

SOVEREIGN: How does scaling top-k selection in SMoES affect inference throughput and MMMU accuracy trade-off under cross-

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: How does scaling top-k selection in SMoES affect inference throughput and MMMU accuracy trade-off under cross-modal adversarial attacks versus dense VLMs?.

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. 0 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.

References

- <http://arxiv.org/abs/2403.10883v2>
- <http://arxiv.org/abs/2410.17954v2>
- <http://arxiv.org/abs/2603.11114v1>