Layer-Specific KV Cache Eviction in CAKE and Its Impact on HumanEval Code Completion

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does CAKE's layer-specific KV cache eviction impact code completion accuracy on the HumanEval benchmark compared to full-cache attention in Mistral-7B. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Make Each Token Count: Towards Improving Long-Context Performance with KV Cache Eviction. Research question: How does CAKE's layer-specific KV cache eviction impact code completion accuracy on the HumanEval benchmark compared to full-cache attention in Mistral-7B?.

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 4.7/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.7/10.

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.

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2503.12491v2
- http://arxiv.org/abs/2605.08840v1