ISSN: 2756-2638
Transactions on Engineering and Computing Sciences; Vol. 14 No. 03 (2026) (01-08)
https://doi.org/10.14738/tmlai.1403.10456

Secure Deployment of Generative AI in Cloud Environments

Mohammed Ketel & Hari Joshi

1. Applied Information Technology Department, University of Baltimore, Baltimore, MD 21201, USA

Abstract: Generative Artificial Intelligence (GenAI) models have become widely adopted through cloud computing platforms such as AWS, Microsoft Azure, and Google Cloud. Models such as ChatGPT and Gemini are transforming industries ranging from education and healthcare to enterprises and public services. Cloud environments provide scalability, cost efficiency, and ease of deployment; however, they also introduce complex privacy and security challenges. GenAI models are susceptible to sophisticated attacks such as prompt injection, model inversion, unauthorized access through insecure APIs, and data leakage. This paper examines security and privacy risks in cloud-hosted GenAI systems across data, training, deployment, and interface stages. It reviews mitigations like AI firewalls, differential privacy, and secure enclaves, and explores secure and trustworthy GenAI deployments.

Keywords: AI, Generative AI, Cloud Security, Large Language Models, Machine Learning.

INTRODUCTION

Generative AI is redefining human interaction with machines by enabling systems to generate text, images, computer code, and other outputs that closely mimic human creativity. These powerful models are increasingly integrated into tools and platforms across various sectors, including education, healthcare, business, and entertainment, making them more accessible and widely adopted than ever before.

To support these capabilities, organizations are rapidly adopting cloud computing platforms that provide scalability, high-performance computing, and flexible infrastructure for deploying and managing large-scale AI models. Services such as AWS, Microsoft Azure, and Google Cloud enable on-demand computing, distributed storage, and managed AI services, allowing organizations to develop and deploy GenAI solutions more efficiently and cost-effectively compared to traditional infrastructure [17]-[19].

In recent years, large language models (LLMs) such as GPT-3, GPT-4, and Gemini have driven significant advances in generative AI. These models power chatbots, virtual assistants, code generation tools, content creation systems, and even applications in legal and healthcare domains. Hyperscale cloud platforms have enabled their deployment by providing the massive compute power, GPU acceleration, and storage required for training and inference [1], [4], [6].

However, this convenience introduces significant security risks. Cloud environments are shared, distributed systems often exposed through public APIs, making them vulnerable to new attack vectors such as prompt injection [7], model inversion attacks [2], and API scraping or abuse [8]. These threats are not purely theoretical; real-world cases

demonstrate that sensitive training data can be leaked, model outputs can be manipulated, and proprietary models can be extracted using relatively simple techniques [2], [7], [8].

This paper examines these evolving threats and explores how cloud-native technologies supporting GenAI systems can be combined with AI-specific privacy and security practices. It also considers emerging regulatory frameworks to align technical solutions with governance requirements.

BACKGROUND AND RELATED WORK

Generative AI is a subfield of machine learning in which models are trained on large datasets to generate novel outputs. Prominent examples include OpenAI's GPT series [1], Google's Gemini, and Stable Diffusion by Stability AI. These models are based on transformer architectures and rely on scaling data, parameters, and compute resources [1].

Cloud computing enables GenAI at scale by providing flexible infrastructure. Platforms such as AWS Bedrock, Azure OpenAI Service, and Google Vertex AI support inference, fine-tuning, and pipeline integration through managed APIs [4], [6], [12]. However, GenAI introduces security concerns that go beyond traditional cloud risks. While conventional cloud threats include misconfigurations, unauthorized access, and data breaches [5], GenAI introduces model-specific vulnerabilities such as inference manipulation, training data leakage, and model extraction [2], [7], [8].

Carlini et al. [2] demonstrated that sensitive information, including personal data, can be extracted from trained models. Reference [8] explored model stealing techniques via public APIs, while F Perez and I Ribeiro [7] introduced prompt injection attacks, where adversarial inputs manipulate model behavior. These studies highlight the lack of robust, GenAI-specific defenses at both model and infrastructure levels.

THREAT LANDSCAPE

A cloud-hosted GenAI system presents a broad attack surface spanning data ingestion, training, deployment, and inference stages. One of the most significant threats is prompt injection, where adversaries craft malicious inputs to override system instructions, bypass safety controls, or manipulate model behavior. Such attacks can lead to unintended or harmful outputs, especially in systems exposed through public interfaces [7], [9].

Model inversion attacks represent another major risk, where attackers attempt to reconstruct sensitive training data from model outputs. This is particularly critical in domains such as healthcare and education, where models may be trained on confidential or personally identifiable information. Research has shown that generative models can inadvertently leak aspects of their training data under certain conditions [2].

