SOVEREIGN: What is the inference throughput and memory efficiency trade-off of SMoES relative to dense baselines and hard

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: What is the inference throughput and memory efficiency trade-off of SMoES relative to dense baselines and hard-coded modality routing across different MoE sparsity levels on multimodal QA tasks?.

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. 14 claims extracted, 0 verified. Tribunal: 1.7/10 → 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
Within-category routing similarities lie between 0.83 and 0.85.	×	0.06
Cross-category routing similarities are between 0.58 and 0.64.	×	0.07
Within-task prompts are more similar than the load-balancing baseline predicts.	×	0.10
Cross-task prompts are less similar than the load-balancing baseline predicts.	×	0.10
Task-conditioned separation grows stronger toward deeper layers, peaking around layer 13.	×	0.11
Early layers capture lexical and local structure, while deeper layers reflect more semantically differentiated computati	×	0.07
The first two principal components of routing signatures explain a substantial fraction of the variance.	×	0.07
PCA projection of routing signatures shows distinct clusters for code, math, story, and factual prompts.	×	0.07
Story prompts occupy a clearly separated region in the PCA projection.	×	0.01
Code and math form different but partially adjacent clusters in the PCA projection.	×	0.02
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
Only 8 of 64 experts are active per token in each MoE layer, corresponding to a sparsity level of 12.5%.	×	0.06
The prompt dataset consists of 80 prompts across four categories: Code (20), Math (20), Story (20), and Factual (20).	×	0.03
Each prompt generated 32 tokens during inference.	×	0.04

References

- http://arxiv.org/abs/2505.22937v1
- http://arxiv.org/abs/2604.12213v1

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