FlowKV Cache Management Effects on Gemma-3-12B MT-Bench Performance in Domain-Specific Dialogues

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the impact of FlowKV's KV cache management on the MT-bench conversation quality scores of Gemma-3-12B when tested on domain-specific dialogues (e.g., technical vs. casual conversations). 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: What is the impact of FlowKV's KV cache management on the MT-bench conversation quality scores of Gemma-3-12B when tested on domain-specific dialogues (e.g., technical vs. casual conversations)?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

7 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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/2402.14762v3
- http://arxiv.org/abs/2505.15347v2