Data leakage can also occur due to insecure logging, poorly configured storage systems, or unencrypted inference pipelines. Improper handling of inputs and outputs may expose sensitive information, especially when cloud services retain request data for monitoring or debugging purposes [5].

Additionally, public APIs used to access GenAI models introduce risks of abuse and model extraction. Without proper authentication, rate limiting, and monitoring, attackers

can repeatedly query models to reverse-engineer their behavior or replicate proprietary systems [8].

Common cloud misconfigurations such as excessive access permissions, exposed storage buckets, or weak identity management can significantly amplify these risks. In some cases, improperly secured model weights or datasets stored in cloud storage can be directly accessed and misused, leading to full model compromise [4], [5].

Table 1 outlines common security threats associated with cloud-hosted Generative AI systems, highlighting their potential impacts and recommended security controls to reduce risk across the AI lifecycle.

Table 1: Threats and Mitigation Strategies in Cloud-Based Generative AI Environments

Threat	Description	Impact	Mitigation
Prompt Injection	Malicious prompts manipulate model behavior or bypass safeguards.	Harmful or unauthorized outputs	Input filtering, guardrails, output monitoring
Model Inversion	Attackers reconstruct sensitive training data from outputs.	Privacy breaches, PII exposure	Differential privacy, restricted outputs, query limits
Data Leakage	Sensitive data exposed via logs or storage.	Unauthorized data disclosure	Encryption, secure logging, access controls
API Abuse / Model Extraction	Repeated API queries used to copy or abuse models.	IP theft, service abuse	Authentication, rate limiting, monitoring
Cloud Misconfigurations	Weak cloud settings expose systems or datasets.	System or data compromise	Least-privilege access, audits, MFA

SECURITY TECHNIQUES

Securing cloud-hosted GenAI systems requires a layered security approach that addresses vulnerabilities across input handling, model training, deployment, and inference stages. AI firewalls [24] are commonly used as a first line of defense by inspecting user prompts and model outputs in real time to detect and block threats such as prompt injection, jailbreak attempts, and malicious instructions. These systems rely on both rule-based and semantic analysis techniques to enforce safety policies and prevent harmful or unintended model behavior [9].

Differential privacy is another key mitigation technique that reduces the risk of sensitive data leakage from training datasets. By introducing carefully calibrated statistical noise during training or inference, it ensures that individual data records cannot be reliably reconstructed from model outputs while still preserving overall model utility [10].

Federated learning further enhances privacy by enabling decentralized training, where data remains on local devices or institutional systems and only model updates are shared with a central server. This reduces the exposure of raw sensitive data but introduces challenges such as vulnerability to model poisoning and requires additional safeguards for secure aggregation [11], [20].

Secure enclaves and confidential computing technologies, such as Intel SGX and AWS Nitro Enclaves, provide hardware-level isolation for AI workloads. These environments protect model parameters and sensitive data even from cloud infrastructure operators by

encrypting memory and verifying execution integrity through remote attestation [13]. NVIDIA also extends this approach through confidential GPU computing for secure AI processing [16].

API security is critical because most GenAI systems are accessed through external interfaces. Strong authentication mechanisms, rate limiting, anomaly detection, and request monitoring are necessary to prevent abuse, scraping, and model extraction attacks. Without these protections, attackers can systematically query APIs to reconstruct or approximate proprietary models [8], [21].

Secure MLOps (Machine Learning Operations) practices ensure that security is embedded throughout the entire machine learning lifecycle. This includes encrypted model storage, cryptographically signed artifacts, secure CI/CD pipelines (Continuous Integration (CI) and Continuous Delivery/Deployment (CD), role-based access control, and continuous monitoring for model drift, adversarial behavior, and operational anomalies [6].

Adversarial training improves model robustness by exposing models to manipulated inputs during training. This enhances resistance to adversarial examples and prompt injection attacks [22], [23]. However, it increases computational overhead and requires continuous retraining to remain effective against evolving threats.

AI firewalls [24] act as a semantic security layer between users and generative AI models. They analyze both input prompts and output responses to detect malicious intent, prevent prompt injection, enforce data loss prevention policies, and ensure safe outputs.

Unlike traditional firewalls, AI firewalls [24] operate on contextual and semantic understanding. Despite their effectiveness, they may introduce latency, false positives, and require continuous updates to address evolving adversarial techniques.

CLOUD SECURITY FOR GENERATIVE AI SYSTEMS

Cloud-based environments play a central role in deploying and scaling generative AI systems due to their computational capabilities, elasticity, and integrated security tooling. However, they also introduce shared-responsibility security challenges, where both providers and users are responsible for securing AI workloads.

Major cloud ecosystems have developed dedicated frameworks for securing generative AI systems, focusing on identity management, model safety, secure APIs, and monitoring capabilities. These approaches align with broader AI risk governance frameworks such as the NIST AI Risk Management Framework [14].

