Sliding Window Attention Effects on Long-Sequence Code LLM Throughput and Memory

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: What is the impact of sliding window attention mechanisms on inference throughput and memory usage for sequence lengths exceeding 32k tokens in code LLMs. Recent advances in language modeling have demonstrated the effectiveness of State Space Models (SSMs) for efficient sequence modeling. While hybrid architectures such as Samba and the decoder-decoder architecture, YOCO, have shown promising performance gains over Transformers. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 1.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Decoder-Hybrid-Decoder Architecture for Efficient Reasoning with Long Generation. Research question: What is the impact of sliding window attention mechanisms on inference throughput and memory usage for sequence lengths exceeding 32k tokens in code LLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 1.5/10.

3 Results

3 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 1.5/10.

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

- https://arxiv.org/abs/2604.18856
- https://www.semanticscholar.org/paper/2713cfcaec422a74fcc32f620dd5db65c32184b0
- https://arxiv.org/abs/2507.06607