



SECURITY IN GENERATIVE AI SYSTEMS: A COMPREHENSIVE LITERATURE REVIEW

Mohammed Basheer*¹ and Ebby Darney²

¹Research scholar, LIPS Research, EIU France.

²Professor, Dept of EEE, Rajarajeswari College of Engineering, Bengaluru, India.

¹<https://orcid.org/0009-0005-7181-8154>^{ID}, ²<https://orcid.org/0000-0002-2559-3442>^{ID}

Email: *basheer_s26@yahoo.com, pebbydarney@gmail.com

ARTICLE INFO

Article History

Received: March 10, 2026

Reviewed: April 10, 2026

Accepted: July 10, 2026

Published: August 31, 2026

Keywords:

Generative AI,
AI security,
Machine learning, privacy,
cybersecurity,
AI governance,
NIST AI-RMF,
risk management,
EU AI Act.

ABSTRACT

Gen AI systems such as LLM's, code assistants, image generators and personal assistants are growingly integrated with productivity and critical business applications across different domains of organizations such as security, testing, code review and automate day to day tasks. Their functionality depends on different training models, inputs and instructions and those models are integrated with various applications via API's and connectors. In addition, there are many practical use cases where Generative AI LLM's are already operationally embedded with many industrial equipment's. This paper presents a detailed literature review of security, privacy concerns and compliance risks that are evolving day to day. This review refers to various published articles and subjects related to AI security from security perspective which addresses various security concerns specific to Gen AI and the need for evolution to adopt and comply with the evolving regulations. Also, the paper refers to various publications related to the role of AI in security, defense mechanisms to safeguard AI models and its applications, different approaches on risk management and so on. Based on these references, we also discussed the advantages and limitations of adopting NIST AI Risk management framework (AI-RMF) which is an adaptive governance guidance for Generative AI security at its lifecycle. The review finally highlights Assessment gaps and proposes a governance-tied adaptive security framework for Generative AI.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Artificial Intelligence (AI) has moved from theoretical constructions and narrow expert systems to large-scale, data driven models. This acts as general-purpose problem solvers across many domains. Recent advances in deep learning, especially transformer-based architecture, have enabled Generative AI (GenAI) systems capable of producing high fidelity text, images, videos, code, and synthetic data. These systems are deployed as conversational agents, content creation tools, decision-support modules, and programmable APIs integrated into broader sociotechnical systems. Even though this technological advancement in Gen AI makes it powerful but also introduces new and even complicated security and privacy risks. GenAI models are typically trained on large, heterogeneous datasets, sometimes scraped from the web or sourced from third-party providers, raising concerns about data provenance, consent, intellectual property, and memorisation of sensitive information. [1-6]. From an organizational viewpoint, Gen AI is now recognised as a paradigm shift because, compared to traditional software building process which encodes logic as explicit instructions the AI systems infer the behavior from the sourced data and optimisation objectives. This behavior makes it force a re-examination of how risk, accountability, and assurance should be defined and managed. Regulatory initiatives such as the EU AI Act explicitly require high-risk AI systems to be robust against cyberattacks and to undergo security-oriented risk assessment and conformity assessment. This paper focuses on security in Generative AI systems—security for AI, rather than AI for security—and proposes a lifecycle- and governance-oriented view of GenAI security. [7-10].

II. FOUNDATIONS: AI, GENERATIVE AI, AND SECURITY CONCEPTS

II.1 AI, MACHINE LEARNING, AND DEEP LEARNING

AI is commonly defined as a machine-based system that receives the set of human-defined objectives, can make predictions, recommendations, or decisions influencing real or virtual environments. Machine learning (ML) and deep learning (DL) are subsets of AI. ML focuses on learning patterns from data, while DL uses layered neural network architectures to capture complex, non-linear relationships. From the security governance, compliance and risk perspective, AI systems should be treated as a complex information system as they are relying on data, user instructions, and work by their own internal logic and memory. This improves business value, also at some point, it impacts business and individuals considerably.

II.2 GENERATIVE AI ARCHITECTURES AND APPLICATIONS

Generative AI models learn from huge subsets of data distribution and generate their own new samples. Generative AI discovers the pattern and learns from the provided rules and provide context rather than copying the initial information provided. It includes architectures like transformer-based models, Flow based models, Neural radiance fields, Latent diffusion models, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs etc. Due to advancement in technologies and research, these architectures are being evolved in such a way as it is widely used in many applications of productivity and entertainment such as text generations, video generations, personal assistance, Anomaly findings etc. Due to this broader applicability, risks like deep-fakes, manipulation and coordinated disinformation are evolving against individuals and organizations. Some surveys recently provided detailed information about the architecture of generative AI with its applications and implications, with notable evidence on adversarial impacts in real time. However, the productivity and benefits remain the same with the mindset of mindful usage of AI across various use cases.

II.3 AI FOR SECURITY VERSUS SECURITY FOR AI

The literature differentiates AI for security from security for AI. AI for security covers AI-enabled threat detection, automated patch management, anomaly detection, endpoint protection, network defense, phishing campaigns, security posture management, vulnerabilities exposure scoring, compliance automation and adaptive authentication. Security for AI focuses on maintaining data integrity, robustness, access control, logging, and governance for AI systems, including protecting training pipelines, preserving architecture, neural thinking ability, model artefacts, and inference endpoints from attack. Most cybersecurity use-cases focus on adapting AI for improving security. This review complements that by concentrating primarily on security for AI in the Generative AI context, while using AI for security literature to understand its interactions and applicability. [11], [12].