A key focus area is secure model deployment and inference protection, where security controls are applied at the API and runtime layers to mitigate prompt injection, data leakage, and abusive usage patterns. These mechanisms are complemented by policy enforcement systems that regulate both input prompts and generated outputs [21].

Another important aspect is AI safety and responsible deployment, where automated filtering systems are used to detect harmful, biased, or unsafe outputs. These systems are increasingly integrated into cloud-based AI services to ensure compliance with ethical and regulatory requirements [17]–[19].

In addition, cloud environments emphasize secure identity and access management (IAM) to enforce strict authorization policies for AI model access. This ensures that only authenticated users and services can interact with sensitive AI resources, reducing the attack surface for unauthorized access and misuse [18], [19].

Continuous monitoring and threat detection systems are used to detect anomalies in API usage, model behavior, and system access patterns. These monitoring mechanisms are essential for identifying early-stage attacks such as model extraction attempts, prompt injection campaigns, and API abuse [17]–[19].

Overall, cloud-based generative AI security relies on a combination of infrastructure-level controls, model-level protections, and application-level safeguards, forming a comprehensive defense-in-depth architecture.

Figure 1 illustrates a defense-in-depth architecture for securing generative AI systems. The model is structured as a hierarchical stack where each layer enforces security controls that protect against threats specific to that level while also supporting adjacent layers.

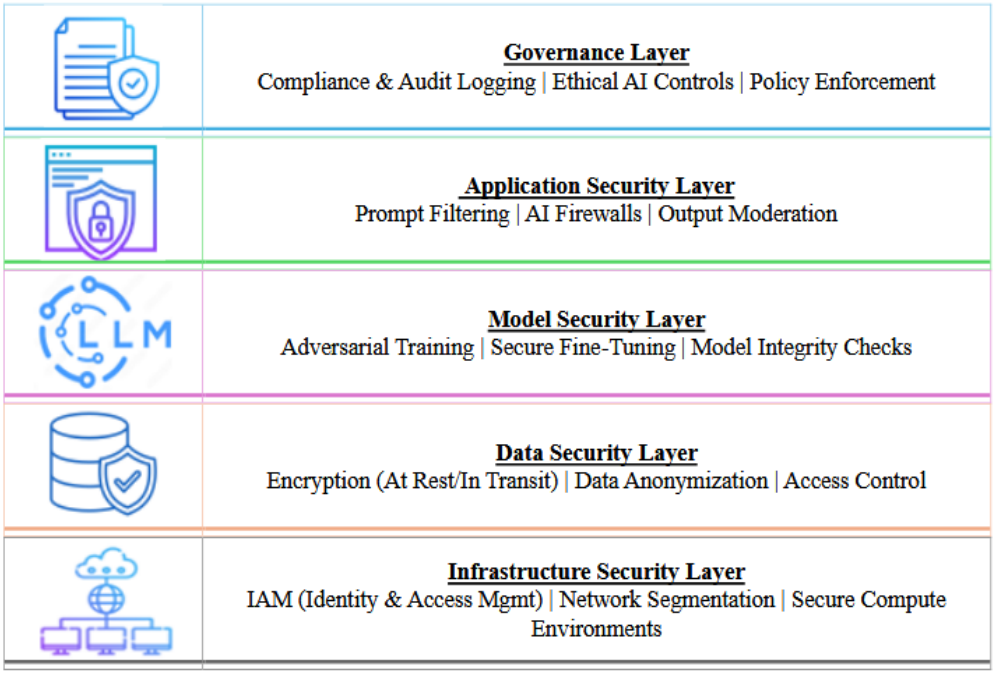

Figure 1: Layered Security Architecture for Generative AI Systems

GOVERNANCE AND POLICY FRAMEWORKS

Governance and policy frameworks are essential for ensuring that cloud-hosted GenAI systems are secure, transparent, and compliant with ethical and legal standards. The NIST AI Risk Management Framework (AI RMF) provides a structured lifecycle approach for managing AI risks through functions such as governance, mapping, measurement, and risk mitigation. It emphasizes core principles including transparency, robustness, accountability, and safety across AI system development and deployment [14].

At the regulatory level, the European Union AI Act establishes a risk-based classification system that categorizes AI applications into unacceptable, high, limited, and minimal risk. GenAI systems, particularly those deployed at scale, fall under stricter regulatory obligations that require transparency, human oversight, cybersecurity protections, and documentation of system behavior. Providers must also ensure that users are informed when content is AI-generated and that systems comply with safety and rights protections [15].

Cloud service providers also contribute to governance through built-in security and compliance tools. Platforms such as AWS, Microsoft Azure, and Google Cloud offer features like identity and access management (IAM), virtual private cloud (VPC) isolation, audit logging, content filtering, and secure private endpoints. These controls help enforce organizational policies and reduce exposure to unauthorized access and data leakage in GenAI deployments [4], [6], [12].

