FlowKV Selective Eviction and Inference Throughput in LLaMA-3 Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of FlowKV's selective eviction on the inference throughput of LLaMA-3 models compared to full-cache attention, measured in tokens per second across varying sequence lengths (e.g.,. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Make Each Token Count: Towards Improving Long-Context Performance with KV Cache Eviction. Research question: What is the impact of FlowKV's selective eviction on the inference throughput of LLaMA-3 models compared to full-cache attention, measured in tokens per second across varying sequence lengths (e.g., 50K, 100K, 150K, 200K)?.

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 5.7/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.7/10.

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.

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2602.08329v1
- http://arxiv.org/abs/2503.08879v1