Adaptive Block Sizing in Sparse Attention for LongBench Perplexity and Retrieval

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does adaptive block sizing in sparse attention mechanisms affect perplexity and retrieval accuracy on the LongBench suite compared to fixed block size baselines at 128k context lengths. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AB-Sparse: Sparse Attention with Adaptive Block Size for Accurate and Efficient Long-Context Inference. Research question: How does adaptive block sizing in sparse attention mechanisms affect perplexity and retrieval accuracy on the LongBench suite compared to fixed block size baselines at 128k context lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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/2605.12110v1
- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2601.15305v1