Cybersecurity Considerations for Generative AI

Introduction

The rapid advancement of artificial intelligence (AI) has ushered in a new era of technological innovation, enabling a wide range of applications across various industries. The AI market is projected to reach a staggering $407 billion by 2027, experiencing substantial growth from its estimated $86.9 billion revenue in 2022.[1]

> The AI market is projected to reach a staggering $407 billion by 2027, experiencing substantial growth from its estimated $86.9 billion revenue in 2022.

Generative AI, Language Models (LLMs), and Chatbots have emerged as powerful tools, capable of generating human-like text and engaging in interactive conversations. While these technologies offer numerous benefits, they also introduce significant cybersecurity considerations necessary to ensure the integrity, confidentiality, and availability of information.

Defining the Technology

Generative AI is a breed of artificial intelligence that excels in creating or 'generating' outputs. These AI models are adept at producing an array of content such as images, music, and text, all meticulously crafted and virtually indistinguishable from human-produced content.

Large Language Models (LLMs), a vital subset of Generative AI, demonstrate the linguistic proficiency and capabilities of this field of technology. These models excel at comprehending, processing, and generating human language in a remarkably intuitive and meaningful way.

Chatbots, the digital conversationalists powered by Generative AI and LLMs, are transforming the landscape of customer interaction and service delivery. These intelligent virtual assistants, capable of simulating human-like conversation, are used to automate customer service, provide round-the-clock support, and deliver personalized customer experiences.

Complex Functionality

Generative AI, Large Language Models (LLMs), and Chatbots all refer to different types of artificial intelligence applications with subtle differences.

While other AI systems classify or analyze data, Generative AI broadly describes any system that uses artificial intelligence to generate content. Generative AI learns and creates from the data it has been trained on, providing the foundational capabilities for creating new, unique outputs that mimic the characteristics of the original data.

Large Language Models (LLMs) focus specifically on textual data (a language). By training on vast quantities of text, they generate responses or create text that mirrors human language, understanding the subtle nuances and contextual relevance.

Chatbots leverage the capabilities of Generative AI and LLMs by providing end-users a "chat-like" interface. They conversationally interact with users, comprehending user inputs and delivering appropriate responses.

Generative AI Cybersecurity Controls

Cybersecurity controls for AI should consist of multiple layers aimed at protecting data, models, and processes from cyber threats.

SaaS Versus Self-Hosted (On-Prem)

In Software as a Service (SaaS) solutions, where the models and user interfaces are hosted by a provider, security controls are largely managed by the service provider. These include physical and network security measures for the cloud infrastructure hosting the AI

1 https://www.forbes.com/advisor/business/ai-statistics/?utm_source=www.joinsuperhuman.ai&utm_medium=newsletter&utm_campaign=7-key-ai-statistics-from-2023-you-should-know#sources_section

models, along with application-level controls such as strong access management, end-to-end encryption of data in transit and at rest, and regular security updates and patches.

Self-hosted AI solutions, meanwhile, require organizations to implement and manage their own security controls. These typically involve network and system security measures such as firewalls, intrusion detection and prevention systems, and secure network design. At the application level, strong user authentication and authorization, data encryption, secure AI model storage, and routine security audits are crucial.

In both cases, considerations should be made for the data life cycle, ensuring data privacy and integrity from collection and storage to processing and disposal. Regular training for staff on security best practices also plays a crucial role.

Threats to AI Systems and Users

Data Privacy Concerns

AI algorithms are essentially fueled by data. This data, often sensitive and private, represents the most critical assets of many organizations. As SaaS models continue to be integrated into business ecosystems, they become lucrative targets for cybercriminals. Attackers who can successfully access training data for AI systems will gain access to much of this data.

> AI algorithms are essentially fueled by data. This data, often sensitive and private, represents the most critical assets of many organizations. As SaaS models continue to be integrated into business ecosystems, they become lucrative targets for cybercriminals.

Information Disclosure Risks

Unauthorized information disclosure, often due to unintended capabilities or programming errors that are discovered only after systems enter production, can lead to serious consequences including reputational damage, regulatory fines, and loss of customer trust. Furthermore, such leaks offer cybercriminals vital insights to plan sophisticated attacks. Trivial leaks can facilitate targeted phishing or identity theft. Hence, AI SaaS platforms must prioritize stringent data security and privacy measures.

LLM Interface Attack Vectors

There are multiple methods that attackers can attempt to compromise AI Systems, and they vary depending on how much of the system is web-exposed or Internet connected.

Application Programming Interfaces, or APIs, represent a critical attack surface for potential cyber threats. Unauthorized access through API exploitation can lead to misuse of the model, extraction of sensitive information, or introduction of adversarial inputs that skew the model's outputs. Attacks might range from simple API key thefts, enabling unauthorized data access, to more sophisticated attacks exploiting flaws in API structure or security protocols. Thus, securing the API of an LLM demands stringent controls, rigorous testing, and continuous monitoring to ensure robust defense against potential cyber threats.

