.11
Routing signatures form distinct clusters when projected using PCA, with separate regions for code, math, story, and fac	×	0.08

References

- http://arxiv.org/abs/2509.09014v1
- http://arxiv.org/abs/2305.14882v2
- http://arxiv.org/abs/2603.11114v1