Impact of Layer-wise KV Cache Reconstruction on Artificially Inflated Needle-in-a-Haystack Scores in Ultra-Long Context Tasks

Assignee Research

June 11, 2026

Abstract

Large Language Models (LLMs) require significant GPU memory when processing long texts, with the key value (KV) cache consuming up to 70\% of total memory during inference. Although existing compression methods reduce memory by evaluating the importance of individual tokens, they overlook critical semantic relationships between tokens, resulting in fragmented context and degraded performance. We introduce ChunkKV, which fundamentally reimagines KV cache compression by treating semantic chunks - rather than isolated tokens - as basic compression units. This approach preserves complete linguisti

1 Introduction

This paper examines: ChunkKV: Semantic-Preserving KV Cache Compression for Efficient Long-Context LLM Inference. Research question: To what extent does layer-wise KV cache reconstruction in methods like ReST-KV artificially inflate needle-in-a-haystack scores relative to standard eviction policies on ultra-long context tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.6/10.

3 Results

11 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.6/10.

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
Large Language Models (LLMs) require significant GPU memory when processing long texts, with the key value (KV) cache co	✓	0.35
Existing compression methods reduce memory by evaluating the importance of individual tokens, but they overlook critical	✓	0.37
ChunkKV treats semantic chunks - rather than isolated tokens - as basic compression units, preserving complete linguisti	✓	0.28
ChunkKV includes a novel layer-wise index reuse technique that exploits the higher cross-layer similarity of preserved i	✓	0.37
Comprehensive evaluations on challenging benchmarks: LongBench, Needle-In-A-HayStack, GSM8K, and JailbreakV demonstrate	✓	0.36
Semantic-aware compression significantly enhances both efficiency and performance for long-context LLM inference, provid	✓	0.38
The code for ChunkKV is available at https://github.com/NVIDIA/kvpress.	✓	0.18

References

- https://openalex.org/W7140347279
- https://openalex.org/W7162149457
- https://doi.org/10.48550/arxiv.2502.00299