Introducing the IBM Framework for Securing Generative AI

While generative artificial intelligence (AI) is becoming a top technology investment area, many organizations are unprepared to cope with the cybersecurity risks associated with it.

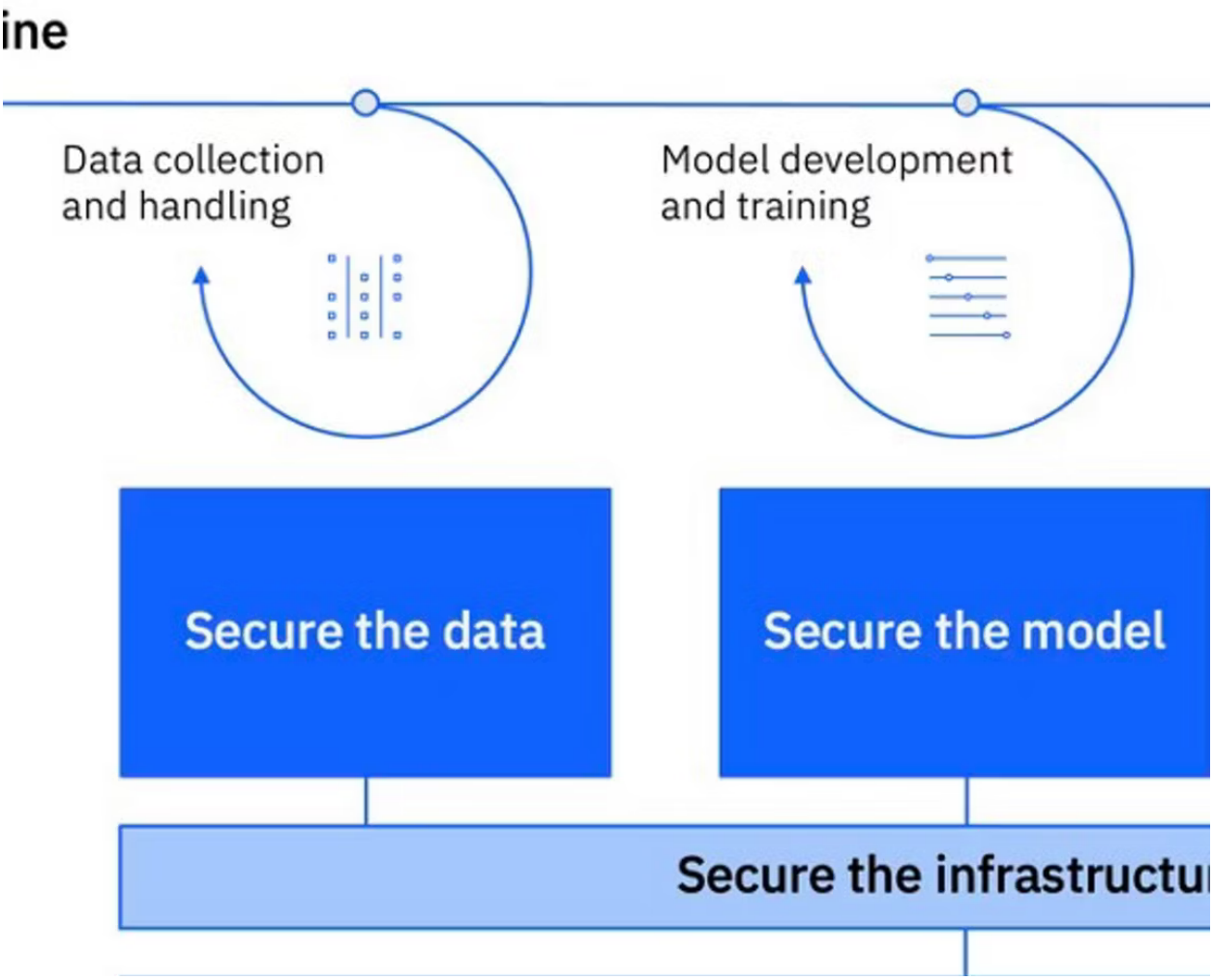
Like with any new technology, it's paramount that we recognize the new security risks generative AI brings, because it's without a doubt that adversaries will try to exploit any weakness in pursuit of their objectives. In fact, according to the IBM Institute for Business Value, 96% of executives say adopting generative AI makes a security breach

likely in their organization within the next three years.

