Phi-3-Mini and Gemma-3-12B Multi-Turn Dialogue Coherence on MT-Bench

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What are the comparative MT-bench conversation quality scores between Phi-3-mini and Gemma-3-12B when evaluated on multi-turn dialogue coherence. 15 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. 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 are the comparative MT-bench conversation quality scores between Phi-3-mini and Gemma-3-12B when evaluated on multi-turn dialogue coherence?.

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 4.3/10.

3 Results

7 papers retrieved. 15 claims extracted; 3 independently verified. Quality review score: 4.3/10.

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 a score of 75.4 on the Multi-Turn Benchmark.	×	0.08
The baseline method achieves a score of 58.7 on the Multi-Turn Benchmark.	×	0.07
FlowKV achieves a score of 10.9 on the KV cache compression on PrefEval.	✓	0.16
The baseline method achieves a score of 10.6 on the KV cache compression on PrefEval.	✓	0.15
FlowKV achieves a score of 75.40% on the Multi-Turn Benchmark.	×	0.11
The baseline method achieves a score of 58.70% on the Multi-Turn Benchmark.	×	0.06
FlowKV achieves a score of 10.90% on the KV cache compression on PrefEval.	✓	0.18
The baseline method achieves a score of 10.60% on the KV cache compression on PrefEval.	×	0.13
FlowKV achieves a score of 76.15% on the LLaMA model for Turn 1.	×	0.02
FlowKV achieves a score of 61.93% on the LLaMA model for Turn 2, which is an improvement of 24.85% over the baseline.	×	0.02
FlowKV achieves a score of 54.95% on the LLaMA model for Turn 3, which is an improvement of 25.56% over the baseline.	×	0.02
FlowKV achieves a score of 76.49% on the Qwen model for Turn 1.	×	0.02
FlowKV achieves a score of 56.72% on the Qwen model for Turn 2, which is an improvement of 39.39% over the baseline.	×	0.02
FlowKV achieves a score of 49.67% on the Qwen model for Turn 3, which is an improvement of 27.71% over the baseline.	×	0.02
FlowKV achieves an average performance improvement of over 20% on the Multi-IF dataset.	×	0.04

References

- http://arxiv.org/abs/2304.00180v1
- http://arxiv.org/abs/2601.06757v1

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