Similar to APIs, Web Application User Interfaces (Web UIs) are another significant attack surface. Cyber adversaries can exploit vulnerabilities in the application's code to inject malicious scripts (Cross-Site Scripting), hijack user sessions (Session Hijacking), or to manipulate database queries (SQL Injection). These attacks could lead to unauthorized access to the LLM, manipulation of its outputs, or potential exfiltration of sensitive data. This makes stringent application security practices such as regular code audits, input validation, and the use of secure, up-to-date frameworks vital in protecting the web application front of LLMs from cyber threats.

Another common feature of LLM infrastructure is the Command Line Interface (CLI). Access to a CLI may allow attackers to exploit vulnerabilities and execute arbitrary commands (e.g. Command Injection), or manipulate the LLM's behavior. They could potentially alter the model's output or allow unauthorized access to the system it runs on. Ensuring rigorous input sanitization, secure command execution, and robust access control mechanisms are crucial in mitigating these risks associated with the CLI of LLMs.

Vulnerabilities in LLMs

LLMs are vulnerable to several emerging attack techniques. LLMs possess access to a diverse range of information that serves as the foundation for their probabilistic modeling, and attacks directed at LLMs exploit this extensive information or exploit the absence of adequate governance measures. The following section details some techniques targeting LLMs.

Prompt Injection

The practice of prompt injections entails circumventing filters or manipulating the LLM by employing meticulously constructed prompts which cause the model to disregard prior instructions or execute unintended actions. Exploiting these vulnerabilities may result in data leakage, unauthorized access, or other breaches of security.

Data Leakage

Data leakage refers to any event when an LLM inadvertently discloses sensitive information, proprietary algorithms, or other confidential data in its responses. Data leakage can also occur on the platform hosting the LLM.

The repercussions of data leakage extend beyond immediate consequences, as the compromised information can be further exploited to perpetrate additional security breaches or enable cybercriminal activities. Therefore, it is imperative to implement robust security measures and rigorous monitoring to prevent data leakage and mitigate its potential impact.

Inappropriate Sandboxing

Inadequate sandboxing occurs when a Language Model (LLM) lacks appropriate isolation measures while being granted access to external resources or sensitive systems. This poses significant risk, as it opens avenues for potential exploitation, unauthorized access, or inadvertent actions by the LLM, thereby compromising the security and integrity of the overall system.

> It is imperative to implement robust security measures and rigorous monitoring to prevent data leakage and mitigate its potential impact.

To mitigate risk, it is imperative to establish stringent sandboxing protocols that effectively isolate the LLM from external resources and sensitive systems. By implementing proper containment measures, such as limiting network access, restricting file system permissions, and employing strong authentication mechanisms, the potential for exploitation and unintended consequences can be significantly reduced. This ensures that the LLM operates within predefined boundaries, safeguarding critical assets and preventing unauthorized activities that could jeopardize security.

Unauthorized Code Execution

Unauthorized code execution occurs when an adversary exploits an LLM to execute. nefarious code, commands, or actions on the underlying system by leveraging natural language prompts. This attack vector poses a significant threat to the security and integrity of the system.

By manipulating the LLM through carefully crafted prompts, an attacker can coerce the model into executing arbitrary code or commands that were never intended by its designers. This unauthorized execution capability grants the attacker unauthorized control over the underlying system, enabling them to carry out malicious activities, compromise sensitive data, or cause system malfunctions.

Training Data Poisoning

Training data poisoning refers to malicious manipulation of the training data or fine-tuning procedures of an LLM to introduce vulnerabilities, backdoors, or biases that have the potential to undermine the model's security, efficacy, or ethical comportment. By tampering with the training data or fine-tuning procedures, an attacker can inject subtle yet impactful modifications that subvert the LLM's intended behavior. This can include introducing malicious patterns, biased information, or intentionally misleading examples into the training data, aiming to bias the model's decision-making process or compromise its ability to generalize effectively.

Training data poisoning can have detrimental consequences across multiple domains. For instance, in the context of security, the attacker may introduce hidden vulnerabilities or backdoors that, when triggered, enable unauthorized access, data exfiltration, or the manipulation of sensitive information. Moreover, the injection of biases can result in discriminatory or unfair outcomes, perpetuating societal inequalities or compromising the model's ethical behavior.

Conclusion

It is clear that Generative AI, Large Language Models, and Chatbots are already having a tremendous impact on multiple facets of everyday life. It is imperative that organizations who produce, maintain, and offer this technology to public, commercial, and government organizations and consumers take steps to manage the risks associated with using the technology. By understanding the risks associated with Generative AI, and developing and maintaining appropriate risk management processes, it is possible to unlock massive benefits that the technology promises.

Contributors

Chase Theodos, Evans Foster, Jason Ingalls, Kim Buckley, and ChatGPT