With AI models ingesting troves of valuable and sensitive data into their training sets—plus business leaders examining how these models can optimize critical operations and outputs—the stakes are incredibly high. Organizations cannot afford to bring unsecured AI into their environments.

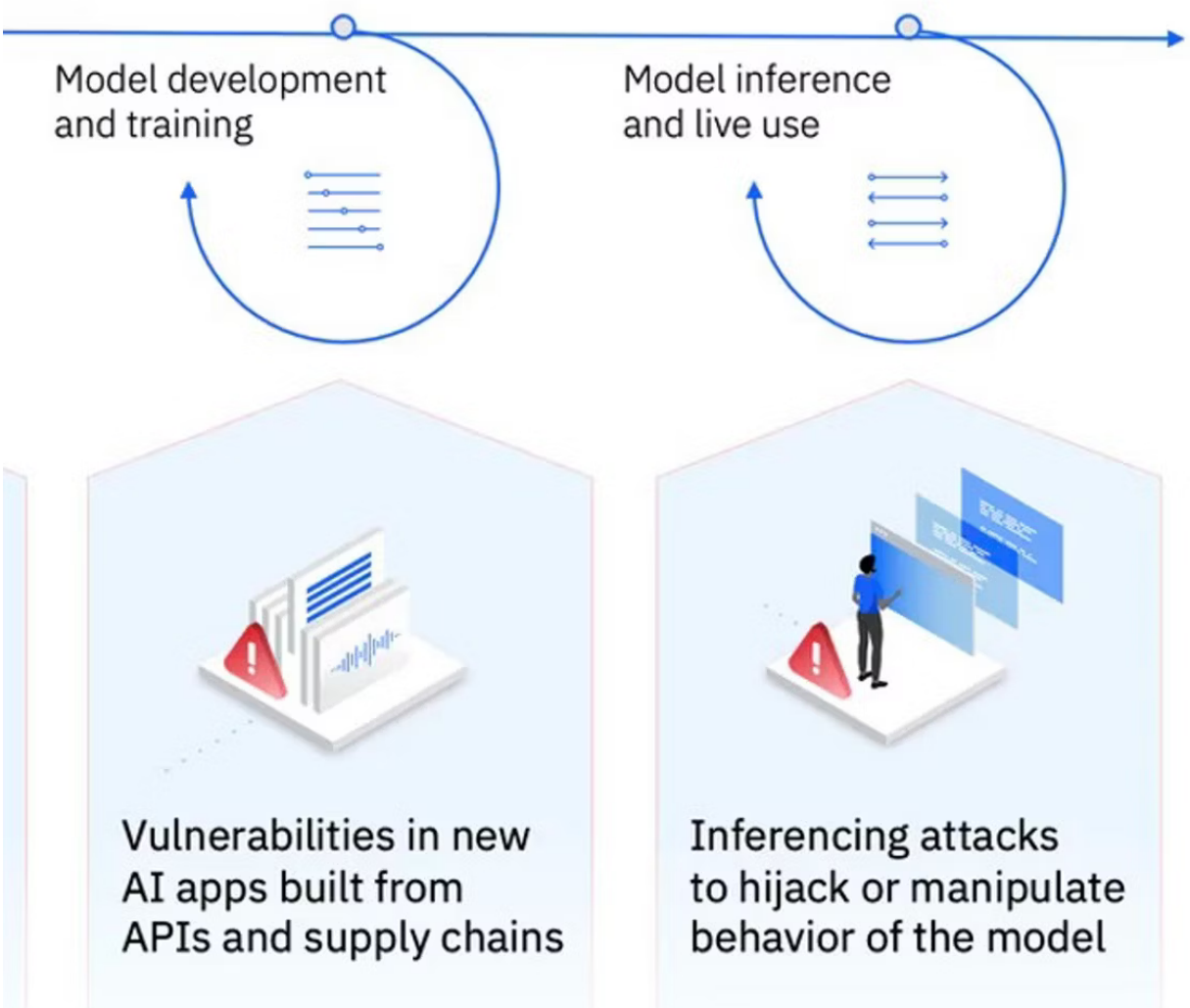
IBM framework for securing generative AI

In this blog, we introduce the **IBM Framework for Securing Generative AI.** It can help customers, partners and organizations around the world better understand the likeliest attacks on AI and prioritize the defensive approaches that are most important to secure their generative AI initiatives quickly.



