Learnable Eviction in FlowKV Preserves Long-Context Code Generation Under Memory Constraints

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: Does the learnable eviction mechanism in FlowKV maintain performance on long-context code generation benchmarks like HumanEval+ under extreme memory constraints compared to static eviction policies. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CAKE: Cascading and Adaptive KV Cache Eviction with Layer Preferences. Research question: Does the learnable eviction mechanism in FlowKV maintain performance on long-context code generation benchmarks like HumanEval+ under extreme memory constraints compared to static eviction policies?.

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 4.0/10.

3 Results

10 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
CAKE achieves an approximate 48.63% reduction in peak memory usage compared to the full cache implementation with a 128K	×	0.06
CAKE demonstrates over 10$\times$ speedup in decoding latency compared to the full cache approach when processing sequences wit	×	0.14
CAKE maintains a relatively stable decoding speed by preserving a fixed amount of KV cache, resulting in significantly l	×	0.11
Methods equipped with CAKE's allocation strategy consistently improve performance across nearly all tasks compared to va	×	0.07
CAKE achieves significant overall performance gains across different eviction methods and tasks.	×	0.07
CAKE's preference-prioritized adaptive allocation strategy considers each layer's unique characteristics and adaptively	×	0.12
CAKE defines a preference metric for each layer's KV cache requirements, considering both the spatial dispersion and tem	×	0.10
CAKE focuses on the submatrix A[-Sw :, : -Sw] of A, representing a recent window of size Sw, inspired by recent research	×	0.02

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2503.12491v2
- http://arxiv.org/abs/2408.03675v2