Generative AI in Hacking: How Cybercriminals Use It Now

July 2025

Abstract

Generative Artificial Intelligence (GenAI) has transformed cybersecurity, offering both innovative defenses and new avenues for cyberattacks. This paper explores how cybercriminals exploit GenAI to enhance hacking techniques, including phishing, malware creation, deepfakes, and social engineering. Through a systematic review of current applications, we analyze the mechanisms, implications, and challenges of GenAI-driven cybercrime. We also discuss emerging defense strategies and propose future research directions to mitigate these threats. This study aims to inform cybersecurity professionals and policymakers about the evolving threat landscape and encourage proactive measures to safeguard digital systems.

Contents

1 Introduction

Generative AI (GenAI) refers to artificial intelligence systems capable of creating new content, such as text, images, audio, or code, by learning patterns from vast datasets. While GenAI powers creative tools like chatbots and art generators, its misuse by cybercriminals has raised significant concerns. In hacking, GenAI enables attackers to automate and enhance malicious activities, making cyberattacks more sophisticated and harder to detect. This paper examines how cybercriminals currently use GenAI in hacking, focusing on its applications, implications, and potential countermeasures.

The rise of GenAI tools, such as large language models (LLMs) like ChatGPT and image generators like DALL-E, has democratized access to advanced technology. Cybercriminals leverage these tools to craft convincing phishing emails, generate malicious code, and create deepfakes, posing unprecedented challenges to cybersecurity. This research addresses the urgent need to understand these threats and develop robust defenses.

2 Background

2.1 What Is Generative AI?

Generative AI encompasses algorithms that produce novel outputs based on learned patterns. Unlike traditional AI, which focuses on classification or prediction, GenAI creates content that mimics human-generated material. Key technologies include:

- **Large Language Models (LLMs)**: Models like GPT-4 generate human-like text for conversations or code.
- **Generative Adversarial Networks (GANs)**: Used for creating realistic images, videos, or audio.

- **Variational Autoencoders (VAEs)**: Employed in data generation tasks, including synthetic malware.

These technologies have legitimate uses but are increasingly exploited for malicious purposes.

2.2 Cybercrime and AI Evolution

Cybercrime has evolved from basic viruses to complex, targeted attacks. The integration of AI, particularly GenAI, has accelerated this evolution by enabling automation and personalization. Recent studies highlight GenAI's role in scaling cyberattacks, reducing the technical expertise required for sophisticated hacks [1, 10].

3 Methodology

This study employs a systematic literature review based on Prisma International Standards. We searched databases like ACM, arXiv, IEEE Xplore, and Springer for articles published between 2018 and 2025, focusing on GenAI in cybersecurity. Out of 936 papers, 46 were selected based on relevance to AI-driven cyberattacks. The review addresses three research questions:

1. How do cybercriminals use GenAI to enhance hacking techniques?
2. What challenges do GenAI-driven attacks pose to cybersecurity?
3. What countermeasures can mitigate these threats?

4 Applications of Generative AI in Hacking

4.1 Phishing and Social Engineering

GenAI enhances phishing by generating highly convincing emails and messages. LLMs can mimic a sender's tone, incorporate personal details, and avoid detection by traditional filters. For example, cybercriminals use tools like WormGPT to automate phishing campaigns, increasing their success rate [11]. Social engineering attacks also benefit from GenAI's ability to create personalized content, such as fake profiles or tailored messages [6].

4.2 Malware Creation

GenAI enables cybercriminals to generate novel malware that evades antivirus software. By using LLMs to write malicious code or GANs to create polymorphic malware, attackers can produce unique attack payloads. Research shows that tools like ChatGPT can be manipulated to generate ransomware scripts with minimal fixes [5, 9].

4.3 Deepfake Attacks

Deepfakes, powered by GANs, create realistic audio or video impersonations. Cybercriminals use deepfakes to trick victims into transferring funds or sharing sensitive data. A notable case involved hackers using deepfake technology to impersonate a Binance CEO during video calls [9].

4.4 Automated Vulnerability Exploitation

GenAI can autonomously identify and exploit vulnerabilities. For instance, GPT-4-powered agents have exploited 87% of real-world vulnerabilities by analyzing CVE descriptions [2]. This capability allows even novice hackers to target unpatched systems.

4.5 Jailbreaking and Prompt Injection

Jailbreaking involves bypassing AI model safeguards to extract malicious outputs. Techniques like prompt injection and reverse psychology manipulate LLMs to generate harmful content, such as instructions for illegal activities [8, 10]. These methods highlight vulnerabilities in current AI systems.

5 Implications of GenAI in Hacking

5.1 Increased Attack Scale and Speed

GenAI reduces the technical barrier for cyberattacks, enabling less-skilled actors to launch sophisticated attacks. Automation also accelerates attack execution, making it harder for defenders to respond in time [5].

5.2 Challenges in Detection

AI-generated content often lacks traditional signatures, complicating detection. For example, polymorphic malware changes its structure to avoid antivirus scans, and deepfakes challenge authentication methods [3].

5.3 Ethical and Regulatory Concerns

The accessibility of GenAI tools raises ethical questions about their regulation. Current safeguards are often bypassed, necessitating stronger oversight to prevent misuse [5, 7].

6 Countermeasures

6.1 AI-Powered Defenses

Cybersecurity professionals are leveraging GenAI to counter threats. Techniques include:

- **Automated Threat Detection**: GenAI models identify patterns in network traffic to flag anomalies.
- **Phishing Detection**: LLMs analyze email content for subtle red flags, improving detection rates [3].
- **Adversarial Training**: Training AI models to recognize manipulated inputs strengthens defenses.

6.2 User Awareness and Training

Educating users about GenAI-driven threats is critical. Tips include verifying email sources, avoiding suspicious media, and using two-factor authentication [4].

6.3 Regulatory Frameworks

Governments and organizations must regulate GenAI to limit misuse. Proposed measures include stricter chatbot safeguards and penalties for malicious AI development [5].

7 Future Research Directions

Future studies should focus on:

- Developing robust ethical guidelines for GenAI use.

- Enhancing AI model resilience against jailbreaking and prompt injection.
- Exploring GenAI's role in autonomous cyber defense systems.

8 Conclusion

Generative AI in hacking represents a dual-edged sword, offering powerful tools for both cybercriminals and defenders. While hackers exploit GenAI for phishing, malware, deepfakes, and vulnerability exploitation, cybersecurity experts can harness it for advanced threat detection and response. Addressing these challenges requires a combination of technological innovation, user education, and regulatory oversight. This paper provides a foundation for understanding GenAI's role in cybercrime and calls for proactive measures to secure the digital landscape.

References

[1] Review of Generative AI Methods in Cybersecurity, arXiv, 2024.

[2] Generative AI in Cybersecurity, CETaS Briefing Paper, 2024.

[3] Generative AI in Cybersecurity: A Comprehensive Review, Artificial Intelligence Review, 2025.

[4] Cybercriminals are Creating Their Own AI Chatbots, The Conversation, 2024.

[5] Ransomware Attacks in the Context of Generative AI, International Cybersecurity Law Review, 2023.

[6] Digital Deception: Generative AI in Social Engineering and Phishing, Artificial Intelligence Review, 2024.

[7] An Assessment of the Use of Generative AI in Cybersecurity, ResearchGate, 2023.

[8] The Hacking of ChatGPT Is Just Getting Started, WIRED, 2023.

[9] Hacking with AI, Atlantic Council, 2024.

[10] From ChatGPT to ThreatGPT, ResearchGate, 2023.

[11] WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks, The Hacker News, 2023.