FlowKV and SmoothFormer Integration for LongBench Accuracy in Mistral-7B Code Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Does integrating FlowKV's selective eviction with SmoothFormer's adaptive attention mechanisms improve the LongBench accuracy of Mistral-7B on code-heavy contexts (e.g., HumanEval-X) compared to. 4 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: Does integrating FlowKV's selective eviction with SmoothFormer's adaptive attention mechanisms improve the LongBench accuracy of Mistral-7B on code-heavy contexts (e.g., HumanEval-X) compared to full-cache attention, measured by accuracy decay across 100K to 200K token lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

15 papers retrieved. 4 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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	×	0.08
CAKE demonstrates over 10$\times$ speedup in decoding latency compared to the full cache approach when processing sequences wit	×	0.14
CAKE's preference-prioritized adaptive allocation strategy improves performance across nearly all tasks compared to vani	×	0.06
The preference score P for an attention layer's cache size is defined as P = H^(1/τ1) $\cdot$ V^(1/τ2), where H and V are meas	×	0.06

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2503.12491v2
- http://arxiv.org/abs/2509.10798v2