ReST-KV Spatial-Temporal Smoothing vs. Attention-Based KV Eviction in Long-Context Language Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does ReST-KV's spatial-temporal smoothing compare to other attention-based KV eviction methods (e.g., FIFO, LRU) in terms of perplexity and accuracy on LongBench at 512K context lengths. 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: LiteCache: A Query Similarity-Driven, GPU-Centric KVCache Subsystem for Efficient LLM Inference. Research question: How does ReST-KV's spatial-temporal smoothing compare to other attention-based KV eviction methods (e.g., FIFO, LRU) in terms of perplexity and accuracy on LongBench at 512K context lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 1 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

1 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

- https://openalex.org/W7106207811