SOVEREIGN: Does task-conditioned routing signatures in SMoE transformers improve compositional reasoning accuracy on NLVR

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: Does task-conditioned routing signatures in SMoE transformers improve compositional reasoning accuracy on NLVR2 compared to fixed-ratio top-k routing at equal total expert parameters, controlling for model scale?.

2 Methodology

Multi-query arXiv search (1 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

3 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
The diagonal entries in the routing signature similarity matrix are consistently higher than off-diagonal entries.	×	0.02
Within-category routing similarities lie between 0.83 and 0.85.	×	0.05
Cross-category routing similarities typically lie between 0.58 and 0.64.	×	0.05
Routing similarity follows the ordering: Across < Load-Balance < Within.	×	0.06
Task separation in routing behavior is weakest in early layers and strongest in deeper layers, peaking around layer 13.	×	0.09
PCA projection of routing signatures shows distinct clusters for code, math, story, and factual prompts.	×	0.07
The model OLMoE-1B-7B-0125-Instruct contains 16 MoE layers with 64 experts per layer.	×	0.14
Each prompt generates 32 tokens during inference.	×	0.02

References

- http://arxiv.org/abs/1403.3007v3
- http://arxiv.org/abs/2601.22323v2
- http://arxiv.org/abs/2603.11114v1