SOVEREIGN: What is the inference efficiency trade-off (throughput vs accuracy) of SMoES compared to modality-agnostic MoE

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

A pivotal advancement in the progress of large language models (LLMs) is the emergence of the Mixture-of-Experts (MoE) LLMs. Compared to traditional LLMs, MoE LLMs can achieve higher performance with fewer parameters, but it is still hard to deploy them due to their immense parameter sizes. Different from previous weight pruning methods that rely on specifically designed hardware, this paper mainly aims to enhance the deployment efficiency of MoE LLMs by introducing plug-and-play expert-level sparsification techniques. Specifically, we propose, for the first time to our best knowledge, post-tr

1 Introduction

Analysis of: Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models. Research goal: What is the inference efficiency trade-off (throughput vs accuracy) of SMoES compared to modality-agnostic MoE routing on ChartQA when varying expert count across different VLM backbone sizes?.

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

12 papers retrieved. 10 claims extracted, 2 verified. Tribunal: 5.0/10 → 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
MoE LLMs achieve a reduction in on-the-fly (active) parameters by choosing only top-k experts for the inference of each	×	0.12
The eight experts constitute around 96% (45B out of 47B) of the total number of parameters in the Mixtral 8x7B model.	×	0.03
Loading the Mixtral 8x7B model in bf16 format requires at least two A100-80G GPUs.	×	0.01
Unlike existing post-training weight pruning schemes for LLMs, our approach focuses on expert-level sparsity for model s	×	0.12
Our proposed method significantly reduces memory usage for deploying MoE LLMs and enhances their inference speed.	×	0.10
We introduce hardware-friendly post-training methods for either permanently removing unimportant experts (expert pruning	×	0.13
We examine expert-level pruning for both task-agnostic and task-specific models.	✓	0.16
In the decoder-only sparse MoE Transformer models, the Feed-Forward Network (FFN) sublayers are replaced with MoE layer	×	0.06
Each token x in the input sequence is routed to the top-2 experts based on the routing weights w in the Mixtral 8x7B mod	×	0.04
This work is the first to systematically explore expert-level sparsity in MoE LLMs and introduce post-training methods f	✓	0.17

References

- http://arxiv.org/abs/2504.13275v4
- http://arxiv.org/abs/2402.14800v2
- http://arxiv.org/abs/2603.11114v1