Securing Cloud AI Workloads: Protecting Generative AI Models from Adversarial Attacks

Author 1: Advait Patel
Affiliation: Broadcom
Email: advaitpa93@gmail.com

Author 2: Pravin Pandey
Affiliation: Tiffany & Co
Email: pravin.pandey@outlook.com

Author 3: Hariharan Ragothaman
Affiliation: Athenahealth
Email:hariharanragothaman@ieee.org

Author 4:Ramasankar Molleti
Affiliation:Options Clearing Corporation
Email: sankar276@gmail.com

Author 5:Ajay Tanikonda
Affiliation: Independent Researcher
Email:ajay.tani@gmail.com

2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC) | 979-8-3315-1888-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICAIC63015.2025.10848877

Abstract—Generative artificial intelligence models have brought about advancements in fields like healthcare and finance, as well as in autonomous systems; however, they also encounter notable security vulnerabilities, primarily when operating in cloud environments. These AI models can be targeted by attacks that involve altering input data to deceive the system into generating harmful or incorrect results. This study delves into the security issues that AI systems face in cloud setups, explicitly focusing on the dangers posed by adversarial manipulation of data integrity and the challenges of utilizing shared resources within multi-user environments. The text covers methods for defending AI models, like training and defensive distillation, to make them more robust against attacks. It also delves into security measures for the cloud, such as encrypted communications and robust authentication systems to safeguard data integrity. Furthermore, the importance of AI explainability and transparency in uncovering vulnerabilities and building trust is highlighted. The outcomes of security breaches emphasize the importance of having AI systems to avoid impacts on decision-making and broader ethical and societal concerns. The document also discusses research areas such as quantum algorithms and decentralized security structures to tackle evolving risks and safeguard the future of secure AI applications that generate content.

Keywords—AI explainability, adversarial attacks, cloud security, defensive techniques, generative AI

I. Introduction

This Generative artificial intelligence models are one of the most appreciated and thought-provoking enabling technologies, creating realistic images, text, media, and more across industries, including healthcare, finance, and autonomous systems. However, using them in the cloud offers different security risks than traditional computing platforms. Adversarial attacks refer to a complex threat in which inputs are slightly modified in a manner that will mislead these compelling models [1]. These remain threats, not limited to outdated computer issues but real-life dangers. For example, adversarial examples can lead to incorrect object detection in the perception systems of an autonomous vehicle or provoke severe consequences in a healthcare context, such as providing a wrong diagnosis. It is even worse at the moment due to the increased use of cloud platforms as the place where these generative AI models are hosted. Despite being flexible, resource-effective options for organizations, cloud environments are intrinsically vulnerable to a sweeping range of risks. Usually, data transfer processes, the sharing of infrastructure, and open interfaces in cloud systems are vulnerable to breaches of AI workloads [2]. When an adversarial attack targets these issues, it does not merely lead to technical failures but also erodes trust in AI-generated results and opens businesses to financial and reputational losses. This paper analyzes how generative AI workloads hosted on the cloud can be protected against adversarial attacks. It explores the threats in training and deploying these models, tactics used to counter adversarial inputs, and how this adversarial threat affects the overall decision-making of artificial intelligence systems. Realistic defense strategies must be emphasized, with generative AI progressing significantly, especially in defending the technology and its users. Therefore, this study aims to develop ideas for advancing the ongoing discussion on enhancing security for AI systems in cloud environments for their dependability and correct usage in risk-sensitive applications.

II. Literature Review

A. Generative AI in the Cloud

Generative AI is assigned as an innovative direction within artificial intelligence research and an area that involves the development of new content, ranging from images, texts, music, and videos. Generative Adversarial Networks (GANs) and other subsequent forms like GPT and DALL·E across industries have undoubtedly improved the quality and realism of outputs generated [3]. However, these models require heavy computations and memory storage during the training and implementation process. This critical element has made cloud computing a must-have in supporting generative AI since the tasks involved require such characteristics from the infrastructure. Amazon Web Services (AWS), Microsoft Azure, and Google Cloud offer services ideal for AI applications [4]. They provide high-horsepower GPUs and TPUs, various data storage choices, and a straightforward and convenient API for integration and development [5]. Table 1 presents vital differences between AWS, Azure, and Google Cloud, which organizations can choose from.

