Phi-3-Mini-8K and Llama-3-8B Conversational Coherence on MT-Bench with FlowKV and Traditional KV Cache Strategies

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the multi-turn conversational coherence of Phi-3-mini-8K compare to Llama-3-8B on the MT-bench benchmark when evaluated with FlowKV's isolated KV cache management versus traditional eviction. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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 the multi-turn conversational coherence of Phi-3-mini-8K compare to Llama-3-8B on the MT-bench benchmark when evaluated with FlowKV's isolated KV cache management versus traditional eviction strategies?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

11 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 LLaMA model using the SKV strategy, FlowKV achieves a Turn 2 score of 61.93%, which is an improvement of 24.85 pe	×	0.02
On the LLaMA model using the SKV strategy, FlowKV achieves a Turn 3 score of 54.95%, which is an improvement of 25.56 pe	×	0.02
On the LLaMA model using the CKV strategy, FlowKV achieves a Turn 2 score of 52.83%, representing a 40.27 percentage poi	×	0.03
On the Qwen model using the SKV strategy, FlowKV achieves a Turn 2 score of 56.72%, an improvement of 39.39 percentage p	×	0.02
FlowKV achieves an average performance improvement of over 20% in subsequent conversation turns compared to the baseline	×	0.05
During the initial turn of conversation, FlowKV's core isolation mechanism is not engaged due to the absence of prior co	×	0.10
In a 3-turn dialogue, Turn 1 Response attention is heavily focused on Turn 1 Query (T1Q) and the local window.	×	0.05
In a 3-turn dialogue, Turn 2 Response attention is heavily focused on T1Q, Turn 1 Response (T1R), and the local window.	×	0.05
Queries in Turns 2 and 3 show increased attention to previous queries and the system prompt.	×	0.03
FlowKV achieves a score of 75.4% on the Full KV metric in the reported benchmark results.	×	0.06
FlowKV achieves a score of 58.7% on the Baseline metric in the reported benchmark results.	×	0.03

References

- http://arxiv.org/abs/2510.17722v2
- http://arxiv.org/abs/2402.14762v3
- http://arxiv.org/abs/2505.15347v2