SOVEREIGN: How robust is ExpertFlow's token scheduling in MoE vision-language models to distribution shifts in attribute

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: How robust is ExpertFlow's token scheduling in MoE vision-language models to distribution shifts in attribute binding tasks on AMBER across different expert activation budgets compared to dense baselines?.

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, 1 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
FLAME-MoE significantly outperforms the dense counterparts with the same pretraining FLOPs on almost every task.	×	0.05
The advantages of FLAME-MoE are more pronounced at larger scales, with more than 3 points of average accuracy improvemen	×	0.08
FLAME-MoE can match or even outperform dense models trained with 2x FLOPs.	×	0.07
Increasing EP generally improves utilization and reduces latency for FLAME-MoE models.	×	0.05
Deeper pipeline parallelism (e.g., PP=2) can further enhance scalability.	×	0.01
FLAME-MoE models achieve better training efficiency, achieving a better speed-quality frontier.	×	0.06
FLAME-MoE is a transparent, robust platform for controlled experimentation across model scales, sparsity levels, and arc	×	0.10
FLAME-MoE is the only MoE platform offering full openness—code, data, checkpoints, routing logs, and evaluation results—	×	0.13
FLAME-MoE includes seven decoder-only MoE models (38M–1.7B active parameters), each with 64 experts per layer, top-8 gat	✓	0.21
Empirical evaluations on 6 downstream tasks show that FLAME-MoE consistently outperforms dense counterparts trained unde	×	0.07

References

- http://arxiv.org/abs/2505.20225v1
- http://arxiv.org/abs/2602.09258v1
- http://arxiv.org/abs/2410.17954v2