Sliding Window Attention Impact on Long-Context LLM Inference Efficiency

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does sliding window attention affect inference latency and memory usage when processing context lengths exceeding 32K tokens in LLM reasoning tasks. The quadratic compute and memory costs of global self-attention severely limit its use in high-resolution images. Local attention reduces complexity by restricting attention to neighborhoods. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Hilbert-Guided Sparse Local Attention. Research question: How does sliding window attention affect inference latency and memory usage when processing context lengths exceeding 32K tokens in LLM reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.8/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.8/10.

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/2203.01882v3
- http://arxiv.org/abs/2511.05832v2
- http://arxiv.org/abs/2201.13027v2