Performance Gaps Between Full-Cache Inference and ReST-KV Eviction at Scale

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: Does the performance gap between full-cache inference and ReST-KV eviction increase with longer sequence lengths (e.g., 200K tokens vs. 50K tokens) on multi-turn code generation tasks, and how does. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ReST-KV: Robust KV Cache Eviction with Layer-wise Output Reconstruction and Spatial-Temporal Smoothing. Research question: Does the performance gap between full-cache inference and ReST-KV eviction increase with longer sequence lengths (e.g., 200K tokens vs. 50K tokens) on multi-turn code generation tasks, and how does this 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 3.5/10.

3 Results

6 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
ReST-KV was evaluated on five open-source LLMs: Llama2-Chat, Gemma-Instruct, Llama3-Instruct, Mistral-Instruct-v0.3, and	×	0.02
Llama2-Chat and Gemma-Instruct utilize Multi-head attention architecture.	×	0.03
Llama3-Instruct, Mistral-Instruct-v0.3, and Qwen2.5-Instruct utilize Grouped-query attention architecture.	×	0.02
ReST-KV was compared against five baseline methods: StreamingLLM, H2O, TOVA, SnapKV, and LaCache.	×	0.04
ReST-KV incorporates adaptive budget strategies from PyramidKV and AdaKV.	×	0.04
ReST-KV was evaluated on the LongBench benchmark, which covers 16 datasets across six categories.	×	0.04
ReST-KV was evaluated on the RULER benchmark, consisting of 4 categories and 13 complex tasks.	×	0.04
ReST-KV was evaluated on the Needle-in-a-Haystack benchmark.	×	0.09
ReST-KV was evaluated on the InfiniteBench benchmark, which includes 10 tasks.	×	0.05
ReST-KV achieves approximately a 36.0% reduction in peak memory usage compared to full cache at a context length of 128k	×	0.06
ReST-KV achieves an approximate 10.61$\times$ speedup in decoding latency over the full cache method at a 128K context length.	×	0.11
ReST-KV yields a Time-To-First-Token (TTFT) speedup of up to 3.42$\times$ when compatible with prefill sparse attention approac	×	0.04
ReST-KV has a computational complexity comparable to that of SnapKV.	×	0.03
KV cache reduces redundant computation in LLMs by storing previously computed keys and values.	×	0.07

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2605.08840v1