PyramidKV Layer-Wise KV Cache Compression and Accuracy Trade-offs in LLaMA-3-70B

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the trade-off between output reconstruction accuracy and KV cache compression ratio in ReST-KV when evaluated on the LLaMA-3-70B model across different LongBench subsets (e.g., narrative,. 6 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling. Research question: What is the trade-off between output reconstruction accuracy and KV cache compression ratio in ReST-KV when evaluated on the LLaMA-3-70B model across different LongBench subsets (e.g., narrative, technical)?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

13 papers retrieved. 6 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
PyramidKV outperforms H2O, SnapKV, and StreamingLLM, especially in small KV cache sizes.	×	0.08
The evaluation results from LongBench are shown in Table 1 and Figure 3.	×	0.04
In Figure 3, the average score across datasets for 64, 96, 128, and 256 case sizes is reported.	×	0.01
In Table 1, results for two different KV cache sizes with 64 and 2048 are reported.	×	0.05
PyramidKV consists of two steps: dynamically allocating different KV cache sizes/budgets across different layers and sel	×	0.11
PyramidKV allocates more KV cache to the lower layers where information is more dispersed and each KV state contains les	✓	0.19

References

- http://arxiv.org/abs/2406.02069v4
- http://arxiv.org/abs/2603.22910v2
- http://arxiv.org/abs/2605.08840v1