FlowKV vs. Recurrent Cache and MCTS-Based Eviction in Long-Form Dialogue Coherence

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does FlowKV's isolated KV cache management compare to state-of-the-art approaches like Recurrent Cache or MCTS-based eviction in maintaining conversational coherence on MT-bench for long-form. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FlowKV: Enhancing Multi-Turn Conversational Coherence in LLMs via Isolated Key-Value Cache Management. Research question: How does FlowKV's isolated KV cache management compare to state-of-the-art approaches like Recurrent Cache or MCTS-based eviction in maintaining conversational coherence on MT-bench for long-form technical dialogues (e.g., 50+ turns)?.

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 5.8/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.8/10.

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.

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2510.17722v2
- http://arxiv.org/abs/2505.15347v2