Cascading Eviction vs. Full-Cache Attention Retrieval Degradation in Mistral-7B Code Documents

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the degradation in LongBench needle-in-a-haystack retrieval accuracy for code-heavy documents when applying CAKE's cascading eviction versus full-cache attention in Mistral-7B. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. 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: What is the degradation in LongBench needle-in-a-haystack retrieval accuracy for code-heavy documents when applying CAKE's cascading eviction versus full-cache attention in Mistral-7B?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

7 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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	×	0.13
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	×	0.06
CAKE achieves significant overall performance gains when compared with vanilla uniform cache allocation.	×	0.07
CAKE's preference-prioritized adaptive allocation strategy demonstrates strong compatibility with existing eviction indi	×	0.05
CAKE's allocation strategy is evaluated on LongBench datasets using Llama2-7B-Chat under Btotal of 128L and 512L.	×	0.06
CAKE's preference metric for each layer's KV cache requirements considers both the spatial dispersion and temporal shift	×	0.12
CAKE's preference score P is defined as P = H^(1/τ1) $\cdot$ V^(1/τ2), where H and V are measures of spatial dispersion and te	×	0.04
CAKE focuses on the submatrix A[-Sw :, : -Sw] of A, representing a recent window of size Sw, for calculating preference	×	0.03

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2103.16669v3
- http://arxiv.org/abs/2503.12491v2