Block-Sparse FlashAttention Latency and Speedup in Long-Context Prefilling

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the top-k block selection latency of Block-Sparse FlashAttention scale relative to sliding window attention on the XSum dataset when context lengths exceed 64k tokens. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sparser Block-Sparse Attention via Token Permutation. Research question: How does the top-k block selection latency of Block-Sparse FlashAttention scale relative to sliding window attention on the XSum dataset when context lengths exceed 64k 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 5.1/10.

3 Results

12 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 5.1/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Llama-3.1-8B(128K) and Qwen-2.5-7B-1M(1M) are state-of-the-art long-context LLMs supporting context lengths above 128K t	×	0.07
LongBench is a collection of 21 long-context understanding tasks in 6 categories with mostly real-world data, with the a	×	0.06
LongBenchv2 further scales the context length, ranging from 8K to 2M, covering various realistic scenarios.	×	0.05
RULER is a synthetic benchmark designed to systematically evaluate long-context LLMs across various context lengths.	×	0.06
PBS-Attn matches full attention on LongBench (33.98 vs. 34.08 average score) and achieves up to a 2.72$\times$ end-to-end speed	×	0.10
PBS-Attn exhibits minimal performance degradation compared to the full attention baseline while consistently surpassing	×	0.15
PBS-Attn consistently outperforms the unpermuted MeanPooling baseline, with a remarkable relative improvement of 31% in	×	0.07
PBS-Attn and PBS-Attn+ consistently outperform their unpermuted baselines, MeanPooling and XAttention, respectively, on	×	0.08
PBS-Attn achieves an end-to-end speedup of up to 2.75$\times$ in long-context prefilling.	✓	0.21

References

- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2512.10411v5

- http://arxiv.org/abs/2510.21270v2