Block-Sparse FlashAttention and Linear Attention Perplexity on LongBench Beyond 32k Tokens

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the perplexity of Block-Sparse FlashAttention models compare to linear attention variants on the LongBench suite for documents exceeding 32k tokens. 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: Efficient Long-Context Modeling in Diffusion Language Models via Block Approximate Sparse Attention. Research question: How does the perplexity of Block-Sparse FlashAttention models compare to linear attention variants on the LongBench suite for documents exceeding 32k tokens?.

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 6.3/10.

3 Results

10 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/2205.14135v2
- http://arxiv.org/abs/2601.15305v1
- http://arxiv.org/abs/2605.19726v1