SOVEREIGN: Can expert caching mechanisms in MoE models be optimized to reduce CPU-GPU memory transfer overhead while main

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Among parallel decoding paradigms, diffusion large language models (dLLMs) have emerged as a promising candidate that balances generation quality and throughput. However, their integration with Mixture-of-Experts (MoE) architectures is constrained by an expert explosion: as the number of tokens generated in parallel increases, the number of distinct experts activated grows nearly linearly. This results in substantial memory traffic that pushes inference into a memory-bound regime, negating the efficiency gains of both MoE and parallel decoding. To address this challenge, we propose Dynamic Exp

1 Introduction

Analysis of: Dynamic Expert Sharing: Decoupling Memory from Parallelism in Mixture-of-Experts Diffusion LLMs. Research goal: Can expert caching mechanisms in MoE models be optimized to reduce CPU-GPU memory transfer overhead while maintaining inference throughput gains on A100/H100 GPUs?.

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. 4 claims extracted, 4 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
Diffusion large language models (dLLMs) have emerged as a promising candidate that balances generation quality and throu	✓	0.30
As the number of tokens generated in parallel increases in MoE dLLMs, the number of distinct experts activated grows nea	✓	0.27
Dynamic Expert Sharing (DES) reduces unique expert activations by over 55% and latency by up to 38%, while retaining 99%	✓	0.34
DES effectively decouples memory overhead from the degree of parallelism.	✓	0.15

References

- http://arxiv.org/abs/2602.03495v1
- http://arxiv.org/abs/2602.00879v1
- http://arxiv.org/abs/2410.17954v2