# SOVEREIGN: How does the inference throughput and accuracy of SMoES-based MoE-VLMs with soft modality-guided routing compa

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

**Abstract**

The transformer architecture has become a cornerstone of modern AI, fueling remarkable progress across applications in natural language processing, computer vision, and multi-modal learning. As these models continue to scale explosively for performance, implementation efficiency remains a critical challenge. Mixture-of-Experts (MoE) architectures, selectively activating specialized subnetworks (experts), offer a unique balance between model accuracy and computational cost. However, the adaptive routing in MoE architectures—where input tokens are dynamically directed to specialized experts base

## 1 Introduction

Analysis of: MoEcho: Exploiting Side-Channel Attacks to Compromise User Privacy in Mixture-of. Research goal: How does the inference throughput and accuracy of SMoES-based MoE-VLMs with soft modality-guided routing compare to dense VLMs of equivalent parameter count on the MMMU benchmark at 7B and 13B scales?.

## 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

9 papers retrieved. 3 claims extracted, 3 verified. Tribunal: 7.3/10 → REVISE (revision_round=1). Policy: SOFT_APPROVE.

## 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 transformer architecture has become a cornerstone of modern AI, fueling progress across applications in natural lang | ✓ | 0.33 |
| Mixture-of-Experts (MoE) architectures selectively activate specialized subnetworks (experts) to offer a balance between | ✓ | 0.33 |
| The adaptive routing in MoE architectures dynamically directs input tokens to specialized experts. | ✓ | 0.24 |

## References


- https://doi.org/10.5281/zenodo.20417084
- https://doi.org/10.5281/zenodo.20417083
- https://doi.org/10.1007/s11704-026-60308-3