

Block-Sparse FlashAttention Throughput vs. Reformer and Linformer at 128K-Token Scale

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the throughput of Block-Sparse FlashAttention compare to Reformer and Linformer on the ConvAITest benchmark when scaled to Llama-3-70b with context lengths exceeding 128k 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 4.7/10. 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 the throughput of Block-Sparse FlashAttention compare to Reformer and Linformer on the ConvAITest benchmark when scaled to Llama-3-70b with context lengths exceeding 128k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

9 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.7/10.

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/2512.07011v1>
- <http://arxiv.org/abs/2601.15305v1>
- <http://arxiv.org/abs/2408.11848v2>