Multimodal vs. Unimodal Models in Adversarial Robustness for Malware Detection

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How do multimodal models (e.g., combining code and behavioral features) with adversarial fine-tuning compare to unimodal models in terms of robustness against evasion attacks on malware detection. 19 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. Research question: How do multimodal models (e.g., combining code and behavioral features) with adversarial fine-tuning compare to unimodal models in terms of robustness against evasion attacks on malware detection benchmarks (e.g., MalConv, DeepMal)?.

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 6.2/10.

3 Results

8 papers retrieved. 19 claims extracted; 9 independently verified. Quality review score: 6.2/10.

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 evolution of Generative AI models was a highlight of digital transformation in 2022.	✓	0.18
ChatGPT and Google Bard are examples of Generative AI models that continue to increase in complexity and capability.	×	0.15
Recent instances have demonstrated the use of Generative AI tools in both defensive and offensive cybersecurity operatio	✓	0.17
ChatGPT possesses vulnerabilities that can be exploited to exfiltrate malicious information by bypassing ethical constra	✓	0.20
The paper demonstrates successful Jailbreak attacks on ChatGPT.	×	0.13
The paper demonstrates successful reverse psychology attacks on ChatGPT.	✓	0.17
The paper demonstrates successful prompt injection attacks on ChatGPT.	✓	0.19
Cyber offenders can use Generative AI tools to develop cyber attacks.	✓	0.16
Adversaries can use ChatGPT to create social engineering attacks.	✓	0.17
Adversaries can use ChatGPT to create phishing attacks.	×	0.11
Adversaries can use ChatGPT for automated hacking.	×	0.11
Adversaries can use ChatGPT for attack payload generation.	×	0.14
Adversaries can use ChatGPT for malware creation.	×	0.11
Adversaries can use ChatGPT to create polymorphic malware.	×	0.11
Generative AI tools can be used to improve security measures including cyber defense automation.	✓	0.23
Generative AI tools can be used to improve security reporting.	×	0.12
Generative AI tools can be used to improve threat intelligence.	×	0.13
Generative AI tools can be used for secure code generation.	✓	0.16
Generative AI tools can be used for secure code detection.	×	0.12

References

- https://doi.org/10.1109/access.2023.3300381
- https://doi.org/10.1109/access.2020.3006143
- https://doi.org/10.1109/tnsm.2020.2971776