SOVEREIGN: Can mixture-of-experts routing strategies trained on imbalanced multimodal data improve inference efficiency a

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 mixture-of-experts routing strategies trained on imbalanced multimodal data improve inference efficiency and maintain accuracy under domain shift on document and infographic understanding benchmarks, and how do these trade-offs compare to uniform modality routing 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

11 papers retrieved. 11 claims extracted, 0 verified. Tribunal: 2.8/10 → 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
Figure 1 shows the routing signature similarity matrix across the four task categories.	×	0.08
Figure 2 shows the routing signature similarity matrix across the four task categories.	×	0.08
Within-category routing signature similarities lie between 0.83 and 0.85.	×	0.05
Cross-category similarities typically lie between 0.58 and 0.64.	×	0.03
Figure 3 shows the layer-wise effect size (Cohen's d) separating within-category and across-category routing similaritie	×	0.09
Figure 4 visualizes routing signatures projected into two dimensions using PCA.	×	0.10
The model OLMoE-1B-7B-0125-Instruct contains 16 MoE layers, 64 experts per layer, and uses top-k routing with k=8.	×	0.15
Each prompt generated 32 tokens during inference.	×	0.02
Story prompts occupy a clearly separated region in the PCA projection.	×	0.01
Code and math prompts form different but partially adjacent clusters in the PCA projection.	×	0.02
Factual prompts form a distinguishable cluster in the routing signature projection.	×	0.03

References

- http://arxiv.org/abs/2406.08610v1

- http://arxiv.org/abs/2603.11114v1
- http://arxiv.org/abs/2508.05993v3