FlowKV and SmoothEvict Effects on Llama-3-70B Adversarial Robustness in Multi-Hop Reasoning

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the comparative impact of FlowKV and SmoothEvict on the adversarial robustness of Llama-3-70B during multi-hop reasoning tasks in the LongBench suite. 5 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. 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 impact of FlowKV and SmoothEvict on the adversarial robustness of Llama-3-70B during multi-hop reasoning tasks in the LongBench suite?.

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 3.0/10.

3 Results

16 papers retrieved. 5 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
DBTrimKV achieves a relative performance of 99.9% on short-form question answering tasks compared to vanilla inference.	×	0.08
DBTrimKV achieves a relative performance of 73.43% on the short-form QA tasks compared to vanilla inference.	×	0.06
TrimKV achieves 99.7% relative performance on short-form question answering tasks compared to vanilla inference.	×	0.08
The memory capacity in the implementation is set to Mglobal = 64*L*H, where L and H are the number of layers and heads.	×	0.09
The retention gate g is implemented as a small MLP with hidden dimension dg = 512.	×	0.02

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2404.14464v1