SOVEREIGN: Does token scheduling in sparse MoE inference improve tokens-per-second throughput for document-based question

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Selective parameter activation provided by Mixture-of-Expert (MoE) models have made them a popular choice in modern foundational models. However, MoEs face a fundamental tension when employed for serving. Batching, critical for performance in serving, forces the activation of all experts, thereby negating MoEs'benefits and exacerbating memory bandwidth bottlenecks. Existing work on efficient MoE inference are unable to resolve this tension even with extensive workload-specific tuning. We present LYNX, a system that enables efficient MoE inference in a workload-agnostic fashion. LYNX leverages

1 Introduction

Analysis of: Lynx: Enabling Efficient MoE Inference through Dynamic Batch-Aware Expert Selection. Research goal: Does token scheduling in sparse MoE inference improve tokens-per-second throughput for document-based question answering tasks (DocVQA, ChartQA) on dense vs. sparse 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

7 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.5/10 → 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
LYNX achieves up to 1.30x improvement in throughput while maintaining accuracy loss of less than 1% points across tasks.	✓	0.25
LYNX is complementary to existing techniques where it additionally boosts their performance by up to 1.38x.	✓	0.26
LYNX leverages a key property of MoE training: load-balancing losses introduce batch-level expert activation skews and r	✓	0.37
LYNX exploits batch-level expert activation skews and redundancy by remapping low-affinity token-to-expert assignments w	✓	0.39
Batching forces the activation of all experts, thereby negating MoEs' benefits and exacerbating memory bandwidth bottlen	✓	0.29
Existing work on efficient MoE inference are unable to resolve this tension even with extensive workload-specific tuning	✓	0.36

References

- https://www.semanticscholar.org/paper/9df454cc131c958d15826a4da36beb97701cd8af
- https://www.semanticscholar.org/paper/ccd794405e9612ab517ccfc54ed94577d5373f6f
- https://www.semanticscholar.org/paper/38f21e3165af040ae48d9f23a5108e4a005cc079