Mixed-Precision KV Cache Storage and Throughput in Llama-2-70B Multi-Query Attention

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of mixed-precision (FP16 vs. BF16) KV cache storage on end-to-end throughput when using LiteCache's GPU-centric management versus CPU-centric offloading during multi-query. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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: What is the impact of mixed-precision (FP16 vs. BF16) KV cache storage on end-to-end throughput when using LiteCache's GPU-centric management versus CPU-centric offloading during multi-query attention inference with Llama-2-70B?.

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 4.8/10.

3 Results

11 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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 and fully utilizing PCIe bandw	✓	0.28
LiteCache improves decoding throughput by 10.7-224.2% on both H100 and A40 GPUs.	✓	0.17
LiteCache supports sequence lengths beyond 1M.	×	0.07
RetroInfer runs 14–79% slower than FullAttn and 15–143% slower than FullAttn+CuGraph.	×	0.02
PQCache performs significantly worse, lagging behind by 168–797% and 176–1001% relative to FullAttn and FullAttn+CuGraph	×	0.01
FullAttn+CuGraph achieves a 1.01–1.75$\times$ speedup over its non-graph counterpart on H100 GPU.	×	0.04
The ratio of cache-related time to GPU kernel execution ranges from 31% to 47% on A40 and increases with sequence length	×	0.06
On A40, the ratio of cache-related time to GPU kernel execution ranges from 31% to 47%.	×	0.04
On H100, the ratio of cache-related time to GPU kernel execution ranges from 45% to 68%.	×	0.04

References

- http://arxiv.org/abs/2412.19442v3
- http://arxiv.org/abs/2605.17170v1
- http://arxiv.org/abs/2511.14510v2