# SOVEREIGN: Does AnyExperts' dynamic expert allocation maintain consistent accuracy improvements over dense baselines when

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 AnyExperts' dynamic expert allocation maintain consistent accuracy improvements over dense baselines when scaling from 8 to 64 experts on challenging reasoning tasks like those found in ScienceQA and ARO datasets?.

## 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

11 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 3.7/10 → 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 model evaluated is OLMoE-1B-7B-0125-Instruct with 16 MoE layers, 64 experts per layer, and top-k routing with k=8. | × | 0.14 |
| Within-category routing similarity is between 0.83 and 0.85, while cross-category similarity is between 0.58 and 0.64. | × | 0.12 |
| Routing similarity follows the ordering: Across < Load-Balance < Within. | × | 0.08 |
| Task separation effect size (Cohen's d) grows stronger toward deeper layers, peaking around layer 13. | × | 0.08 |
| The first two principal components of PCA projection of routing signatures explain a substantial fraction of the varianc | × | 0.07 |
| Story prompts form a clearly separated cluster, while code and math form different but partially adjacent clusters in PC | × | 0.02 |

## References


- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2002.00741v1
- http://arxiv.org/abs/2603.11114v1