SOVEREIGN: How does ExpertFlow's offloading and caching mechanism compare to static cache baselines in terms of inference

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 ExpertFlow's offloading and caching mechanism compare to static cache baselines in terms of inference throughput and POPE hallucination scores when evaluated on larger MoE-VLM architectures such as Mixtral 8x22B or Qwen2-VL-MoE?.

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. 13 claims extracted, 0 verified. Tribunal: 3.7/10 → 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
Experiments were conducted on a single NVIDIA A40 GPU with 48 GB of memory and Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	×	0.06
The expert-to-total parameter ratio (E.P) for Mixtral-8 is 2/8	×	0.07
The expert-to-total parameter ratio (E.P) for Switch-32 is 1/32	×	0.02
The expert-to-total parameter ratio (E.P) for Switch-64 is 1/64	×	0.02
The expert-to-total parameter ratio (E. P) for Switch-128 is 1/128	×	0.02
The expert-to-total parameter ratio (E.P) for Qwen1.5 is 4/60	×	0.02
The expert-to-total parameter ratio (E.P) for Deepseek-MoE is 6/64	×	0.04
Cache-MoE [7] maintains a fixed per-layer expert cache with LRU replacement	×	0.07
SE-MoE [35] preloads experts for multiple layers and employs ring scheduling to overlap compute and data movement	×	0.04
Pregated-MoE [12] trains MLP-based routers to select experts without runtime gating	×	0.05
Mixtral-8 achieves an expert-to-total parameter ratio (E.P) of 2.04	×	0.05
Qwen1.5 has an expert-to-total parameter ratio (E.P) of 1.85	×	0.02
Deepseek-MoE achieves an expert-to-total parameter ratio (E.P) of 4.14	×	0.03

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2510.26730v1