TABLE I. Differences Between AWS, AZURE, and Google Cloud

979-8-3315-1888-2/25/$31.00 ©2025 IEEE

Aspect	AWS	Azure	Google Cloud
Integration	Extensive third-party support	Best for Microsoft-based environments	Best for data-centric applications
Compute Services	EC2 instances with high flexibility	Virtual Machines with enterprise support	Compute Engine focused on AI/ML
Storage	S3, EBS with global reach	Blob Storage, SQL Database	Cloud Storage, BigQuery for analytics
Pricing	Pay-as-you-go with reserved instances	Pay-as-you-go with hybrid models	Pay-as-you-go, competitive discounts
Market Leader	Largest cloud provider	Strong in hybrid cloud and enterprise	Leading in data analytics and AI

Besides, generative AI models are built through multiple current technologies, such as cloud hubs, by training and deploying them in their organization without buying costly proprietary systems. For example, the process of training models such as OpenAI's GPT-4 requires processing large datasets and the execution of complex algorithms—something that is only possible using cloud resources [6].

Cloud-generative AI can be applied across industries to change the functioning of companies and organizations. In healthcare, these models diagnose images, create datasets for experimentation, and present individual health recommendations. The finance sector enjoys AI-created finance reports and uses AI to generate market reports and detect fraud. In entertainment, generative AI is at the forefront of containing new ideas in script image and video writing. It describes how the AI models hosted in the cloud, which control autonomous systems ranging from vehicles to robotics, continually perform real-time simulations to generate synthetic data to train AI models.

The benefits of generative AI hosted on the cloud include the following: The cloud's flexible resource provision makes it easy for an organization to provide high numbers of resources during the significant training phase and fewer resources during less demanding phases. The affordability of cloud services, which eliminates the need for substantial upfront hardware investments by allowing businesses to pay for the resources they use merely, complements this elasticity.

B. Adversarial Attacks on Generative AI

The emerging generative AI models have gained rapid popularity across the fields, which has created a massive potential for novelty. But it has simultaneously made these systems vulnerable to complex threats, especially adversarial ones. These attacks include invading and manipulating parameters in an AI model to derive a wrong or damaging result. It is essential to classify these adversarial attacks and know how the attackers work if we need to design some countermeasures against them to make AI systems reliable.

a. Types of Adversarial Attacks

Adversarial attacks on generative AI models generally fall into three broad categories: they have been further divided into evasion attacks, poisoning attacks, and inference attacks [7]. Each presents different complexities to the security and effectiveness of the AI systems being developed today. Evasion attacks are characterized by manipulating the input data feeds to introduce slight variations while feeding them to the AI model to fool the model. For instance, an image might have slight modifications that the attacker makes; it would result in a generative AI model failing to classify the image correctly or even providing poor-quality output. This kind of attack is quite a threat in all the related applications, such as automobiles.

Poisoning attacks begin at the training phase of an AI model [8]. Deliberately, the attacker poisons the training data set, which creates a different model behavior. For example, in a financial application, an attacker can inject fraud patterns into the training data, and as a result, the generated model will generate the wrong market signals. Poisoning attacks continue to erode the model's trustworthiness over its lifetime and are very dangerous. Also, inference attacks are directed at stealing helpful information from an AI model [9]. While the model's output is obliterated, the attacker can deduce features regarding the dataset used during training. This attack is brutal in healthcare: an innocent generative AI model may leak specific data about a particular individual from the training set, violating the patient's rights to privacy.

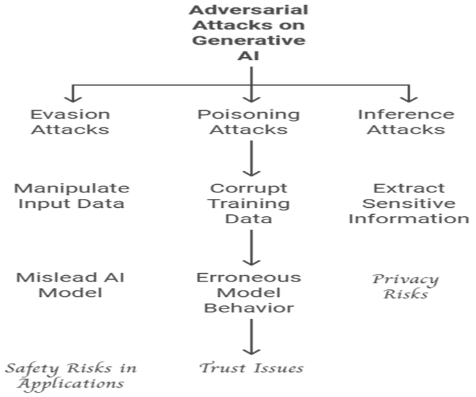

Figure 1. Types of Adversarial Attacks on Generative AI

b. Methods Used by Attackers to Target Generative AI Systems

Attackers also use various strategies to attack generative AI systems depending on the specificities of the generative model's architecture and the environment where the model is being used. Gradient-based optimization is a popular approach, where attackers use model parameter files and generate adversarial inputs [10]. Since the gradient representing the direction in which the given model adjusts during training can be computed, the input data can be tweaked towards a specific, unsafe outcome. For example, in an evasion attack, the attacker might use gradient information to modify an image in a way that is classified as entirely different by the generative model but looks the same to the human observer (see Figure 2).

