Adversarial Robustness of Claude-3.5-Sonnet and Quantized Mobile Models in MobileAloha Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How robust are instruction-following capabilities of Claude-3.5-Sonnet and quantized mobile models when tested with adversarial perturbations in the MobileAloha dataset, measured by success rate and. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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 robust are instruction-following capabilities of Claude-3.5-Sonnet and quantized mobile models when tested with adversarial perturbations in the MobileAloha dataset, measured by success rate and contextual understanding scores?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

16 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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.	×	0.08
The temperature parameter was set to 0.0 for all evaluations to minimize response variability.	×	0.02
Response timeout limits were established at 10 seconds.	×	0.03
Primary metrics included binary pass/fail determination, response time measurement, and token usage statistics.	×	0.01
Results were compiled into a structured Excel workbook containing an overview sheet, individual test sheets, and a model	×	0.02
The verification process operates in two stages: primary verification for strict matching and secondary verification for	×	0.01
The evaluation encompassed 331 models available via OpenRouter as of October 14, 2025.	×	0.13
256 out of 331 models passed basic functionality verification and were subsequently evaluated.	×	0.08
Twenty diagnostic prompts were used to evaluate the verified models.	×	0.08
Stage 1 verification used the query 'What is the capital of France?' to test basic responsiveness.	×	0.02
Model metadata, including supported parameters, was retrieved via the OpenRouter API prior to testing.	×	0.04
Verification payloads were constructed dynamically to include only parameters supported by each specific model.	×	0.02
Where supported, models were tested with max_tokens=150 and seed=42.	×	0.01
One benchmark result listed in Table (p20) shows a value of 80.1%.	×	0.02
One benchmark result listed in Table (p20) shows a value of 70.7%.	×	0.02
One benchmark result listed in Table (p20) shows a value of 70.3%.	×	0.02

References

- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2308.10819v3
- http://arxiv.org/abs/2510.18892v1