SOVEREIGN: To what extent does expert-level quantization (e.g., INT8 vs FP16) combined with CPU-GPU expert caching reduce

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: To what extent does expert-level quantization (e.g., INT8 vs FP16) combined with CPU-GPU expert caching reduce memory transfer overhead in MoE diffusion LLMs, and what is the trade-off in generation accuracy on the GSM8K and HumanEval benchmarks?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 4.8/10 → 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.

References

- http://arxiv.org/abs/2605.23078v1
- http://arxiv.org/abs/2602.00879v1
- http://arxiv.org/abs/2502.05370v2