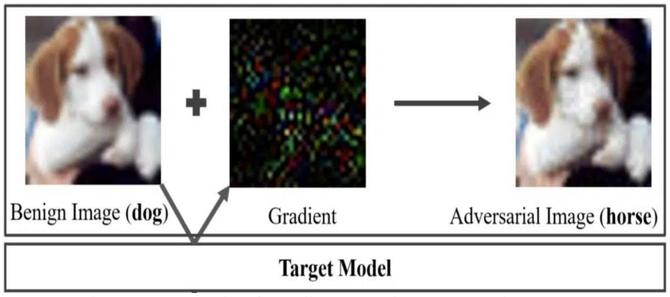

Figure 2. Example of Gradient Attack on Generative AI [10]

As presented in the Figure above, gradient-based attacks can craft adversarial images to fool the target model, for instance, changing the predicted class from dog to horse.

Besides, some methods target vulnerabilities in open APIs and public interfaces, which almost every generative AI system makes public to facilitate integration. Using these interfaces, attackers can continuously query the model and analyze the result to generate more perturbed inputs. Such query-based methods are frequently used in inference attacks. For instance, an adversary might feed carefully crafted inputs into a text generation system to learn the training data, such as business information that a company invested in developing the model.

Also, poisoning attacks frequently use data injection approaches. Adversaries, therefore, inject contaminated data into a repository or dataset where developers may later incorporate it. Regarding a real-world application, the data set of autonomous vehicle systems might be corrupted by providing fake patterns or labels [11]. After this dubious data feeds the model to develop its decisions, there is a high possibility of unpredictable or dangerous emotional responses when utilized. The methods are more hazardous because generative AI systems are relatively difficult to understand and interpret. Most of them are based on big-sized data and multi-layer convolutional neural network strategies for more secure filtering and handling of adversarial interference, thus generating problems in efficiently realizing those techniques.

Security Challenges in Cloud AI Workloads

Since organizations continue to utilize cloud environments to host and run their AI models, the specific risks associated with such a deployment model are prime to create security issues. Generative AI models hosted in the cloud are deployed in data-sharing environments, multi-tenancy-based systems, mutual resource access, etc. Nevertheless, these features also open networked environments to various security risks. This section discusses the issues related to cloud-based AI models regarding data risks, data integrity in the training process and during inference, and risks arising from multitenant environments.

a. Unique Vulnerabilities of Cloud-Hosted AI Models

AI models hosted in the cloud work in contexts where continuous data exchange is paramount; thus, data security during data transfer is paramount. Transfers between storage, computation nodes, and client applications are over networks; they introduce the risks of eavesdropping or even data tampering by attackers. For instance, the compromisers may insert adversarial noise or steal information from information transmission flows if data transmission is dynamically unencrypted. Such risks are even more significant in applications in which the data is real-time, including trading and self-driving vehicles, where even latency or manipulation can cause severe problems [12].

This is compounded by the fact that most cloud-hosted systems use open interfaces and APIs. While APIs provide an interface for applications to work together, they also offer a gateway for hackers. Finally, these interfaces allow attackers to perform unauthorized queries, overload systems with DoS attacks, and potentially retrieve API misconfiguration to obtain model parameters and datasets. For instance, a vulnerability in an API link of a generative AI sector may lead to unauthorized modification of data generated by the AI applying given input or even malicious probe into the AI model's activities, which is a potent recipe for breaking user privacy and stealing intellectual property.

