Sparse Mixture-of-Experts Accuracy and Throughput Trade-offs in Multimodal VQA Benchmarks

Assignee Research

May 29, 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

1 Introduction

This paper examines: FLAME-MoE: A Transparent End-to-End Research Platform for Mixture-of-Experts Language Models. Research question: How does the accuracy of sparse Mixture-of-Experts multimodal models on VQAv2 and OK-VQA benchmarks degrade when the number of active experts per token is reduced below the standard threshold, and what is the corresponding inference throughput gain?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

4 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
FLAME-MoE significantly outperforms dense counterparts with the same pretraining FLOPs on almost every task.	×	0.03
FLAME-MoE achieves more than 3 points of average accuracy improvements over dense baselines under both 8.0e19 and 2.4e20	×	0.08
FLAME-MoE matches or outperforms dense models trained with 2x FLOPs in the 400M-4x configuration.	×	0.05
FLAME-MoE substantially improves pretraining efficiency, achieving a better speed-quality frontier.	×	0.03
Increasing expert parallelism (EP) generally improves utilization and reduces latency in FLAME-MoE.	×	0.04
Deeper pipeline parallelism (e.g., PP=2) can further enhance scalability in FLAME-MoE.	×	0.02
FLAME-MoE adopts the best-performing configuration of PP=1 and EP=8 for training.	×	0.05
FLAME-MoE models demonstrate great utilization under EP=8 as presented in Appendix A.	×	0.03
The overall FLOPs throughput of FLAME-MoE still lags behind dense models.	×	0.05
FLAME-MoE includes seven decoder-only MoE models (38M–1.7B active parameters), each with 64 experts per layer, top-8 gat	✓	0.20
FLAME-MoE is the only MoE platform offering full openness—code, data, checkpoints, routing logs, and evaluation results—	×	0.12
Empirical evaluations on 6 downstream tasks show that FLAME-MoE consistently outperforms dense counterparts trained unde	×	0.06

References

- http://arxiv.org/abs/2410.07348v1
- http://arxiv.org/abs/2605.15484v1
- http://arxiv.org/abs/2505.20225v1