SOVEREIGN: Can dynamic expert caching strategies in MoE-based diffusion LLMs achieve comparable throughput to static expe

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Sparse Mixture-of-Experts (MoE) models can outperform dense large language models at similar computation by activating only a small set of experts per token. However, stacking many expert modules introduces substantial parameter memory, which makes MoE models difficult to deploy in memory-constrained environments such as single-GPU devices. Offloading alleviates this issue by storing inactive experts in CPU memory and loading them on demand, but existing methods remain limited: static caches disregard input-dependent routing, and methods that train separate models to predict expert usage ahead

1 Introduction

Analysis of: ExpertFlow: Efficient Mixture-of-Experts Inference via Predictive Expert Caching and Token Scheduling. Research goal: Can dynamic expert caching strategies in MoE-based diffusion LLMs achieve comparable throughput to static expert allocation on A100 GPUs while maintaining perplexity within 1% on the Natural Questions benchmark?.

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. 7 claims extracted, 0 verified. Tribunal: 4.2/10 → 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
Our experiments were conducted on a single NVIDIA A40 GPU with 48 GB of memory and Intel(R) Xeon(R) Gold 6338 CPU @ 2.00	×	0.07
Cache-MoE maintains a fixed per-layer expert cache with LRU replacement, falling back to CPU on misses.	×	0.06
SE-MoE preloads experts for multiple layers and employs ring scheduling to overlap compute and data movement.	×	0.03
Pregated-MoE trains MLP-based routers to select experts without runtime gating.	×	0.04
ExpertFlow achieves 1.85x to 5.27x speedup over Cache-MoE depending on model/dataset.	×	0.04
Qwen1.5-MoE has 27 layers, 2.80B total parameters and 16.40B activated parameters.	×	0.03
Mixtral-8×7B has 32 layers, 12.90/46.70 B total/active parameters, 2/8 activated experts per token ratio.	×	0.04

References

- http://arxiv.org/abs/2506.14646v2
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2406.17716v3