SOVEREIGN: What is the scaling behavior of SMoES-based VLMs from 1B to 13B parameters in terms of accuracy-latency Pareto

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Mixture-of-Experts (MoE) architectures enable conditional computation by routing inputs to multiple expert subnetworks and are often motivated as a mechanism for scaling large language models. In this project, we instead study MoE behavior in an image classification setting, focusing on predictive performance, expert utilization, and generalization. We compare dense, SoftMoE, and SparseMoE classifier heads on the CIFAR10 dataset under comparable model capacity. Both MoE variants achieve slightly higher validation accuracy than the dense baseline while maintaining balanced expert utilization th

1 Introduction

Analysis of: Mixture-of-Experts Models in Vision: Routing, Optimization, and Generalization. Research goal: What is the scaling behavior of SMoES-based VLMs from 1B to 13B parameters in terms of accuracy-latency Pareto frontier on multimodal benchmarks compared to dense and hard-routing MoE counterparts?.

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. 0 claims extracted, 0 verified. Tribunal: 4.7/10 → REVISE (revision_round=1). 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/2601.15021v1
- http://arxiv.org/abs/2303.07226v1
- http://arxiv.org/abs/2502.03009v2