7 Essential Layers for Generative AI Security and Governance

By Gopal Wunnava
Founder, DataGuard AI Consulting | Managing Director, XiPhi.ai
Originally published on DataGuard AI Consulting; co-published on XiPhi.ai with canonical attribution.

Generative AI: New Opportunities Lead to New Risks

Generative AI applications can be a game changer for organizations by creating new opportunities through the use of unstructured data, such as text and images. However, these applications also introduce new risks and challenges, as the underlying large language models (LLMs) are susceptible to various types of malicious attacks.

As the generative AI space rapidly innovates, the threat landscape expands. Organizations must implement a comprehensive security and governance strategy from the beginning to safeguard against malicious attacks. However, many organizations today have limited awareness of which strategies to adopt and the types of frameworks that need to be created to protect themselves from these threats right from the outset.

This leads to critical gaps in areas that will always be prone to malicious attacks unless steps are taken from the outset. In this article, I'll demonstrate how an enterprise can proactively protect itself from emerging threats and minimize risks by adopting seven essential layers. These foundational layers are critical for establishing a comprehensive generative AI security and governance framework.

Current Gaps and Limitations

The rapid growth and adoption of generative AI applications increase the risk of malicious attacks on enterprise data, leading to potential events such as jailbreaks, data poisoning, model theft, and exposure of sensitive data. The threat landscape is expanding daily, both in scale and in the variety of malicious attacks and infiltration points that can harm an

organization. Many of these threats, along with the extent of the potential damage, remain unknown at this time.

Organizations tend to respond only after a risk event has taken place and damage has been incurred. Such events may not only result in the loss of sensitive data but also cause harm to the organization's reputation, increase customer dissatisfaction, and lead to possible litigation. There is a critical need for organizations to establish a well-thought-out AI security and governance framework that will mitigate risks associated with generative AI implementation from the outset.

The generative AI security and governance frameworks that organizations have in place today are rarely comprehensive. Many organizations typically address risks in just a few areas of their applications, such as by placing controls to protect their data and application layers. More mature organizations may implement access control mechanisms to prevent attacks from unauthorized users.

Organizations further along in the maturity spectrum may have monitoring capabilities for their applications to identify and mitigate risks quickly. On the other hand, larger organizations may have implemented certain aspects of an enterprise AI security and governance framework.

As mentioned earlier, organizations place controls in response to events that may have resulted in a loss to the organization or to peers within their industry. Not only are such controls limited in nature, but they also lack the breadth and depth needed to provide protection from future attacks at every stage of the process. These attacks can find new ways to target the data, application, and infrastructure resources of an organization. Therefore, there is a critical need for a comprehensive strategy that can address these risks from the outset.

The 7 Essential Layers: A Comprehensive Framework

The "7 Essential Layers for Generative AI Security and Governance" are listed below. Implementing these layers enables a comprehensive, end-to-end security and governance framework that safeguards against emerging threats while ensuring regulatory compliance.

1) Infrastructure Controls

Infrastructure controls establish a robust defense mechanism for infrastructure resources such as networks, compute, data, and storage components of an organization. This prevents threats such as malicious attacks and unauthorized access.

This foundational layer includes the following:

- Methods to secure physical and virtual infrastructure that support generative AI systems and applications

- Networking components and private link infrastructure

- Encryption techniques to ensure data remains secure at rest and during transit

- AI firewall protection techniques

- Protection against DoS and DDoS attacks

- Infrastructure as Code (IaC) and managed/cloud-native services from AWS, Azure, and GCP

2) Access Controls

Access management is central to maintaining security and governance for generative AI applications. Access controls implement strict authentication and authorization processes to ensure that only authorized individuals or resources can view or modify AI resources.

This layer includes the following:

- Identity and Access Management (IAM)

- Role-Based Access Control (RBAC) and/or Attribute-Based Access Control (ABAC)

- Strong authentication techniques, including Multi-Factor Authentication (MFA)

- Authorization controls based on the principle of least privilege

- Secure handling of API tokens to prevent unauthorized access

3) Metadata Controls

Metadata controls govern the lifecycle of metadata to track, secure, and audit AI systems, ensuring compliance. Metadata is crucial for securing access to PII and sensitive data by applying tagging techniques.

This layer involves the following:

- Management and security of metadata, including metadata for users and data

- Classification and tags for various data types, including sensitive data

- Preservation of data privacy using techniques such as anonymization

- Providing context to raw data to help LLMs improve accuracy and relevance

- Facilitating organization and control over enterprise data assets

- Enabling collaboration for federated AI and learning

4) Data Controls

Data controls ensure the protection, security, and integrity of data. The underlying controls prevent data loss and ensure the privacy of data and compliance with regulations.

This layer includes the following:

- Secure storage and sharing of data in object storage and vector databases

- Securely ingest, version, and query datasets

- Secure retrieval of data using embedding techniques

- Data validation including input sanitization and output moderation

- Mechanisms to prevent data poisoning, prompt injection, and unauthorized modifications

- Integrity and confidentiality via audit trails and cryptographic hashing

- Privacy-preserving techniques such as anonymization to protect PII

5) Application Controls

Application controls address security and governance at both the front-end and back-end layers of a generative AI application. The front-end includes the UI, browser, and web framework. The back-end includes LLM models and RAG techniques.

This layer includes the following:

- Protection against jailbreaks

- Prevention of leaks involving sensitive data

- Protection from supply chain risks and harmful plugins

- Techniques to protect LLM models from vulnerabilities

- Federated AI and learning techniques to train models using decentralized data

- Incorporation of differential privacy techniques

- Regular audits to meet regulatory and compliance requirements

- Thorough validation of interfaces to ensure robustness of AI applications

6) LLM Ops / Observability

The LLM Ops (Operations) and Observability layer provides continuous monitoring and observability features for generative AI applications. This ensures high system performance while maintaining robustness against emerging threats.

This layer includes the following:

- Ensuring CI/CD for generative AI applications

- Evaluation of metrics for performance, efficiency, and security using automation

- Securing integration points with internal/external agents, APIs, and data exchanges

- Mitigation of risks through rigorous testing and validation

- Monitoring and logging through AI telemetry

- Controls to prevent unauthorized data enrichment or synthetic data generation

- Integration of DLP policies into output workflows to prevent exposure of sensitive data

- Secure handling of retrieval pipelines including audit trails and watermarking

7) AI Governance / Center of Excellence (CoE)

The AI Governance layer provides the foundation for ethical and responsible AI usage and encompasses governance, risk, and compliance (GRC) policies and procedures. The CoE for generative AI is an extension of this layer, providing best practices, fostering innovation, and ensuring compliance with legal and ethical guidelines.

The layer includes the following:

- Responsible and ethical use of AI across the organization

- Promotion of transparency and accountability

- Mitigation of legal and ethical risks; alignment with legal frameworks

- Establishment of a data governance framework across the organization

- Integration with human-in-the-loop techniques and processes

- Data access, retention, and usage policies

- AI training, education, and certification programs

- Risk assessment frameworks

- Best practices, ethical oversight, policy advocacy; performance assessment and collaboration

- R&D for AI initiatives across the enterprise

In summary, these seven essential layers provide a structured approach to securing and governing AI systems, addressing potential vulnerabilities at every stage, and delivering the breadth and depth of protection needed to ensure the integrity and reliability of generative AI operations from the outset.

Canonical Reference:
This article was originally published on DataGuard AI Consulting and is mirrored on XiPhi.ai with canonical attribution.
<link rel='canonical' href='https://dataguardaiconsulting.com/blog/7-essential-layers-for-generative-ai-security-and-governance/' />
