# Block-Sparse FlashAttention vs. Sparse Mechanisms: Throughput and Memory at 100K-Token Scale

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does Block-Sparse FlashAttention compare to other sparse attention mechanisms (e.g., LongNet, H3) in terms of throughput and memory efficiency on the PG-19 benchmark when scaling to 100k 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 6.8/10. 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 Block-Sparse FlashAttention compare to other sparse attention mechanisms (e.g., LongNet, H3) in terms of throughput and memory efficiency on the PG-19 benchmark when scaling to 100k tokens?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 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/2512.07011v1
- http://arxiv.org/abs/2510.18830v2
- http://arxiv.org/abs/2307.02486v2