FlowKV vs. SmoothEvict and RingQKV on LongBench for Long-Context Llama-3-70B

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the accuracy of FlowKV compare to other KV cache eviction methods (e.g., SmoothEvict, RingQKV) on the LongBench benchmark when processing 200K+ tokens in Llama-3-70B. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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 the accuracy of FlowKV compare to other KV cache eviction methods (e.g., SmoothEvict, RingQKV) on the LongBench benchmark when processing 200K+ tokens in Llama-3-70B?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

14 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 experiments use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-8B models.	×	0.10
Llama-3.1-8B-Instruct provides a maximum context length of 128K tokens.	×	0.12
Mistral-7B-Instruct-v0.3 provides a maximum context length of 32K tokens.	×	0.07
Qwen3-8B provides a maximum context length of 32K tokens.	×	0.06
The method is compared against FullKV, StreamingLLM (SLLM), SnapKV, AdaKV, CriticalKV, and CAKE baselines.	×	0.02
The evaluation uses fixed absolute cache sizes rather than ratios relative to the full KV cache.	×	0.07
The main experiments include the LongBench and RULER benchmarks.	×	0.03
The evaluation includes the InfiniteBench benchmark.	×	0.02
Evaluations on LongBench were conducted across 16 datasets.	×	0.03
Cache budgets for LongBench evaluations ranged from 128 to 1024 tokens.	×	0.03
Table 1 details performance across three models at a budget of 128 tokens.	×	0.03
LaProx consistently outperforms previous works in nearly every LongBench dataset.	×	0.07
The performance gap between LaProx and 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 (Output = Σ HiW^i_	×	0.10
Algorithm 1 computes eviction scores by multiplying the L2 norm of attention weights by the L2 norm of projected values	×	0.06
In Algorithm 1, tokens within the observation window w are assigned an infinite score to prevent eviction.	×	0.03

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2605.08840v1
- http://arxiv.org/abs/2605.07234v1