Output-Aware Eviction in FlowKV Preserves Reasoning Performance Under Extreme Context Pressure

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: Does the output-aware eviction strategy in FlowKV maintain reasoning performance on the LongBench suite for Llama-3-8b better than head-wise eviction methods under extreme context pressure. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. 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: Does the output-aware eviction strategy in FlowKV maintain reasoning performance on the LongBench suite for Llama-3-8b better than head-wise eviction methods under extreme context pressure?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

5 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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.	×	0.03
Llama-3.1-8B-Instruct provides a maximum context length of 128K tokens.	×	0.04
Mistral-7B-Instruct-v0.3 provides a maximum context length of 32K tokens.	×	0.04
Qwen3-8B provides a maximum context length of 32K tokens.	×	0.04
The method is compared against FullKV, StreamingLLM (SLLM), SnapKV, AdaKV, CriticalKV, and CAKE baselines.	×	0.03
The evaluation uses fixed absolute cache sizes rather than ratios relative to the full KV cache.	×	0.08
The main experiments include the LongBench and RULER benchmarks.	×	0.02
The evaluation includes the InfiniteBench benchmark.	×	0.00
LaProx was evaluated across 16 datasets in the LongBench benchmark.	×	0.03
Cache budgets for the LongBench evaluation ranged from 128 to 1024 tokens.	×	0.03
Table 1 details performance across three models at a cache budget of 128 tokens.	×	0.05
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.	×	0.06
Standard MHA output is defined as the concatenation of all head outputs followed by a linear projection.	×	0.05
MHA computation can be exactly decomposed into a sum of independent head-wise contributions.	×	0.08
Algorithm 1 computes eviction scores using the formula p[i] = A[:, i]2 $\cdot$ H[i, :]2 for tokens outside the observation	×	0.04
In Algorithm 1, tokens within the observation window (i >= T - w) are assigned an eviction score of infinity.	×	0.03

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2508.12485v1
- http://arxiv.org/abs/2605.07234v1