FlowKV Isolated KV Cache Management: Perplexity and Coherence in Multi-Turn Conversations

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does FlowKV's isolated KV cache management affect perplexity and response coherence in multi-turn conversations compared to VAttention and PageAttention on the LongBench dataset under. 14 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: FlowKV: Enhancing Multi-Turn Conversational Coherence in LLMs via Isolated Key-Value Cache Management. Research question: How does FlowKV's isolated KV cache management affect perplexity and response coherence in multi-turn conversations compared to VAttention and PageAttention on the LongBench dataset under high-concurrency scenarios?.

2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 papers retrieved. 14 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
On the PrefEval benchmark, the Full KV method achieves a score of 75.4.	×	0.05
On the PrefEval benchmark, the FlowKV method achieves a score of 58.7.	×	0.03
On the PrefEval benchmark, the Baseline method achieves a score of 10.6.	×	0.03
In Turn 1 responses, attention is heavily focused on the Turn 1 Query (T1Q) and the local window.	×	0.03
In Turn 2 responses, attention is heavily focused on T1Q, T1R, and the local window.	×	0.02
In Turn 3 responses, attention is heavily focused on T1Q, T1R, T2Q, T2R, and the local window.	×	0.02
Queries in Turns 2 and 3 show increased attention to previous queries and the system prompt.	×	0.02
For the LLaMA model using the SKV strategy, the Baseline Turn 2 IFR is 37.08%.	×	0.03
For the LLaMA model using the SKV strategy, FlowKV improves Turn 2 IFR by 24.85 percentage points compared to the Baseli	×	0.03
For the LLaMA model using the CKV strategy, FlowKV improves Turn 2 IFR by 40.27 percentage points compared to the Baseli	×	0.04
For the Qwen model using the SKV strategy, FlowKV improves Turn 2 IFR by 39.39 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
During the initial turn of conversation, the core isolation mechanism of FlowKV is not engaged due to the absence of pri	×	0.09
With FlowKV, SnapKV and ExpectedAttention exhibit minimal performance degradation compared to the Full KV baseline in la	×	0.06

References

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

- http://arxiv.org/abs/2504.03775v1
- http://arxiv.org/abs/2505.15347v2