SOVEREIGN: What is the trade-off between accuracy and tokens-per-second on the GSM8K benchmark for Qwen3 under dynamic ex

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Mixture-of-Experts (MoE) has become a practical architecture for scaling LLM capacity while keeping per-token compute modest, but deploying MoE models on a single, memory-limited GPU remains difficult because expert weights dominate the HBM footprint. Existing expert offloading and prefetching systems reduce the resident set, yet they often pay expert-loading costs on the critical path when activation becomes dense. Post-training quantization (PTQ) lowers the footprint without transfers, but prevailing pipelines fix expert bit-widths offline and assume routing remains stable, even though MoE e

1 Introduction

Analysis of: Dynamic Expert Quantization for Scalable Mixture-of-Experts Inference. Research goal: What is the trade-off between accuracy and tokens-per-second on the GSM8K benchmark for Qwen3 under dynamic expert allocation compared to top-1 routing with varying expert capacity factors?.

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

7 papers retrieved. 9 claims extracted, 0 verified. Tribunal: 0.8/10 $\rightarrow$ 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
DynaExq is evaluated on Qwen3-30B-A3B-Instruct, Qwen3-80B-A3B-Instruct, and Phi-3.5-MoE-instruct models.	×	0.03
Task performance is reported on WikiText-2 (perplexity), MMLU-Pro, GPQA, AIME25, GSM8K, and HumanEval benchmarks.	×	0.05
All experiments are run on a single RTX A6000 (48 GB) GPU.	×	0.04
Hot experts are assigned a higher-precision version (usually FP16, while INT4 for Qwen3-80B), and cold experts remain in	×	0.09
On Qwen3-MoE-30B, static Int4 reduces the average score from 65.29 (FP16) to 63.98.	×	0.02
DynaExq is compared against a static quantization baseline (Int4 for Qwen3-30B and Phi-3.5-MoE; Int2 for Qwen3-80B) and	×	0.08
Performance metrics include prefill time-to-first-token (TTFT), decode time-per-output-token (TPOP), end-to-end request	×	0.03
Offloading and prefetching are most effective when each iteration touches a small, stable working set of experts.	×	0.09
MoE activation can become substantially denser during prefill and at larger batch sizes, which expands the per-iteration	×	0.04

References

- `http://arxiv.org/abs/2402.14800v2`
- `https://arxiv.org/abs/2511.15015`
- `http://arxiv.org/abs/2601.21337v2`