SOVEREIGN: What is the impact of SMoES routing on cross-dataset generalization robustness (ANLS accuracy) in zero-shot se

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Sparsely activated Mixture-of-Experts (SMoE) has shown promise to scale up the learning capacity of neural networks, however, they have issues like (a) High Memory Usage, due to duplication of the network layers into multiple copies as experts; and (b) Redundancy in Experts, as common learning-based routing policies suffer from representational collapse. Therefore, vanilla SMoE models are memory inefficient and non-scalable, especially for resource-constrained downstream scenarios. In this paper, we ask: Can we craft a compact SMoE model by consolidating expert information? What is the best re

1 Introduction

Analysis of: Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy. Research goal: What is the impact of SMoES routing on cross-dataset generalization robustness (ANLS accuracy) in zero-shot settings for unseen multimodal benchmarks (e.g., TabFact, VisualMRC) relative to dense and hard-routed MoE baselines?.

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

10 papers retrieved. 7 claims extracted, 0 verified. Tribunal: 2.2/10 → 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
MC-SMoE reaches up to an 80% memory saving with only a negligible compromise in performance on the COPA benchmark with t	×	0.04
Transformers have become the de facto network architecture in various natural language processing scenarios and computer	×	0.03
Parameter counts of transformer models are commonly measured in billions rather than millions.	×	0.04
Empirical scaling laws reveal a power-law relationship between final model quality and the amount of data, model capacit	×	0.04
Training a GPT-based model typically leads to thousands of GPU days.	×	0.04
Sparse Mixture-of-Experts (SMoE) was proposed to trim down computing cost while enabling efficient scaling of network ca	×	0.09
SMoE leverages input-dependent conditional computation for predictions of a given input.	×	0.05

References

- http://arxiv.org/abs/2602.09258v1
- http://arxiv.org/abs/2004.03143v1
- http://arxiv.org/abs/2310.01334v2