SOVEREIGN: How does the inference throughput of MoE-based multimodal streaming recommenders compare to dense baselines (e

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 inference throughput of MoE-based multimodal streaming recommenders compare to dense baselines (e.g., DeepFM, DCN) under varying numbers of active experts on the Amazon Reviews and MovieLens-20M benchmarks?.

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. 6 claims extracted, 1 verified. Tribunal: 6.3/10 → REVISE (revision_round=1). 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	×	0.06
We evaluate on four datasets: Alpaca for chat, WMT16 for translation, XSUM for summarization, and AIME2024 for problem s	×	0.02
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.	×	0.04
Pregated-MoE trains MLP-based routers to select experts without runtime gating.	×	0.04
ExpertFlow uses predictive expert caching and token scheduling for efficient MoE inference.	✓	0.22

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2412.11557v1
- http://arxiv.org/abs/2508.05993v3