SOVEREIGN: How does the latency-accuracy trade-off of Qwen3's dynamic expert specialization compare to fixed routing in M

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Mixture-of-Experts (MoE) networks promise favorable accuracy-compute trade-offs, yet practical vision deployments are hindered by expert collapse and limited end-to-end efficiency gains. We study when sparse top-k routing with hard capacity constraints helps in vision classification, evaluated under multi-seed protocols on four benchmarks (CIFAR-10/100, Tiny-ImageNet, ImageNet-1K). We observe a \emph{compute-leverage pattern}: positive accuracy gaps require a substantial fraction ρ of total FLOPs to be routed; at ImageNet scale this is necessary but not sufficient, as multi-expert routing

1 Introduction

Analysis of: When Does Sparse MoE Help in Vision? The Role of Backbone Compute Leverage in Sparse Routing. Research goal: How does the latency-accuracy trade-off of Qwen3's dynamic expert specialization compare to fixed routing in MoE models on multi-step reasoning tasks across varying batch sizes on the GQA benchmark?.

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. 0 claims extracted, 0 verified. Tribunal: 2.2/10 → 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/2604.23996v1
- http://arxiv.org/abs/2605.15484v1
- http://arxiv.org/abs/2603.11114v1