b. *Challenges in Ensuring Data Integrity During Model Training and Inference*

In training and, mainly, inference processes, maintaining data integrity in cloud settings constitutes a persistent concern. During training, models depend on large volumes of data that must be compiled, uploaded, and passed through a secure system. However, these datasets can be easily poisoned where the attackers ensure the model learns from improper or motivated data. It is made worse as the poisoning attacks impact the model even during training, thereby negatively influencing decision-making throughout the model's life cycle. Concerns about data integrity are not limited to the learning phase; the real-time inputs also affect the outputs produced by the AI models [13]. For instance, generative AI models in manipulation use patient data to produce diagnostics regarding disease. If attackers modify these inputs during transmission, the obtained outputs may confuse healthcare professionals and threaten lives. Further, since inference happens in real-time, especially in cloud-hosted systems, the available time to counter such manipulations is slim, which increases the risk. Yet another problem relates to the use of third-party data as well as pre-trained models, which is typical for many industries.

c. *Issues Related to Multi-Tenancy and Shared Resources in Cloud Environments*

One fundamental characteristic differentiating cloud computing from traditional systems is multi-tenancy, whereby multiple organizations or applications share the same physical equipment. This feature offers cost savings and possibilities for scale, but it also presents dramatic security considerations. Storage, processing power, and networking infrastructure are examples of shared resources that can lead to cross-tenant attacks, in which one tenant takes advantage of weaknesses to access the data or workloads of another [2]. With generative AI workloads, cross-tenant risks can be presented in one or many forms. For example, suppose an attacker wants to compromise another tenant's AI model. In that case, one may use side-channel attacks to observe resource usage patterns and predict certain information about the desired model

Vulnerabilities in the cloud also mean the environment is communal, thus making it difficult to identify malicious practices [14]. One tenant's application can impact others, often leading to security breaches that generate cascading consequences. For instance, an attacker taking advantage of a vulnerability in a shared hypervisor could get complete control of many virtual machines, which can be those that work with crucial AI models. While they adopt measures to segment and secure tenant workloads, these measures could be better designed.

III. Methods: Techniques for Securing Generative AI Models

A. *Defensive Training*

a. Adversarial Training

Adversarial training is a preventive approach to reinforcing generative AI models against adversarial perturbation attacks. This technique adds new values or inputs known as adversarial examples to the training data set. Feeding the model such inputs during training helps it learn how attackers' inputs work to avoid them during deployment [15]. For instance, an adversarial attack in an image generation system can be achieved by adding noise to the almost invisible image without affecting the model. The inputs can be classified correctly by repeatedly exposing cases to the model, and the adversarial noise does not contaminate the output.

b. Defensive Distillation

Another method known to help prevent adversarial attacks is defensive distillation. It does this by training the student network archetype to mimic the behavior of a more intricate teacher network in return for smoothing the decision boundaries of its field [16]. This reduces the model's susceptibility to input perturbations, making it difficult for many adversarial examples to work. For example, in performing defensive distillation of a text generation model, knowledge transfer may involve transferring features learned by a large language model to a less complex model that can resist adversarial inputs. The simplified decision boundaries also allow the student model to maintain stability, although the attackers try to add some sophistication to the input text.

c. Model verification and certification

Model verification and certification ensure that a generative AI model works as required without compromising. Verification, conversely, scrutinizes the model against predetermined security standards to look for flaws. At the same time, certification is a more formal way of ensuring that a model conforms to requisite security standards. For instance, a healthcare generative AI model might be certified to ensure it complies with the data protection rules and is resistant to adversarial examples. Such processes improve confidence in the model's output, particularly in critical applications.

B. *Cloud-Specific Measures*

a. Using Encrypted Communications and Storage

Data encryption is essential for protecting data during its transfer and storage in systems hosted in the cloud, including AI systems. End-to-end encryption retains data confidentiality between users, storage systems, and computational nodes during transit. Similarly, encrypted data at rest safeguards the stored information, including the databases or other storage media [17]. For instance, outputs and inputs can be encrypted in generative AI models for financial applications running on the cloud. If, for example, an attacker gets access to the storage system, the data/components encrypted will be difficult, if not impossible, to interpret without the decryption keys. Cloud providers also provide essential encryption services, such as Amazon Web Services Key Management Service (KMS), to help in the process.

