Multi-Query Attention Throughput-Latency Trade-offs in Code Models at Scale

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the throughput-latency trade-off of multi-query attention in code models scale with batch size on the MBPP benchmark during inference. 5 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference. Research question: How does the throughput-latency trade-off of multi-query attention in code models scale with batch size on the MBPP benchmark during inference?.

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 6.6/10.

3 Results

14 papers retrieved. 5 claims extracted; 2 independently verified. Quality review score: 6.6/10.

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
ShadowKV can reduce the GPU memory footprint of the KV cache by over 6$\times$ without accuracy degradation on a wide range of	✓	0.19
ShadowKV supports 6$\times$ larger batch sizes and increases the inference throughput by up to 3.04$\times$ without compromising model	×	0.12
ShadowKV outperforms other methods and maintains the accuracy on benchmarks including RULER and LongBench.	×	0.11
ShadowKV can support 6$\times$ larger batch sizes and boost throughput by up to 3.04$\times$ compared to small batches on an A100 usin	✓	0.18
ShadowKV increases throughput up to 2.97$\times$ across different models and context lengths.	×	0.05

References

- http://arxiv.org/abs/2602.00426v1
- http://arxiv.org/abs/2602.06072v1
- http://arxiv.org/abs/2410.21465v3