SOVEREIGN: Does the layer-dependent expert specialization learned by SMoES generalize to out-of-distribution multimodal r

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Mixture-of-Experts (MoE) Multimodal large language models (MLLMs) excel at vision-language tasks, but they suffer from high computational inefficiency. To reduce inference overhead, expert skipping methods have been proposed to deactivate redundant experts based on the current input tokens. However, we find that applying these methods-originally designed for unimodal large language models (LLMs)-to MLLMs results in considerable performance degradation. This is primarily because such methods fail to account for the heterogeneous contributions of experts across MoE layers and modality-specific b

1 Introduction

Analysis of: MoDES: Accelerating Mixture-of-Experts Multimodal Large Language Models via Dynamic Expert Skipping. Research goal: Does the layer-dependent expert specialization learned by SMoES generalize to out-of-distribution multimodal reasoning tasks (e.g., VCR or SNLI-VE) without retraining, measured by accuracy degradation compared to in-distribution GQA/NLVR2 scores?.

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. 0 claims extracted, 0 verified. Tribunal: 1.8/10 → 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.

References

- http://arxiv.org/abs/2604.23996v1
- https://arxiv.org/abs/2511.15690
- https://arxiv.org/abs/2604.19503