b. Implementing Robust Authentication and Authorization Systems

Authentication controls the AI model's interaction with only those allowed, and authorization controls what the individual or system can do. Multi-factor authentication (MFA) is a process where users can corroborate their identities in multiple ways, for instance, through password and fingerprint scans [18]. Moreover, when implementing role-based access control (RBAC), permissions are also limited according to user roles, practically decreasing the possibility of abuse. For instance, a developer best develops some areas of a model, while other people, such as the data owners, can work on different places, like the training data. These measures help protect the organization from unauthorized access from extraneous attackers and intruders.

c. Continuous Monitoring for Anomalies in Model Behavior

The monitoring type ensures that the performance of the generative AI models is checked in real-time to identify any anomalies. A situation where such hosts are detected may depict an extension of a continued attack, a data integrity problem, or a system failure. For instance, in an image synthesis application, a spike in the number of generated images that look unnatural, distorted, etc., may indicate an attack. Supervisory tools, such as Google Cloud AI Platform Monitoring, are offered in the cloud. This tool tracks behavior and detects anomalies in one's model. Other features include the ability to be alerted and log threats more automatically.

IV. Results and Discussion

A. *Role of AI Explainability (XAI) and Transparency in Enhancing Security*

AI explainability and transparency have become foundational requirements for ensuring the safe implementation of AI. Interpretability allows developers, users, and auditors to understand how specific AI models made decisions throughout the entire AI development and implementation process [19]. Transparency means we must keep track of the model design, the training mechanisms, and the various model setup processes for maximal accountability.

a. Enhancing Threat Detection

XAI approaches can be used to understand model behavior weaknesses due to their ability to explain the reasoning. For example, in a text generation system, XAI can uncover that specific input tokens have a particular weight, which an attacker can misuse. Knowing these decision paths is also essential, as it lets developers strengthen the weak points and protect them from adversarial tricks.

b. Building User Trust

Transparency increases the reliability of the information between users and is especially important for crucial and frequently discussed topics such as health and economics. Written documentation and written or verbal description of strengths, weaknesses, and security features for the model butt down their expectations [19]. They inform best practices for transparent combativeness against adversarial attacks or system failure diagnosis and remediation.

c. Regulatory support

Various industries are more concerned with data usage and system security rules in existing organizations. Under explainability and transparency, proof of compliance with such standards becomes far more accessible. For instance, an ML model in a healthcare provider can demonstrate to auditors how data is protected and how generative AI prevents inaccuracies in its results.

B. Integrating Techniques for Coprehensive Techniques

For generative AI models to be secure, organizations should combine these defensive strategies with cloud-specific safeguards into a unified security architecture.

Initial Risk Assessment

Find threats in the model of artificial intelligence and threats in the cloud.

Implementation of Defensive Techniques

Apply adversarial training during model development and apply defensive distillation and verification processes.

Cloud Security Configuration

Encryption of data transfer and storage should be possible for all data. Develop authentication and authorization services that are unique to the users.

Deployment and Monitoring

Never let up in the surveillance of model behavior for the least sign of deviation from the norm. Always use the explainable tool to understand why some outputs differ from the expected ones.

Periodic Review and Updates

If necessary, risk assessment, security measures change, and the model's re-certification.

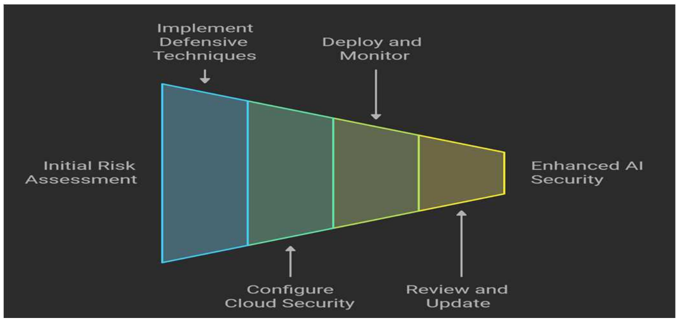

Figure 4. Comprehensive Security Integration for AI

C. Impact of Security Breaches on AI Outputs

The risk of security breaches can also present generative AI systems with adverse outcomes that compromise the reliability, safety, and ethics of programs' operations. Sneaky interference intrudes on mechanisms in AI models, creating false or destructive outputs that disseminate across numerous fields. For example, in the healthcare system, an adversarial attack could change the medical imaging data a little, thereby giving an AI diagnostic system the wrong information about growths and labeling some benign instead of malignant [20]. For instance, such mistakes can delay the chances of treatment and result in patient health issues. In self-driving cars, actual inputs—the traffic signs changed—may lead to misinterpretation of the surroundings and accidents. In the same way, in the finance area, a potential attacker can shift trading algorithms with fake market data, which makes trading incorrect and causes considerable losses.

