Multi-Query Attention in 15B-Parameter Code Models: Throughput and Latency on The Stack

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does multi-query attention in 15B parameter code models affect throughput and latency compared to standard attention mechanisms during large-batch inference on The Stack dataset. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs. Research question: How does multi-query attention in 15B parameter code models affect throughput and latency compared to standard attention mechanisms during large-batch inference on The Stack dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.0/10.

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.

References

- http://arxiv.org/abs/2507.14301v1
- http://arxiv.org/abs/2401.10065v3
- http://arxiv.org/abs/2602.00426v1