# SOVEREIGN: Can task-conditioned routing signatures improve the robustness of sparse MoE transformers to distribution shif

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: Can task-conditioned routing signatures improve the robustness of sparse MoE transformers to distribution shifts in multimodal reasoning benchmarks (e.g., VQAv2 vs. Visual7W), measured via accuracy degradation and inference stability under varying batch sizes?.

## 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

6 papers retrieved. 7 claims extracted, 0 verified. Tribunal: 6.5/10 → REVISE (revision_round=1). 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 within-category routing signature similarity is between 0.83 and 0.85. | × | 0.07 |
| The cross-category routing signature similarity is between 0.58 and 0.64. | × | 0.08 |
| Routing similarity follows the ordering: Within > LoadBalance > Across. | × | 0.05 |
| Task separation in routing similarity is weakest in early layers and strongest in deeper layers, peaking around layer 13 | × | 0.11 |
| PCA projection of routing signatures shows distinct clusters for code, math, story, and factual prompt categories. | × | 0.08 |
| The model used is OLMoE-1B-7B-0125-Instruct with 16 MoE layers and 64 experts per layer. | × | 0.14 |
| Each prompt category consists of 20 prompts, with categories including Code, Math, Story, and Factual. | × | 0.05 |

## References


- http://arxiv.org/abs/2603.03437v1
- http://arxiv.org/abs/2504.16021v1
- http://arxiv.org/abs/2603.11114v1