FlowKV Learnable Eviction Strategy and Llama-3-70B Accuracy in Multi-Modal Long-Context Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of FlowKV's learnable eviction strategy on the accuracy of Llama-3-70b in multi-modal long-context tasks like MMBench, compared to full-cache attention. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLMs Know What to Drop: Self-Attention Guided KV Cache Eviction for Efficient Long-Context Inference. Research question: What is the impact of FlowKV's learnable eviction strategy on the accuracy of Llama-3-70b in multi-modal long-context tasks like MMBench, compared to full-cache attention?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.5/10.

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/2503.14478v2
- http://arxiv.org/abs/2503.08879v1