SOVEREIGN: Can AnyExperts' dynamic expert allocation maintain consistent accuracy improvements over dense baselines when

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Despite their remarkable achievement, gigantic transformers encounter significant drawbacks, including exorbitant computational and memory footprints during training, as well as severe collapse evidenced by a high degree of parameter redundancy. Sparsely-activated Mixture-of-Experts (SMoEs) have shown promise to mitigate the issue of training efficiency, yet they are prone to (1) redundant experts due to representational collapse; and (2) poor expert scalability for inference and downstream fine-tuning, primarily due to overfitting of the learned routing policy to the number of activated exper

1 Introduction

Analysis of: Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers. Research goal: Can AnyExperts' dynamic expert allocation maintain consistent accuracy improvements over dense baselines when scaling from 8 to 64 experts on challenging reasoning tasks like those found in ScienceQA and ARO datasets?.

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. 6 claims extracted, 6 verified. Tribunal: 7.5/10 → 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
Gigantic transformers exhibit severe collapse evidenced by a high degree of parameter redundancy.	✓	0.23
Sparsely-activated Mixture-of-Experts (SMoEs) are prone to redundant experts due to representational collapse.	✓	0.24
SMoEs suffer from poor expert scalability for inference and downstream fine-tuning due to overfitting of the learned rou	✓	0.32
SMoE-Dropout consists of a randomly initialized and fixed router network to activate experts and gradually increases the	✓	0.36
Transformers trained by SMoE-Dropout exhibit a self-slimmable property subject to resource availability.	✓	0.30
SMoE-Dropout offers smooth and consistent performance boosts with an increase in activated experts during inference or f	✓	0.29

References

- https://www.semanticscholar.org/paper/1462a0e5b7db47301bb0995db56426e1f4a0ac7d
- http://arxiv.org/abs/2511.18314v1
- https://www.semanticscholar.org/paper/0ff36f535ac5c37507fd84ecdfd9fb1be849d9c4