Instruction-Tuned Llama3 and DeepSeek R1 Robustness Against Adversarial Code Security Perturbations

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How do instruction-tuned Llama3 and Deepseek R1 models compare in robustness scores when evaluated against taxonomy-specific adversarial perturbations in code security benchmarks. 7 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLMs Caught in the Crossfire: Malware Requests and Jailbreak Challenges. Research question: How do instruction-tuned Llama3 and Deepseek R1 models compare in robustness scores when evaluated against taxonomy-specific adversarial perturbations in code security benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

8 papers retrieved. 7 claims extracted; 6 independently verified. Quality review score: 7.3/10.

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
MalwareBench is a benchmark dataset containing 3,520 jailbreaking prompts for malicious code-generation.	✓	0.32
MalwareBench is based on 320 manually crafted malicious code generation requirements.	✓	0.32
MalwareBench covers 11 jailbreak methods.	×	0.14
MalwareBench covers 29 code functionality categories.	✓	0.15
The average rejection rate for malicious content by mainstream LLMs is 60.93%.	✓	0.22
The average rejection rate for malicious content drops to 39.92% when jailbreak methods are combined with jailbreak atta	✓	0.27
Prior research on LLM security has largely not explored their specific susceptibility to jailbreak attacks in code gener	✓	0.27

References

- https://doi.org/10.48550/arxiv.2411.15594
- https://doi.org/10.18653/v1/2025.acl-long.1350
- https://doi.org/10.3390/info16110926