

Claude-3.5-Sonnet vs. Quantized Mobile Models in Multi-Turn Robotic Instruction Following

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the performance delta in context window utilization efficiency between Claude-3.5-Sonnet and quantized mobile models when executing multi-turn robotic instruction following on the MobileAloha. 19 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: When Models Can't Follow: Testing Instruction Adherence Across 256 LLMs. Research question: What is the performance delta in context window utilization efficiency between Claude-3.5-Sonnet and quantized mobile models when executing multi-turn robotic instruction following on the MobileAloha dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

15 papers retrieved. 19 claims extracted; 2 independently verified. Quality review score: 4.4/10.

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
The evaluation framework was implemented using the OpenRouter API, which provides unified access to multiple language models	×	0.06
The evaluation system executes each test systematically across all target models, maintaining consistent parameters to ensure	×	0.04
The temperature was set to 0.0 to minimize response variability and enable more deterministic evaluation of instruction-	×	0.08
Response timeout limits were established at 10 seconds to balance comprehensive model coverage with practical execution	×	0.04
Primary metrics include binary pass/fail determination based on adherence to specified instructions, response time measurement	×	0.03
Results are automatically compiled into a structured Excel workbook with multiple analytical views.	×	0.02
The overview sheet provides a high-level pass/fail matrix across all models and tests, enabling rapid identification of	×	0.03
Individual test sheets contain detailed response data, allowing deeper investigation of specific failure modes.	×	0.02
A model summary sheet aggregates performance statistics, including success rates, average response times, and total tokens	×	0.02
Each test prompt includes programmatically verifiable success criteria.	×	0.04
The verification approach accounts for common variations in model outputs while maintaining objective assessment standards	×	0.03
The verification process operates in two stages: primary verification applies strict matching criteria to determine exact	×	0.01
The comprehensive evaluation of instruction-following capabilities encompassed all 331 models available via OpenRouter	✓	0.19
331 models were verified for basic functionality, of which 256 passed verification and were subsequently evaluated using	×	0.10
The two-stage verification protocol first assessed basic endpoint functionality and subsequently evaluated all verified	×	0.09
The evaluation of instruction-following capabilities was conducted on October 14, 2025.	✓	0.21
The evaluation included 331 models across diverse providers and architectures.	×	0.06
The verification step used a simple factual query: 'What is the capital of France?'	×	0.02

References

- <http://arxiv.org/abs/2312.10793v3>
- <http://arxiv.org/abs/2510.18892v1>
- <http://arxiv.org/abs/2409.02392v2>