Claude-3.5-Sonnet vs. Open-Source Multimodal Models on MobileAloha Robotic Benchmark

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the performance of Claude-3.5-Sonnet compare to state-of-the-art open-source multimodal models on the MobileAloha benchmark when evaluated for instruction adherence in robotic manipulation. 17 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: How does the performance of Claude-3.5-Sonnet compare to state-of-the-art open-source multimodal models on the MobileAloha benchmark when evaluated for instruction adherence in robotic manipulation tasks?.

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 3.8/10.

3 Results

14 papers retrieved. 17 claims extracted; 1 independently verified. Quality review score: 3.8/10.

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 mo	×	0.07
The evaluation system executes each test systematically across all target models, maintaining consistent parameters to e	×	0.04
Temperature was set to 0.0 to minimize response variability and enable more deterministic evaluation of instruction-foll	×	0.09
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 measu	×	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	×	0.04
Individual test sheets contain detailed response data, allowing deeper investigation of specific failure modes.	×	0.03
A model summary sheet aggregates performance statistics, including success rates, average response times, and total toke	×	0.02
Each test prompt includes programmatically verifiable success criteria.	×	0.04
The verification process operates in two stages: primary verification applies strict matching criteria to determine exac	×	0.01
The comprehensive evaluation of instruction-following capabilities encompassed all 331 models available via OpenRouter a	✓	0.19
256 models passed verification and were subsequently evaluated using twenty diagnostic prompts.	×	0.06
The two-stage verification protocol first assessed basic endpoint functionality and subsequently evaluated all verified	×	0.09
On October 14, 2025, 331 models were available across diverse providers and architectures.	×	0.11
The verification step used a simple factual query: 'What is the capital of France?'	×	0.01
Models received temperature=0.0 and max tokens=150 where supported, and seed=42 for reproducibility where supported.	×	0.02

References

- http://arxiv.org/abs/2310.04793v2
- http://arxiv.org/abs/2510.11852v1
- http://arxiv.org/abs/2510.18892v1