LiteCache vs. CPU Offloading: Memory and Latency in Long-Context LLM Inference

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the performance of LiteCache's GPU-centric KV cache management compare to CPU-centric offloading strategies in terms of peak memory utilization and latency during batched inference with. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference. Research question: How does the performance of LiteCache's GPU-centric KV cache management compare to CPU-centric offloading strategies in terms of peak memory utilization and latency during batched inference with long-context LLMs (context lengths >8K tokens) on NVIDIA A100 GPUs?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.3/10.

3 Results

10 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
ShadowKV reduces the GPU memory footprint of the KV cache by over 6$\times$ without accuracy degradation on a wide range of mod	×	0.13
ShadowKV supports 6$\times$ larger batch sizes compared to baseline methods.	×	0.12
ShadowKV increases inference throughput by up to 3.04$\times$ without compromising model quality.	×	0.08
ShadowKV outperforms other methods and maintains accuracy on the RULER and LongBench benchmarks.	×	0.06
ShadowKV was evaluated on models including Llama-3-8B-1M, Llama-3.1-8B, GLM-4-9B-1M, Yi-9B-200K, Phi-3-Mini-128K, and Qw	✓	0.29
ShadowKV was evaluated using benchmarks including RULER, LongBench, and Needle In A Haystack with contexts up to 1M.	×	0.14
ShadowKV boosts throughput by up to 3.04$\times$ compared to small batches on an A100 GPU using Llama-3.1-8B.	×	0.12
ShadowKV increases throughput up to 2.97$\times$ across different models and context lengths.	×	0.07
ShadowKV outperforms prior works referenced as [23, 45] and scales better to larger KV caches.	×	0.06
During pre-filling, ShadowKV offloads the value cache to the CPU while retaining low-rank pre-RoPE keys, compressed land	×	0.10
During decoding, ShadowKV uses landmarks to select chunk indices for key cache recovery and value cache fetching.	×	0.06
ShadowKV performs accurate sparse attention computation with selected KV pairs and static outliers.	×	0.08

References

- http://arxiv.org/abs/2511.11907v2
- http://arxiv.org/abs/2410.21465v3

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