FlowKV and Sliding Window KV Cache Performance on LongBench Multi-Turn Accuracy Beyond 128K Tokens

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does FlowKV's isolated KV cache management compare to sliding window eviction on LongBench multi-turn accuracy when context exceeds 128K tokens. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: How does FlowKV's isolated KV cache management compare to sliding window eviction on LongBench multi-turn accuracy when context exceeds 128K tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

8 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
FlowKV achieves an Instruction Following Rate (IFR) of 76.15% on Turn 1 for the LLaMA model using the SKV strategy.	×	0.05
For the LLaMA model with the SKV strategy, FlowKV improves Turn 2 IFR by 24.85 percentage points compared to the Baselin	×	0.02
For the LLaMA model with the SKV strategy, FlowKV improves Turn 3 IFR by 25.56 percentage points compared to the Baselin	×	0.03
FlowKV achieves an average performance improvement of over 20% in subsequent conversation turns compared to the baseline	×	0.04
On the PrefEval benchmark, the Full KV method achieves a score of 75.4.	×	0.07
On the PrefEval benchmark, the FlowKV method achieves a score of 58.7.	×	0.03
In a 3-turn dialogue, Turn 1 Response attention is heavily focused on the Turn 1 Query (T1Q) and the local window.	×	0.03
In a 3-turn dialogue, Turn 2 Response attention is heavily focused on T1Q, T1R, and the local window.	×	0.03
Queries in Turns 2 and 3 show increased attention to previous queries and the system prompt.	×	0.03
For the Qwen model using the CKV strategy, FlowKV improves Turn 2 IFR by 28.80 percentage points compared to the Baselin	×	0.02
The core isolation mechanism of FlowKV is not engaged during the initial turn of the conversation due to the absence of	×	0.09
SnapKV and ExpectedAttention exhibit minimal performance degradation when used with FlowKV compared to their baseline co	×	0.02

References

- http://arxiv.org/abs/2601.02872v1
- http://arxiv.org/abs/2302.11081v1
- http://arxiv.org/abs/2505.15347v2