SOVEREIGN: FLAME-MoE: A Transparent End-to-End Research Platform for Mixture-of-Experts Lan

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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: What is the impact of token scheduling in ExpertFlow on attribute binding accuracy (e.g., on AMBER) relative to dense baselines under varying expert activation budgets in MoE vision-language models?.

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. 8 claims extracted, 8 verified. Tribunal: 7.8/10 → 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
FLAME-MoE is a completely open-source research suite composed of seven decoder-only models, ranging from 38M to 1.7B act	✓	0.35
The architecture of FLAME-MoE uses 64 experts with top-8 gating and 2 shared experts.	✓	0.19
All training data pipelines, scripts, logs, and checkpoints are publicly available.	✓	0.25
Across six evaluation tasks, FLAME-MoE improves average accuracy by up to 3.4 points over dense baselines trained with i	✓	0.30
Experts increasingly specialize on distinct token subsets.	✓	0.21
Co-activation matrices remain sparse, reflecting diverse expert usage.	✓	0.21
Routing behavior stabilizes early in training.	✓	0.19
All code, training logs, and model checkpoints are available at https://github.com/cmu-flame/FLAME-MoE.	✓	0.33

References

- http://arxiv.org/abs/2603.11114v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2505.20225v1