Block Sparse Flash Attention and Dynamic Memory Allocation in Long-Context Legal Document Processing

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the trade-off in throughput and accuracy when combining BSFA's top-k block selection with dynamic memory allocation from FlowKV for long-context legal document processing in Llama-3-70b,. 17 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. 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: What is the trade-off in throughput and accuracy when combining BSFA's top-k block selection with dynamic memory allocation from FlowKV for long-context legal document processing in Llama-3-70b, benchmarked on the LongBench legal dataset?.

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 5.5/10.

3 Results

12 papers retrieved. 17 claims extracted; 4 independently verified. Quality review score: 5.5/10.

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
Block Sparse Flash Attention achieves up to 1.10$\times$ speedup on real-world reasoning tasks while maintaining 99% of baselin	✓	0.25
Block Sparse Flash Attention achieves up to 1.24$\times$ speedup for needle-in-a-haystack retrieval tasks.	✓	0.18
Block Sparse Flash Attention substantially outperforms methods that approximate attention scores.	×	0.11
Block Sparse Flash Attention provides a CUDA kernel implementation that extends FlashAttention-2.	✓	0.16
Block Sparse Flash Attention is a production-ready solution that can be immediately deployed in real-world applications.	×	0.08
Transformers use multi-head scaled dot-product attention to process sequences of tokens.	×	0.03
The computational costs of standard attention implementations include linear projections, score computation, softmax nor	×	0.03
For long sequences where N dmodel, the operations quadratic in N dominate: both the QK score computation and PV aggreg	×	0.03
In Llama-3.1-8B with dmodel = 4096 (d = 128, H = 32), processing a sequence of N = 128K tokens requires N^2dmodel $\approx$ 6.7$\times$	×	0.03
The linear projections require only Nd^2model operations.	×	0.02
Block Sparse Flash Attention partitions the query sequence into MQ = N/BM blocks of size BM and the key/value sequence	×	0.07
Block Sparse Flash Attention uses online softmax with incremental updates instead of computing and storing the full atte	×	0.06
Block Sparse Flash Attention maintains running statistics (maximum values and normalizers) that are updated incrementall	×	0.06
Block Sparse Flash Attention computes the exact attention scores between query and key blocks within FlashAttention's ti	✓	0.16
Block Sparse Flash Attention skips loading and processing value blocks whose maximum scores fall below calibrated thresh	×	0.15
Block Sparse Flash Attention exploits the observation that blocks with uniformly low scores contribute negligibly after	×	0.06
Block Sparse Flash Attention preserves exact score computation while selectively skipping value operations, achieving si	×	0.07

References

- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2508.06447v2
- http://arxiv.org/abs/2310.05276v1