SOVEREIGN: Does the Lynx token scheduling approach generalize to other sparse MoE architectures (e.g., Mixtral 8x7B, Deep

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recent large language models such as Gemini-1.5, DeepSeek-V3, and Llama-4 increasingly adopt Mixture-of-Experts (MoE) architectures, which offer strong efficiency-performance trade-offs by activating only a fraction of the model per token. Yet academic researchers still lack a fully open, end-to-end MoE platform for investigating scaling, routing, and expert behavior. We release FLAME-MoE, a completely open-source research suite composed of seven decoder-only models, ranging from 38M to 1.7B active parameters, whose architecture–64 experts with top-8 gating and 2 shared experts–closely refle

1 Introduction

Analysis of: FLAME-MoE: A Transparent End-to-End Research Platform for Mixture-of-Experts Language Models. Research goal: Does the Lynx token scheduling approach generalize to other sparse MoE architectures (e.g., Mixtral 8x7B, DeepSeek-MoE) for document-based QA tasks, and what is the accuracy-throughput trade-off relative to static expert routing methods?.

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. 9 claims extracted, 1 verified. Tribunal: 3.5/10 → 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
FLAME-MoE significantly outperforms the dense counterparts with the same pretraining FLOPs on almost every task.	×	0.06
FLAME-MoE achieves more than 3 points of average accuracy improvements over dense baselines under both 8.0e19 and 2.4e20	×	0.10
FLAME-MoE can match or even outperform dense models trained with 2x FLOPs (e.g., in 400M-4x).	×	0.06
FLAME-MoE includes seven decoder-only MoE models (38M–1.7B active parameters), each with 64 experts per layer, top-8 gat	✓	0.21
FLAME-MoE is the only MoE platform offering full openness—code, data, checkpoints, routing logs, and evaluation results—	×	0.13
Empirical evaluations on 6 downstream tasks show that FLAME-MoE consistently outperforms dense counterparts trained unde	×	0.08
Expert specialization emerges early and intensifies over time; expert co-activation occurs.	×	0.04
Increasing EP generally improves utilization and reduces latency, while deeper pipeline parallelism (e.g., PP=2) can fur	×	0.03
The overall FLOPs throughput of MoE models still lags behind dense models due to inherent sparsity.	×	0.05

References

- http://arxiv.org/abs/2505.20225v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2401.04088v1