SOVEREIGN: What is the throughput vs. accuracy trade-off of ExpertFlow's token scheduling in MoE vision-language models o

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 throughput vs. accuracy trade-off of ExpertFlow's token scheduling in MoE vision-language models on the AMBER benchmark when scaling expert activation budgets from 25% to 100% relative to dense models?.

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. 8 claims extracted, 0 verified. Tribunal: 3.3/10 → 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 signature similarities lie between 0.83 and 0.85.	×	0.04
Cross-category routing similarities are typically between 0.58 and 0.64.	×	0.06
Within-category prompt pairs show higher routing similarity than the load-balancing baseline.	×	0.12
Cross-category prompt pairs show lower routing similarity than the load-balancing baseline.	×	0.14
Task-conditioned separation in routing behavior grows stronger toward deeper layers, peaking around layer 13.	×	0.12
The first two principal components of PCA projection of routing signatures show distinct clusters for each task category	×	0.12
Story prompts occupy a clearly separated region in the PCA projection.	×	0.02
Code and math prompts form different but partially adjacent clusters in the PCA projection.	×	0.03

References

- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2205.15237v1
- http://arxiv.org/abs/2603.11114v1