SOVEREIGN: How does the expert explosion phenomenon in MoE diffusion LLMs scale with the number of parallel tokens genera

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: How does the expert explosion phenomenon in MoE diffusion LLMs scale with the number of parallel tokens generated (e.g., 64 vs 128 tokens) and what is the resulting impact on inference latency and memory bandwidth utilization on H100 GPUs measured against the Llama-2-70B baseline?.

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

9 papers retrieved. 12 claims extracted, 0 verified. Tribunal: 4.0/10 → 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
The experiments were conducted on a single NVIDIA A40 GPU with 48 GB of memory and Intel(R) Xeon(R) Gold 6338 CPU @ 2.00	×	0.06
Cache-MoE maintains a fixed per-layer expert cache with LRU replacement, falling back to CPU on misses.	×	0.05
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 5.86× throughput improvement over Cache-MoE for Mixtral-8 on XSUM task.	×	0.03
ExpertFlow achieves 2.04× throughput improvement over SE-MoE for Switch-32 on WMT16 task.	×	0.03
ExpertFlow achieves 2.01× throughput improvement over Pregated-MoE for Switch-64 on WMT16 task.	×	0.03
ExpertFlow achieves 3.19× throughput improvement over Cache-MoE for Qwen1.5 on Alpaca task.	×	0.03
ExpertFlow achieves 1.99× throughput improvement over SE-MoE for Deepseek-MoE on Alpaca task.	×	0.04
Switch Transformer models used: Switch-32, Switch-64, and Switch-128.	×	0.02
Models used in evaluation: Qwen1.5-MoE, Deepseek-MoE, Mixtral-8×7B, and Switch Transformer with 32, 64, and 128 experts.	×	0.03
Routing path predictors are trained on the same dataset used for inference (in-domain) or on a different dataset (cross-	×	0.05

References

- http://arxiv.org/abs/2402.14800v2
- http://arxiv.org/abs/2410.17954v2

- http://arxiv.org/abs/2603.11114v1