CAKE Cascading Eviction Scales Zero-Shot Code Generation in Low-Memory Inference

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does CAKE's cascading eviction mechanism improve the zero-shot code generation performance (measured via HumanEval+ pass@1) of smaller models (e.g., 7B parameters) proportionally more than larger. 8 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: CAKE: Cascading and Adaptive KV Cache Eviction with Layer Preferences. Research question: Does CAKE's cascading eviction mechanism improve the zero-shot code generation performance (measured via HumanEval+ pass@1) of smaller models (e.g., 7B parameters) proportionally more than larger models (e.g., 70B parameters) in low-memory inference scenarios?.

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. 8 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
CAKE achieves an approximate 48.63% reduction in peak memory usage compared to the full cache implementation with a 128K	×	0.07
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	×	0.12
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 demonstrates strong compatibility with existing eviction indi	×	0.05
CAKE's preference metric for each layer's KV cache requirements considers both the spatial dispersion and temporal shift	×	0.13
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/2402.06262v2
- http://arxiv.org/abs/2503.12491v2
- http://arxiv.org/abs/2605.09649v1