Survey on Offences and Defenses Against Generative AI and Large-Language Models

University of Massachusetts Amherst

Jacqueline Bradley, Xingyu Cai, Turag Ikbal

Ph.D. Mentor: Dayeon Kang

Introduction

Artificial intelligence is being widely used in life, which brings greater risks of cyber attack and leads to problems like personal data leakage which concerns all of us. We want to understand what offences against generative AI and specifically large-language models exist, how they work, and what are the mitigations. We therefore set out to conduct a survey on this topic.

Background

Artificial Intelligence (AI): Computer learning models that can solve complex tasks at a faster rate than humans

Large Language Models (LLMs): Machine Learning Models trained on textual data that can give a text response given a text input

Generative AI(GenAI): Machine Learning Models that can generate other media outside of but also including text, such as code or images

Distinction between LLMs and GenAI: LLMs focus on natural language processing, while GenAI has a wider range of capabilities. All LLMs are GenAI, but not all GenAI are LLMs

Figures

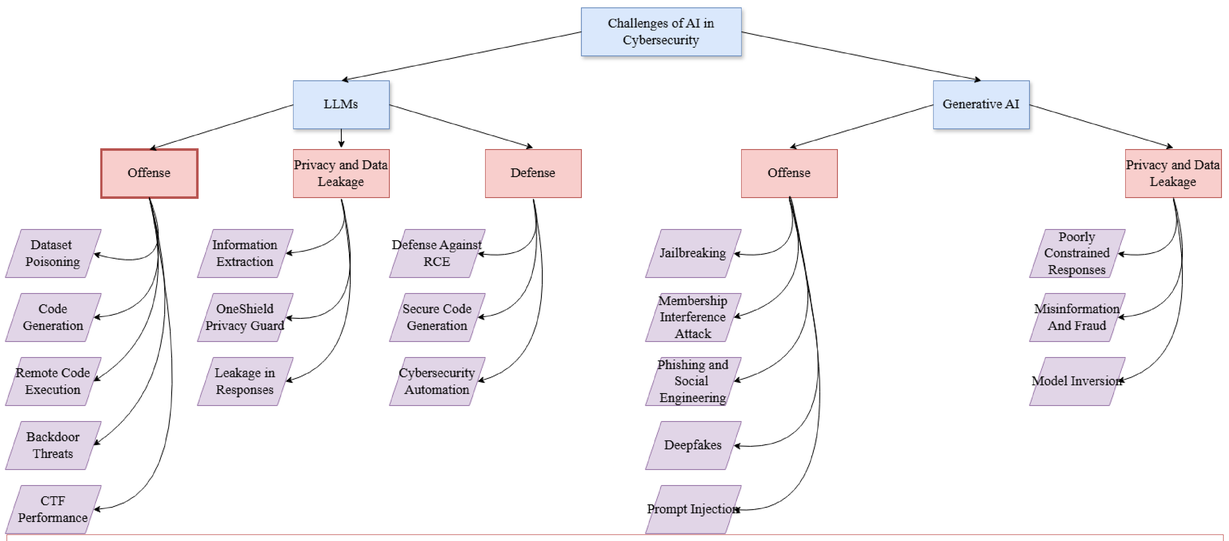

Figure 1: Resulting Offense and Defense Vectors

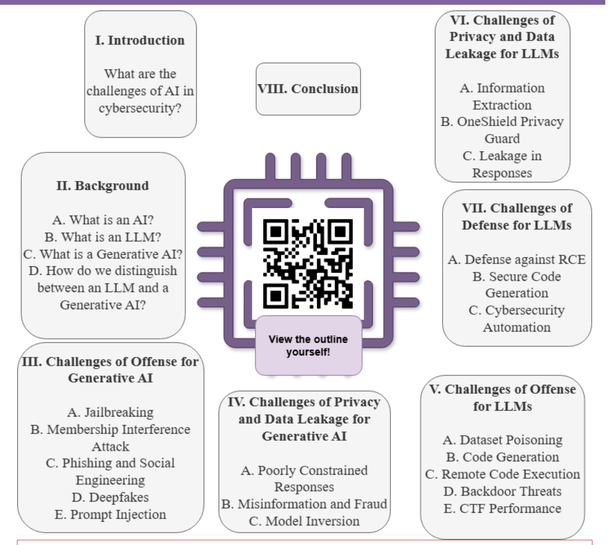

Figure 2: Survey Outline

Papers	GenAI	LLM	Offences	Defences	Privacy/Data Leakage
Asthana et al (2025)		✓		One Shield Privacy Guard	Peronsal Identifiable Information Leakage
Calini et al (2024)	✓		Data Poisoning	SHA-256 Hash, Vary Snapshot Time	
Duan et al (2023)	✓		Membership Inference Attack	Data Augmentation	Peronsal Identifiable Information Leakage
Gupta et al. (2023)	✓		Attack Vector Creation	Automated Incident Response,	Prompt Injection
Liu et al (2024)		✓	Remote Code Execution	Permission Management, Environment Isolation, Prompt Analysis	
Marchal (2024)	✓		Deepfakes, Prompt Injection, Data Poisoning		Model Extraction, Private Conversations
Motlag (2024)		✓	PassGPT, Malware Generation	Process Security Logs	
Pearce et al. (2023)		✓			Information Leakage
Sandoval et al. (2023)		✓	Generation of Insecure Code		
Shao et al (2024)		✓	CTF Performance	Train Security Teams, Vulnerability detection	
Wei (2023)		✓	Jailbreaking	Safety-Capability Parity*	
Yang et al. (2024)		✓	Backdoor Attacks		

Figure 3: Table of Paper Coverage

Scan for Paper References

Summary

The integration of AI into our daily life comes with significant cyber security challenges. Forms of attack include jailbreaking, data poisoning, remote code execution, backdoor attacks, and so on, which brings great risk of privacy violation. While AI enhances many aspects of our life, it also broadens the attack surface. It is vital to understand how to respond to these threats. We have learned that they can be addressed through stronger safety training, improved model transparency, and robust defense mechanisms. Future research should focus on secure AI model development, adversarial defense techniques, and regulatory frameworks to ensure we deplore AI in a responsible and secure way.