SOVEREIGN: Mixture-of-Experts Models in Vision: Routing, Optimization, and Generalization

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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: Does AnyExperts' dynamic expert specialization improve compositional generalization on multi-step reasoning tasks compared to fixed routing, as measured by accuracy on the GQA and NLVR2 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

8 papers retrieved. 10 claims extracted, 10 verified. Tribunal: 8.7/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_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
Mixture-of-Experts (MoE) architectures enable conditional computation by routing inputs to multiple expert subnetworks.	✓	0.30
MoE architectures are often motivated as a mechanism for scaling large language models.	✓	0.19
The study compares dense, SoftMoE, and SparseMoE classifier heads on the CIFAR10 dataset under comparable model capacity	✓	0.24
Both MoE variants achieve slightly higher validation accuracy than the dense baseline.	✓	0.27
Both MoE variants maintain balanced expert utilization through regularization, avoiding expert collapse.	✓	0.24
SoftMoE exhibits higher sharpness by Hessian-based metrics (largest eigenvalue and trace of the loss Hessian) compared t	✓	0.26
Dense and SparseMoE lie in a similar curvature regime based on Hessian-based sharpness metrics.	✓	0.27
All models achieve comparable generalization performance.	✓	0.17
Loss surface perturbation analyses reveal qualitative differences in non-local behavior under finite parameter perturbat	✓	0.34
Naively implemented conditional routing does not yield inference speedups on modern hardware at this scale.	✓	0.27

References

- http://arxiv.org/abs/2508.17298v2
- http://arxiv.org/abs/2602.09258v1
- http://arxiv.org/abs/2601.15021v1