Reasoning-Focused Training Enhances Jailbreak Resistance in Code-Generating Large Language Models

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of reasoning-focused training on the jailbreak resistance of code-generating LLMs when evaluated on malware prompt datasets. 5 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. Research question: What is the impact of reasoning-focused training on the jailbreak resistance of code-generating LLMs when evaluated on malware prompt datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

12 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 8.2/10.

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
Generative AI requirements are categorized into three distinct categories: hardware, software, and user experience.	✓	0.24
Generative AI models described in the literature include variational autoencoders (VAEs), generative adversarial network	✓	0.34
The study presents a taxonomy of generative AI models based on architectural characteristics.	✓	0.19
The research proposes a classification system for generative AI based on output types.	✓	0.24
The study provides a comprehensive classification of input and output formats used in generative AI systems.	✓	0.33

References

- https://doi.org/10.1007/s10462-024-10888-y
- https://doi.org/10.3390/fi15080260
- https://doi.org/10.1145/3658644.3670388