Instruction-Following Accuracy and Latency Trade-offs in Claude-3.5-Sonnet vs. Quantized Llama-3 on MobileAloha Multi-Turn Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the trade-off between instruction-following accuracy and inference latency when comparing Claude-3.5-Sonnet with quantized versions of Llama-3 on the Multi-Turn Robotic Instruction Following. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. 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: What is the trade-off between instruction-following accuracy and inference latency when comparing Claude-3.5-Sonnet with quantized versions of Llama-3 on the Multi-Turn Robotic Instruction Following subset of the MobileAloha dataset?.

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 4.0/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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.

References

- http://arxiv.org/abs/2601.06757v1
- http://arxiv.org/abs/2409.18216v1
- http://arxiv.org/abs/2505.15347v2