SOVEREIGN: How does varying the number of active experts in a Mixture-of-Experts Transformer affect pass@k accuracy on Hu

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Mixture-of-experts (MoE) models that employ sparse activation have demonstrated effectiveness in significantly increasing the number of parameters while maintaining low computational requirements per token. However, recent studies have established that MoE models are inherently parameter-inefficient as the improvement in performance diminishes with an increasing number of experts. We hypothesize this parameter inefficiency is a result of all experts having equal capacity, which may not adequately meet the varying complexity requirements of different tokens or tasks. In light of this, we propos

1 Introduction

Analysis of: Towards Being Parameter-Efficient: A Stratified Sparsely Activated Transformer with Dynamic Capacity. Research goal: How does varying the number of active experts in a Mixture-of-Experts Transformer affect pass@k accuracy on HumanEval Pro and MBPP Pro compared to dense models of equivalent total parameter count?.

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

2 papers retrieved. 5 claims extracted, 1 verified. Tribunal: 6.3/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
SMoE outperforms multiple state-of-the-art MoE models with the same or fewer parameters on three multilingual machine tr	✓	0.38
Switch Transformer only outperforms T5 by an average of 0.7 on the GLUE benchmark despite being 35× larger	×	0.03
A MoE model with 20 times more parameters offers an average improvement of 0.3 BLEU on its ablation dataset compared to	×	0.04
The dense model architecture used as backbone consists of a Transformer model with 12 layers (6 on encoder and 6 on deco	×	0.03
In SMoE, each expert processes at most 2 × Ti/Ei tokens where Ti is the number of tokens in the mini-batch sent to layer	×	0.08

References

- https://arxiv.org/abs/2305.02176
- https://www.semanticscholar.org/paper/ad82ba4ce47672e2d30890e1c684cf8efbe3bb07