FlowKV, H2O, and SnapKV Inference Latency and Memory Footprint on LLaMA-3-70B with Limited VRAM

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the comparative inference latency and memory footprint of FlowKV versus H2O and SnapKV when deploying LLaMA-3-70B on consumer-grade GPUs with limited VRAM for 200K token sequences. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. 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 comparative inference latency and memory footprint of FlowKV versus H2O and SnapKV when deploying LLaMA-3-70B on consumer-grade GPUs with limited VRAM for 200K token sequences?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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/2503.11816v3
- http://arxiv.org/abs/2508.06447v2