FlowKV Layer-Wise Approximation Effects on Perplexity and Needle-In-A-Haystack Accuracy in Llama-3-8B

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does FlowKV's layer-wise matrix multiplication approximation affect perplexity and answer accuracy on the Needle-In-A-Haystack benchmark for Llama-3-8b at 200K context length compared to sliding. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Optimal Approximate Matrix Multiplication over Sliding Window. Research question: How does FlowKV's layer-wise matrix multiplication approximation affect perplexity and answer accuracy on the Needle-In-A-Haystack benchmark for Llama-3-8b at 200K context length compared to sliding window attention?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

4 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
In the condensed SVD of a real matrix A, the diagonal matrix Σ contains singular values arranged in non-increasing order	×	0.03
The Frobenius norm of a matrix A is equal to the square root of the sum of the squares of its singular values.	×	0.03
The spectral norm of a matrix A is equal to its largest singular value (σ1).	×	0.02
The nuclear norm of a matrix A is equal to the sum of its singular values.	×	0.02
The product of matrices A (n$\times$l) and B (m$\times$l) transposed can be expressed as the sum of the outer products of their corres	×	0.01
An algorithm yields an ε-approximation for AMM over a sliding window if the spectral norm error of the approximation is	×	0.12
It is assumed without loss of generality that the squared norms of the columns of input matrices X and Y are normalized	×	0.04
Maintaining O(1/ε) λ-snapshot structures combined with the MG-sketch algorithm achieves an ε-approximation for the frequ	×	0.05
For an element e, the difference between its true frequency f(e) and the estimated frequency f(e) derived using the λ-s	×	0.02
The SO-COD algorithm registers a snapshot and removes a direction from the sketch when the product of the norms of the c	×	0.04

References

- http://arxiv.org/abs/2206.00583v2
- http://arxiv.org/abs/2502.18830v1
- http://arxiv.org/abs/2502.17940v1