Sliding Window Attention and Full Attention Retrieval Accuracy on Long-Context Code Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does sliding window attention impact retrieval accuracy on the Needle In A Haystack benchmark for code contexts exceeding 32k tokens compared to full attention. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Block Sparse Flash Attention. Research question: How does sliding window attention impact retrieval accuracy on the Needle In A Haystack benchmark for code contexts exceeding 32k tokens compared to full attention?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.3/10.

3 Results

8 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.3/10.

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/2604.25452v1
- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2601.15305v1