It is crucial that we secure AI at each stage of the AI pipeline—this includes during data collection and handling, model development and training, and model inference and use. As such, organizations need to secure the data, the model and the model's usage. They must also secure the infrastructure on which the AI models are being built and run. Finally, they need to establish AI governance and monitor for fairness, bias and drift over time.

Below we detail the risks in each stage of the AI pipeline and how to protect it against the primary identified attacks.



Secure the data

During the data collection and handling phase, not only do you need to collect mounds and mounds of data to feed an AI model, but you're also providing access to lots of different people, including data scientists, engineers, developers and others. There is an inherent risk presented in centralizing all that data in one place and granting various stakeholders—most of whom don't have security experience—access to it.

Just consider if intellectual property (IP) that is fundamental to the business is exposed due to mishandling of the training data, potentially creating an existential threat to the business. Leveraging troves of data for an AI model means organizations need to assess the varying risk tied to personally identifiable information (PII), privacy concerns and other sensitive information, and then place the proper security controls around that data.

Safeguards and defenses against likeliest attacks

The primary target in the data collection phase are the underlying data sets, with data exfiltration deemed as the likeliest technique that attackers will seek to employ to get their hands on valuable and monetizable information. As attackers seek the path of least resistance, underlying data sets are a blinking light promising a high yield.

Organizations must not overlook the importance of security fundamentals—in fact, they should be prioritized. If applied correctly, these fundamentals can have a substantial impact on an organization's security posture. This includes focusing on data discovery and classification, encryption at rest and in transit, and key management delivered from data security platforms such as IBM Security® Guardium®. This also means focusing on identity and access management fundamentals enforced by solutions such as IBM Security® Verify, which help ensure that no single entity has unrestricted access to the AI model. Finally, organizations must raise security awareness with the data scientists and researchers and make sure security teams work closely with those teams to ensure proper guardrails.

Secure the model

Within model development, you're building applications in a new way, and that often involves introducing new, exploitable vulnerabilities that attackers can use as entry

points into the environment and, in turn, into your AI models. Considering that organizations have historically struggled with managing a growing debt of known vulnerabilities found within their environments, this risk will carry over to AI.

Developing AI applications often starts with data science teams repurposing pretrained, open-source machine learning (ML) models from online model repositories, which often lack comprehensive security controls. However, the value they provide organizations, such as dramatically reducing the time and effort required for generative AI adoption, often outweighs that risk, ultimately passing it on to the enterprise. The general scarcity of security around ML models, coupled with the increasingly sensitive data that ML models are exposed to, means that attacks targeting these models have a high potential for damage.

Safeguards and defenses against likeliest attacks

The primary attack techniques during model development are supply chain attacks due to the heavy reliance on pretrained, open-source ML models from online model repositories used to accelerate development efforts. Attackers have the same access to these online repositories and can deploy a backdoor or malware into them. Once uploaded back into the repository, they can become an entry point to anyone that downloads the infected model. If these models are infected, it can be incredibly difficult to detect. Organizations must be very cautious about where they consume models and how trusted the source is.

Application programming interface (API) attacks are another concern. Organizations without the resources or expertise to build their own large language models (LLMs) rely on APIs to consume the capabilities of prepackaged, pretrained models. Attackers recognize this will be a major consumption model for LLMs and will look to target the API interfaces to access and exploit data being transported across the APIs.

Attackers may also seek to exploit LLM agents or plug-ins with excessive permissions to access open-ended functions or downstream systems that can perform privileged actions in business workflows. If an attacker can compromise privileges that are granted to AI agents, the damage could be destructive.

