FlowKV Selective Eviction vs. KV Cache Optimizations on LongBench for LLaMA-3

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does FlowKV's selective eviction strategy compare to other KV cache optimization methods (e.g., LongNet, SmoothFormer) in terms of perplexity and answer accuracy on the LongBench suite for. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: How does FlowKV's selective eviction strategy compare to other KV cache optimization methods (e.g., LongNet, SmoothFormer) in terms of perplexity and answer accuracy on the LongBench suite for LLaMA-3 at 200K+ token contexts?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

10 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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-Instruct has a maximum context length of 128K tokens.	×	0.13
Mistral-7B-Instruct-v0.3 has a maximum context length of 32K tokens.	×	0.07
Qwen3-8B has a maximum context length of 32K tokens.	×	0.05
The evaluation benchmarks include LongBench, RULER, and InfiniteBench.	×	0.04
LaProx is evaluated against 16 datasets in the LongBench benchmark.	×	0.03
LaProx is compared against FullKV, SLLM, SnapKV, AdaKV, CriticalKV, and CAKE.	×	0.02
LaProx consistently outperforms previous works in nearly every LongBench's dataset.	×	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.	×	0.10
The eviction score computation involves calculating attention weights and projected values.	×	0.07
The eviction score computation uses a sliding window of size w for the observation window.	×	0.06
Tokens are evicted based on their importance scores, with the top Btotal tokens being retained.	×	0.08

References

- `http://arxiv.org/abs/2605.09649v1`

- http://arxiv.org/abs/2605.08840v1
- http://arxiv.org/abs/2605.07234v1