End users and industries dependent on AI techniques will feel grave penalties. To consumers, trust in AI technologies is reduced when the offered output is unreliable or dangerous. To businesses, the breaches will attract fines and loss of reputation and might hinder their operations. For instance, the attacker can alter one AI content creation model to create fake news and/or negatively report on a particular company brand to the public, thereby causing more social harm. When AI security is threatened, issues are addressed on a larger ethical and societal scale. Specifically, malicious actions that circulate prejudice, fake news, or objectionable information widen social gaps and undermine people's trust in AI solutions. Furthermore, relying on AI outputs in decision-making—concerning hiring policing or public health—reinforces bias and unfairness. Proper security measures for AI systems, vigorous testing to avoid loopholes, and appropriate ethical standards are needed to manage these risks. The lack of these protections means that daily utilization of AI in sensitive areas potentially poses a massive loss to everyone.

D. Future Directions and Research Gaps

Protection of AI workloads requires ongoing advancements to counter advanced risks and risks in the application. Writing, it is realized that as subversion techniques develop in form, so do the functions, which act as counter iterations to the modern adversary waves, consequently leading to the advancement of technologies and methodologies that work towards improving the strength within generative AI systems. New techniques like quantum-resistant algorithms can enhance encryption protection against possible quantum-based cyber assaults [21]. These algorithms safeguard encrypted conversations for telecommunication networks and guard model parameters against quantum threats in the future. Another promising trend is integration with decentralized AI security measures. Drawing from ideas in blockchain, decentralized frameworks can offer audit trails for data and model usage to check for change and ensure that nobody has manipulated the AI systems. Moreover, XAI tools must extend their capabilities to provide more profound explanations of model reasoning to improve anomaly and adversarial manipulation identification. Federated learning, which trains models by periodically exchanging updates between the devices without sending original data to a central point, also addresses concerns revolving around centralized data stores.

Future research should look at integrating adversarial training with other security frameworks, such as those that employ real-time systems monitoring. Further, it is crucial to examine what happens when AI ethics and security are taken together, especially to understand what would happen if he, she, or it were to incorporate society's values into the models while at the same time reducing bias and identities that make models insecure. Another area of study is the use of AI in threat detection systems that would detect and prevent attacks even before being instigated. If these gaps are not addressed, it will be impossible to guarantee that AI systems are resistant, trustworthy, and secure in the growing interconnected and adversarial environment.

V. Conclusion

Protecting generative AI models in the cloud is a challenge and, at the same time, an urgent need. This paper has underscored the weaknesses of cloud-hosted AI workloads and the threats of adversarial incidents, data integrity, multi-tenancy, and shared assets. The threats above can be well defended by adversarial training, defensive distillation, and model verification. Additional and specific security measures for the cloud environment, such as encrypted communication, strict and multi-layered user authentications, and real-time monitoring, also add strength to the security layer, while explainability and transparency help to increase the ability of threat detection and user confidence and trust. The consequences of security breaches are deep-rooted, as adversarial manipulations may lead to wrong decisions, compromise trust, and worsen ethical dilemmas. Mitigating these risks requires a systems approach involving multiple layers and constant change as threats emerge. Protecting generative AI models in a cloud environment is as much a necessity of technology as it is of society. The drive for AI security and stability emerges because AI is increasingly implemented in industries sensitive to disruptions, including healthcare, finance, and transportation. Continuing research in quantum-resistant algorithms, decentralized security mechanisms, and compound defenses will help protect these revolutionary technologies.

References

[1] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial Attacks and Defenses in Deep Learning: from a Perspective of Cybersecurity," *ACM Computing Surveys*, vol. 55, no. 8, Jul. 2022, doi: https://doi.org/10.1145/3547330.

