FlowKV Layer-Wise Approximation Boosts Llama-3-8B Throughput in Long-Context Inference

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of FlowKV's output-aware, layer-wise matrix multiplication approximation on the throughput of Llama-3-8b during long-context inference (200K+ tokens) compared to traditional. 9 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: Reformulating KV Cache Eviction Problem for Long-Context LLM Inference. Research question: What is the impact of FlowKV's output-aware, layer-wise matrix multiplication approximation on the throughput of Llama-3-8b during long-context inference (200K+ tokens) compared to traditional head-wise eviction methods like sliding window or block-sparse attention?.

2 Methodology

Systematic literature search across multiple databases yielded 15 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

15 papers retrieved. 9 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.

5 Extracted Claims

Claim	Verified	Confidence
The models Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B have maximum context lengths of 128K, 32K, and	×	0.05
The baselines compared against include FullKV, StreamingLLM (SLLM), SnapKV, AdaKV, CriticalKV, and CAKE.	×	0.02
The evaluation benchmarks used are LongBench, RULER, and InfiniteBench.	×	0.03
LaProx is evaluated against SOTA KV cache eviction techniques across 16 datasets in the LongBench benchmark with cache b	×	0.11
LaProx consistently outperforms previous works in nearly every LongBench's dataset, leading to significant improvements	×	0.06
The performance gap between LaProx and the baselines widens as the memory budget becomes more constrained.	×	0.05
The output of a standard MHA layer can be expressed as the sum of independent head-wise contributions.	×	0.09
The eviction score computation algorithm takes as input the query Q, KV cache (K, V), projection WO, budget Btotal, and	×	0.11
The eviction score computation involves computing attention weight and projected values, scoring tokens, and evicting to	×	0.06

References

- http://arxiv.org/abs/2605.19726v1
- http://arxiv.org/abs/2502.18830v1

- http://arxiv.org/abs/2605.07234v1