SOVEREIGN: What is the impact of sparsity ratio in token-level routing on the trade-off between inference latency and mul

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 impact of sparsity ratio in token-level routing on the trade-off between inference latency and multimodal reasoning performance in MoE VLMs on SEED-Bench and MMBench benchmarks?.

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. 7 claims extracted, 1 verified. Tribunal: 3.0/10 → 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
Mixture-of-Experts (MoE) architectures offer an alternative design by introducing conditional computation, routing input	×	0.13
In large language models, MoEs are primarily motivated by efficiency and scalability.	×	0.07
This work studies Mixture-of-Experts models in an image classification context, deliberately decoupling MoE behavior fro	×	0.11
Using the CIFAR-10 dataset, this work compares dense classifier heads with SoftMoE and SparseMoE variants built on a sha	×	0.11
This controlled setup allows isolating the effects of expert routing, load balancing, and sparsity on predictive perform	×	0.07
Beyond validation accuracy, this work analyzes generalization-related behavior through the geometry of the loss landscap	×	0.12
This work computes Hessian-based sharpness metrics at convergence.	✓	0.18

References

- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2603.11114v1