[2] M. Dawood, S. Tu, C. Xiao, H. Alasmary, M. Waqas, and S. U. Rehman, "Cyberattacks and Security of Cloud Computing: A Complete Guideline," *Symmetry*, vol. 15, no. 11, pp. 1–33, Nov. 2023, doi: https://doi.org/10.3390/sym15111981.

[3] S. Bengesi, H. El-Sayed, M. K. Sarker, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and...," *ResearchGate*, 2024. https://www.researchgate.net/publication/380392586_Advancements_in_Generative_AI_A_Comprehensive_Review_of_GANs_GPT_Autoencoders_Diffusion_Model_and_Transformers

[4] Digital Ocean, "Comparing AWS, Azure, GCP | DigitalOcean," *Digitalocean.com*, 2024. https://www.digitalocean.com/resources/articles/comparing-aws-azure-gcp

[5] P. Borra, "COMPARISON AND ANALYSIS OF LEADING CLOUD SERVICE PROVIDERS (AWS, AZURE AND GCP)," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 15, no. 3, pp. 266–278, 2024, doi: https://doi.org/10.17605/OSF.IO/T2DHW.

[6] S. M. Kerner, "GPT-4o explained: Everything you need to know," *WhatIs*, Jul. 19, 2024. https://www.techtarget.com/whatis/feature/GPT-4o-explained-Everything-you-need-to-know

[7] Palo Alto, "What Is Adversarial AI in Machine Learning?," *Palo Alto Networks*. https://www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning

[8] N. Hassan, "What is data poisoning (AI poisoning) and how does it work?," *Enterprise AI*, May 2024. https://www.techtarget.com/searchenterpriseai/definition/data-poisoning-AI-poisoning

[9] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models," Feb. 2019, doi: https://doi.org/10.14722/ndss.2019.23119.

[10] M. Ivezic, "Gradient-Based Attacks: A Dive into Optimization Exploits," *Securing.AI - Marin Ivezic*, Jan. 03, 2023. https://securing.ai/ai-security/gradient-based-attacks/

[11] The Lasso Team, "What is Data Poisoning? Types, Examples & Best Practices," *Lasso.security*, Aug. 21, 2024. https://www.lasso.security/blog/data-poisoning

[12] A. Alquwayzani, R. Aldossri, and M. Frikha, "Prominent Security Vulnerabilities in Cloud Computing," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, Jan. 2024, doi: https://doi.org/10.14569/ijacsa.2024.0150281.

[13] Z. Chain and M. Sharif, "Data Integrity Challenges and Solutions in Machine Learning- driven Clinical Trials," Aug. 19, 2023. https://www.researchgate.net/publication/373214664_Data_Integrity_Challenges_and_Solutions_in_Machine_Learning-_driven_Clinical_Trials

[14] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions," *Electronics*, vol. 12, no. 6, pp. 1–42, Mar. 2023, doi: https://doi.org/10.3390/electronics12061333.

[15] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial Training Methods for Deep Learning: A Systematic Review," *Algorithms*, vol. 15, no. 8, p. 283, Aug. 2022, doi: https://doi.org/10.3390/a15080283.

[16] "defensive distillation - Ella," *Ella*, Nov. 14, 2023. https://ella-group.io/en/encyclopedia/defensive_distillation/ (accessed Nov. 25, 2024).

[17] IBM, "What is encryption? Data encryption defined," *www.ibm.com*, 2022. https://www.ibm.com/topics/encryption

[18] M. O'Connor, "What is Multi-Factor Authentication (MFA) and How does it Work?," *RSA*, Oct. 28, 2021. https://www.rsa.com/resources/blog/multi-factor-authentication/what-is-mfa/

[19] D. Rathnayake, “AI Transparency: Why Explainable AI Is Essential for Modern Cybersecurity | Tripwire,” *www.tripwire.com*, 2024. https://www.tripwire.com/state-of-security/ai-transparency-why-explainable-ai-essential-modern-cybersecurity

[20] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: https://doi.org/10.1126/science.aaw4399.

[21] NIST, “NIST Releases First 3 Finalized Post-Quantum Encryption Standards | NIST,” *NIST*, Aug. 13, 2024. https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards