Qwen3 Mixture-of-Experts and Dense Models: Latency and Throughput Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the inference latency and token throughput of Qwen3 MoE architectures compare to dense models of similar parameter counts on standard LLM evaluation suites. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MoE-Lightning: High-Throughput MoE Inference on Memory-constrained GPUs. Research question: How does the inference latency and token throughput of Qwen3 MoE architectures compare to dense models of similar parameter counts on standard LLM evaluation suites?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

14 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 5.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
MoE-Lightning achieves higher throughput with far less CPU memory, enabled by CGOPipe and HRM.	✓	0.19
Existing solutions for memory-constrained batch inference often fall short of effectively overlapping computations with	×	0.10
The GPU may remain idle as it awaits a small yet crucial piece of data such as intermediate results for the upcoming bat	×	0.03
Transferring the weights for subsequent layers may take a long time and potentially block both the GPU and CPU from proc	×	0.04
Increasing I/O utilization and other resource utilization is critical in achieving high throughput.	×	0.14
When a layer's weights are loaded onto the GPU, a common strategy to increase throughput is to process as many requests	×	0.04
Lower I/O utilization means higher I/O overhead of weights' transfer, requiring greater CPU memory to reach peak generat	×	0.08
Existing solutions tend to generate sub-optimal policies with smaller GPU batch sizes which lead to resource under-utili	×	0.07
Existing solutions fail to take into account that changes in the workload can lead to changes in the bottleneck resource	×	0.05
MoE-Lightning is a new inference system developed to address challenges in achieving high-throughput inference with limi	✓	0.16

References

- http://arxiv.org/abs/2304.11414v1
- http://arxiv.org/abs/2411.11217v1
- http://arxiv.org/abs/2405.15052v2