

# FlowKV Enhances Accuracy and Coherence in LongBench with RAG Beyond 32k Tokens

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of FlowKV on answer accuracy and coherence in LongBench when combined with retrieval-augmented generation (RAG) for extended multi-turn conversations beyond 32k tokens. 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 on answer accuracy and coherence in LongBench when combined with retrieval-augmented generation (RAG) for extended multi-turn conversations beyond 32k tokens?.

## 2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 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/2505.00570v3>
- <http://arxiv.org/abs/2505.15347v2>