Organizations deploying GenAI systems are also expected to establish internal governance structures that define ethical usage, data handling policies, model auditing procedures, and incident response strategies. These frameworks often include AI ethics committees and require periodic third-party security assessments to ensure accountability and compliance with regulatory standards.

In addition, modern governance approaches increasingly emphasize transparency and explainability. Organizations are expected to maintain documentation of training data sources, model behavior, and decision-making processes to ensure auditability and build trust among users and regulators.

CONCLUSION

Generative AI is a transformative technology, but its deployment in cloud environments introduces significant security and privacy challenges. As adoption increases, it is critical to address emerging threats such as prompt injection, model extraction, and data leakage.

Combining AI-specific defenses such as firewalls and privacy-preserving techniques with secure cloud architectures like enclaves and federated learning can significantly enhance system resilience. Additionally, aligning technical practices with governance frameworks ensures that GenAI systems remain secure, ethical, and compliant with regulatory standards.

REFERENCES

[1] T. Brown et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS 2020), vol. 33, pp. 1877–1901.

[2] N. Carlini et al. (2023). Extracting training data from diffusion models. Proc. 32nd USENIX Security Symposium (USENIX Security ’23), pp. 5253–5270.

[3] OpenAI, (2022). Introducing ChatGPT. Retrieved from https://openai.com/blog/chatgpt

[4] A. Qayyum et al. (2020). Securing machine learning in the cloud: A systematic review of cloud machine learning security. Frontiers in Big Data, vol. 3, 2020, Art. no. 587139.

[5] Cloud Security Alliance, (2024). Top Threats to Cloud Computing 2024. Retrieved from https://cloudsecurityalliance.org/artifacts/top-threats-to-cloud-computing-post-pandemic-eleven-survey-report

[6] A. Luqman et al. (2024). Privacy and security implications of cloud-based AI services: A survey. arXiv preprint arXiv:2402.00896.

[7] F. Perez and I. Ribeiro, (2022). Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527.

[8] K. Zhao et al., (2025). A survey on model extraction attacks and defenses for large language models. arXiv preprint arXiv:2506.22521.

[9] S. Ahmadi, (2023). Next generation AI-based firewalls: A comparative study. International Journal of Computer (IJC), vol. 49, no. 1, pp. 245–262.

[10] M. Abadi et al. (2016). Deep learning with differential privacy. Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS '16), Vienna, Austria, 2016, pp. 308–318.

[11] K. Bonawitz et al. (2019). Towards federated learning at scale: System design. Proc. Machine Learning and Systems (MLSys), 2019, pp. 374–388.

[12] Microsoft, (2023). Confidential AI. Microsoft Learn, 2023. Retrieved from https://learn.microsoft.com/en-us/azure/confidential-computing/confidential-ai

[13] G. Chen et al. (2018). SGXpectre attacks: Stealing Intel secrets from SGX enclaves via speculative execution. arXiv preprint arXiv:1802.09085.

[14] National Institute of Standards and Technology, (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, 2023. Retrieved from https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

[15] European Parliament and Council of the European Union, (2024). Artificial Intelligence Act. Retrieved from https://artificialintelligenceact.eu/

[16] Microsoft Research, (2022). Powering the next generation of trustworthy AI in a confidential cloud using NVIDIA GPUs. Retrieved from https://www.microsoft.com/en-us/research/blog/powering-the-next-generation-of-trustworthy-ai-in-a-confidential-cloud-using-nvidia-gpus/

[17] Microsoft, (2025). Protect AI assets from emerging threats and vulnerabilities using Microsoft Defender. Microsoft Learn, 2025. Retrieved from https://learn.microsoft.com/en-us/defender-xdr/security-for-ai/defender-security-for-ai

[18] Amazon Web Services, (2024). Layer 3: Security and governance for generative AI platforms on AWS. AWS Prescriptive Guidance, 2024. Retrieved from https://docs.aws.amazon.com/prescriptive-guidance/latest/strategy-enterprise-ready-gen-ai-platform/security.html

[19] Google Cloud, (2025). Well-Architected Framework: Security, privacy, and compliance pillar. Retrieved from https://cloud.google.com/architecture/framework/security

[20] P. Kairouz et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning. vol. 14, no. 1–2, pp. 1–210.

[21] OWASP Foundation, (2024). Top 10 risks for large language model applications. Retrieved from https://owasp.org

[22] I. Goodfellow, J. Shlens, and C. Szegedy, (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[23] N. Papernot et al., (2017). Practical black-box attacks against machine learning. Proc. ACM Asia Conf. Computer and Communications Security (AsiaCCS), 2017, pp. 506–519.

[24] Akamai Technologies. Firewall for AI. Retrieved from https://www.akamai.com/products/firewall-for-ai