Organizations' focus should include:

- Continuously scanning for vulnerabilities, malware and corruption across the AI/ML pipeline
- Discovering and hardening API and plug-in integrations to third-party models

- Configuring enforcing policies, controls and RBAC around ML models, artifacts and data sets so that no one person or thing has access to all the data or all of the model functions

Secure the usage

During inferencing and live use, attackers can manipulate prompts to jailbreak guardrails and coax models into misbehaving by generating disallowed responses to harmful prompts that include biased, false and other toxic information. This can inflict reputational damage on the enterprise. Attackers might also seek to manipulate the model and analyze input/output pairs to train a surrogate model to mimic the behavior of the target model, effectively "stealing" its capabilities and costing an enterprise its competitive advantage.

Safeguards and defenses against likeliest attacks

Several types of attacks are concerning in this stage of the AI pipeline. First, prompt injections—where attackers use malicious prompts to jailbreak models and get unwarranted access, steal sensitive data or introduce bias into outputs. Another concern involves model denial of service, where attackers overwhelm the LLM with inputs that degrade the quality of service and incur high resource costs. Organizations should also prepare for and defend against model theft, where attackers craft inputs to collect model outputs to train a surrogate model that mimics the behavior of the target model.

Our best practices include monitoring for malicious inputs such as prompt injections and outputs containing sensitive data or inappropriate content, and implementing new defenses that can detect and respond to AI-specific attacks such as data poisoning, model evasion and model extraction. New AI-specific solutions have entered the market under the name of machine learning detection and response (MLDR). Alerts generated from these solutions can be integrated into security operations solutions, such as IBM Security® QRadar®, enabling security operations center (SOC) teams to quickly initiate response playbooks that deny access, quarantine or disconnect compromised models.

Secure the infrastructure

One of the first lines of defense is a secure infrastructure. Organizations should leverage existing expertise to optimize security, privacy and compliance standards across distributed environments hosting the AI systems. It's essential that they harden network security, access control, data encryption, and intrusion detection and

prevention around AI environments. They should also consider investing in new security defenses specifically designed to protect AI.

Establish governance

IBM provides not only security for AI but also operational governance for AI. IBM is leading the industry in AI governance to reach trustworthy AI models. As organizations offload operational business processes to AI, they need to make sure the AI system is not drifting and is acting as expected. This makes operational guardrails central to an effective AI strategy. A model that operationally strays from what it was designed to do can introduce the same level of risk as an adversary that's compromised your infrastructure.

The leader of responsible AI

IBM has a long trust heritage and deep commitment to AI built with security, ethics, privacy and governance at its core. These principles underpin our AI, which is why we know that *how* AI models are built and trained are critical to achieving successful, and responsible, outcomes powered by AI.

Data quality, data lineage and data protection across our foundation models is one of our top priorities. At IBM, we implement strong controls and meticulous processes over the training pipeline of our models. They are trained on highly curated data to achieve data accuracy, completeness and provenance while lowering the risk of model hallucination.

Our commitment is evident through the introduction of IBM® watsonx.governance™, a product designed to help companies that use large AI models get that results are unbiased, factually correct and explainable.

We've also structured processes to address data transparency and full traceability to our customers, with an ability to showcase our data sources. We recently expressed our commitment to transparency and responsible AI by publishing the details of training data sets for Granite models. IBM also provides an IP indemnity (contractual protection) for its foundation models.

IBM is continuing to demonstrate and further its commitment to the effective and responsible deployment of AI, with a USD500 Million Enterprise AI Venture Fund to not

only fuel innovation but also invest in capabilities that help secure AI and build responsible solutions to customers' evolving needs.

For more information on how organizations can securely adopt generative AI, check out:

- A CISO's guide: Cybersecurity in the era of generative AI
- How to establish secure AI+ business models
- AI to accelerate your security defenses
- Watch: How Generative AI Changes the Cybersecurity Landscape (link resides outside ibm.com)
- The power of AI: Security

Author

Ryan Dougherty
Program Director
Emerging Security Technology, IBM Security

Explore cybersecurity in the era of generative AI →