LiteCache GPU-Centric KV Cache Robustness Under Dynamic Batch Workloads on H100 GPUs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the robustness of LiteCache's GPU-centric KV cache management compare to CPU-centric offloading when handling dynamic batch sizes (1-16 concurrent requests) in terms of throughput stability. 9 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: LiteCache: A Query Similarity-Driven, GPU-Centric KVCache Subsystem for Efficient LLM Inference. Research question: How does the robustness of LiteCache's GPU-centric KV cache management compare to CPU-centric offloading when handling dynamic batch sizes (1-16 concurrent requests) in terms of throughput stability and memory fragmentation on H100 GPUs?.

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 4.2/10.

3 Results

11 papers retrieved. 9 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
LiteCache achieves comparable accuracy to baselines while sharply minimizing CPU overhead, fully utilizing PCIe bandwidt	✓	0.34
RetroInfer runs 14–79% and 15–143% slower than FullAttn and FullAttn+CuGraph, respectively.	×	0.03
PQCache performs significantly worse, lagging behind by 168–797% and 176–1001% relative to the two baselines.	×	0.01
FullAttn+CuGraph achieves a 1.01–1.75$\times$ speedup over its non-graph counterpart.	×	0.02
On A40, the ratio of cache-related time to GPU kernel execution ranges from 31% to 47% and increases with sequence lengt	×	0.07
The performance gap between existing KVCache subsystems and FullAttn+CuGraph widens significantly on modern high perform	×	0.06
PQCache consistently underperforms RetroInfer.	×	0.00
Each decoding step triggers four major CPU assisted cache operations, namely, lookup, transfer, update, and merge.	×	0.04
Update operations can overlap with data transfer and GPU attention computation, whereas the remaining operations lie dir	×	0.09

References

- http://arxiv.org/abs/2603.27138v1
- http://arxiv.org/abs/2511.14510v2

- http://arxiv.org/abs/2411.01142v1