SOVEREIGN: What is the impact of ExpertFlow's token scheduling policy on downstream task accuracy (e.g., VQA v2, MMBench)

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: What is the impact of ExpertFlow's token scheduling policy on downstream task accuracy (e.g., VQA v2, MMBench) across varying expert cache sizes and memory budgets for MoE-VLMs?.

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

11 papers retrieved. 8 claims extracted, 0 verified. Tribunal: 5.0/10 → 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
ExpertFlow achieves 1.85x to 5.86x throughput improvement over baselines on Switch Transformer models.	×	0.05
ExpertFlow achieves 1.85x to 2.16x throughput improvement over Cache-MoE on Mixtral-8 and Qwen1.5 models.	×	0.05
SE-MoE overlaps compute and data movement using ring scheduling.	×	0.02
Cache-MoE uses fixed per-layer expert cache with LRU replacement.	×	0.06
Pregated-MoE uses MLP-based routers to select experts without runtime gating.	×	0.05
ExpertFlow pre-schedules experts to GPU based on predicted routing paths.	×	0.09
ExpertFlow can correct mispredictions during execution and perform prioritized swaps.	×	0.04
The expert cache engine (ECE) workflow includes expert pre-scheduling, prefetching, and offloading based on predicted ro	×	0.09

References

- http://arxiv.org/abs/2605.22903v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2603.11114v1