CAKE KV Cache Eviction and Performance Scaling in Large-Code Language Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: Does CAKE's KV cache eviction improve code completion performance on the MBPP benchmark when applied to larger models (e.g., Mistral-13B or Llama-3-70B), and how does the trade-off between eviction. 7 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: WKVQuant: Quantizing Weight and Key/Value Cache for Large Language Models Gains More. Research question: Does CAKE's KV cache eviction improve code completion performance on the MBPP benchmark when applied to larger models (e.g., Mistral-13B or Llama-3-70B), and how does the trade-off between eviction rate and accuracy scale with model size?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

6 papers retrieved. 7 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
WKVQuant achieves almost comparable memory savings to weight-activation quantization, while also approaching the perform	✓	0.35
The memory footprint of temporary activations is significantly smaller compared to the memory usage of weights and the K	×	0.05
Temporary activations are highly sensitive to quantization, leading to high accuracy drop.	×	0.05
In decoding phase of LLM, as the computation is bound by memory access of weights and KV cache, quantizing temporary act	×	0.06
LLaMA-13b with 13 billion weights occupies around 26GB of memory in FP16 format.	×	0.01
The memory footprint of temporary activations is relatively small, especially in decode phase.	×	0.04
The memory consumption of the KV cache increases with larger batch sizes and longer input sequences as the model needs t	×	0.06

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2503.12491v2
- http://arxiv.org/abs/2402.12065v2