SOVEREIGN: How does dynamic batch-aware expert selection in MoE inference affect token-per-second throughput compared to

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: How does dynamic batch-aware expert selection in MoE inference affect token-per-second throughput compared to dense models of equivalent total parameter count on DocVQA and ChartQA benchmarks under varying batch sizes?.

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. 15 claims extracted, 1 verified. Tribunal: 2.7/10 → 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
LYNX is the first system to enable efficient MoE inference through dynamic batch-aware expert selection.	✓	0.26
Existing techniques for reducing data movement in MoE models depend on extensive offline calibration that assumes expert	×	0.10
Existing techniques permanently alter the model: experts are discarded, merged, or compressed at compile time and cannot	×	0.06
An MoE with E experts each with P parameters generally underperforms a dense model with E × P parameters.	×	0.05
For larger batch sizes, MoE performance converges to a large dense model, but at lower accuracy.	×	0.07
Qwen3-30B (MoE, 3B active) has 100 unique experts per layer.	×	0.08
Qwen3-4B (dense) accuracy is 72.6.	×	0.04
Qwen3-32B (dense) accuracy is 69.5.	×	0.04
Qwen3-30B (MoE, 3B active) accuracy is 60.7.	×	0.07
At batch size 64, P99 TPOT for MoE is approximately 40 ms.	×	0.04
At batch size 64, experts per token for MoE is approximately 50/128.	×	0.09
37.8% and 30.5% are values from a table on page 5.	×	0.02
12.2% and 6.1% are values from a table on page 5.	×	0.00
58.5% and 53.1% are values from a table on page 5.	×	0.00
11.5% and 9.2% are values from a table on page 5.	×	0.00

References

- http://arxiv.org/abs/2602.07616v1
- http://arxiv.org/abs/2411.08982v3
- http://arxiv.org/abs/2511.02237v1