III. RESEARCH METHODOLOGY

To ensure a comprehensive and clear understanding of the current state of security in Generative AI, this review follows a structured literature search methodology. The objective was to identify appropriate peer reviewed articles, conference proceedings, current trends and reputable industry reports that address the intersection of Generative AI with cybersecurity, privacy, integrity, compliance and governance along with its applications and challenges.

III.1 SEARCH STRATEGY AND DATA SOURCES

We have conducted this review by searching through major academic databases, including Springer, IEEE, ACM Digital Library, arXiv and Google Scholar. The search strings combined keywords related to AI architectures such as “Generative AI”, “Large language models”, “Security risks in AI”, “Generative AI models” with security and risk terms such as “Deep fakes”, “Phishing and malware crafting”, “Code of evasion”, “Misinformation Ops”, “Polymorphic malware”, “Jail breaking”, “Data Poisoning”, “Prompt Injection”, “AI Regulation” [6], [11-15].

III.2 INCLUSION AND EXCLUSION CRITERIA

The retrieved journal publications are filtered via Prisma method to include only the last 3 years and latest articles to make sure the content to be included are up to the recent research and use cases. However, the concept been understood from multiple sources, but the references are included limited to 7 papers to narrow down and mimic the ideas shared by different authors. We only choose papers highlighting only Generative AI related issues and not on the entire AI as a subject to make sure that the issues are highlighted within the boundary of Gen AI. We have eliminated papers which are too generic to AI subject as a whole or those without any ties with Generative AI, its applications and implications. The key qualifier includes security concerns on Generative AI rather than application of AI in security unless it was linked to governance and security concerns of Generative AI. Based on our cybersecurity view, we explored studies on Gen AI relevant security gaps, grey areas, compliance concerns, misuse and impact which makes the primary filtering point of inclusions.

III.3 THEMATIC ANALYSIS

From my cybersecurity experience, grouping papers into threat patterns, data related risks, compliance mismatch, AI's pros and cons, and management frameworks gives the complete picture of GenAI dangers. In accordance with the experience, the selected findings were analyzed and organized into topic groupings and content clustering based on below focus.

(1) Current impact and implications of AI; (2) Security of AI from Governance, Risk and compliance perspective; (3) Widely impacting security concerns of Generative AI; (4) Challenges of AI to adapt with compliance standards and frameworks; (5) Implications and adversarial impact of AI in its most common applications; (6) Evolving usage and application of Generative AI in real world; (7) Methods to optimize AI to comply with GRC (Governance, Risk Management and Compliance) [16], [8].

IV. THREAT LANDSCAPE FOR GENERATIVE AI

As always, technology comes with challenges and those challenges pave way for resolutions. So as is Generative AI, which is nowadays a more adaptable technology which occupies its applicability in almost most of the areas of professional, personal and social usage which makes it open all kinds of fresh ways to attacks, impersonations, misuses and other impacts. This makes it questionable from the aspects of technological, ethical, social, legal viewpoints. In a simplified way, the below depicted figure explains the various security threats and manipulation possibilities of generative AI in its lifecycle from technological, human and data perspective. I have seen these risks play out in social media nowadays where people have started impersonating, providing fake information, news and other contents. Even I have encountered similar limitations in the structure and well created prompt which I have created for the healthcare IT service provider.

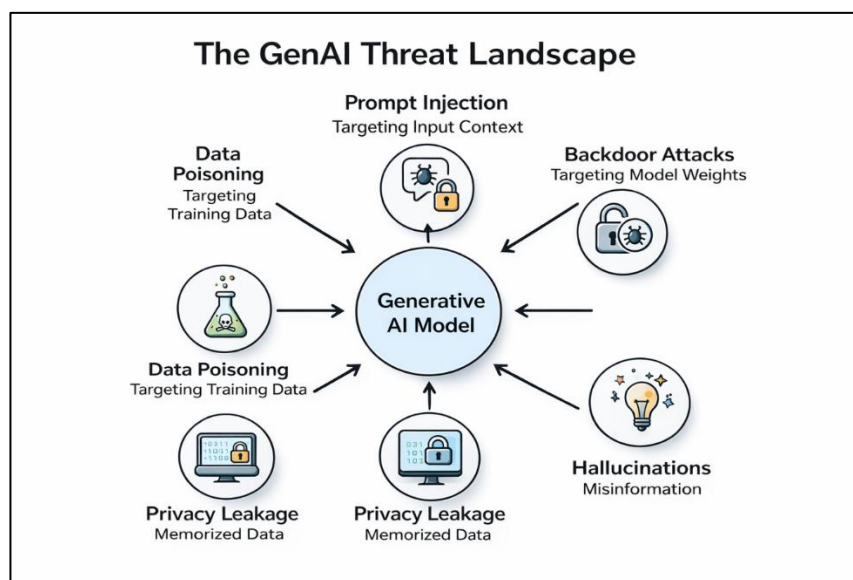


Figure 1: Various security threats targeting Generative AI lifecycle.
Source: Authors (2026).

IV.1 DATA POISONING AND BACKDOOR ATTACKS

Data poisoning and back door attacks are most common issues of generative AI nowadays where people started adapting and even got addicted to its usage without verification or cross review. With this open trust issue, hackers and falsifiers inject malicious samples of training data which in turn alter the model and its associated data to corrupt its samples out of its originally stored / trained data. Some attackers trick users to capture ideal and sensitive information via backdoor and make use of it to impact personal, professional and social aspects. These techniques are largely achieved by using triggers which appear to be activated upon certain situations without the knowledge of users who are interacting with the Generative AI. Those triggers degrade the quality of output and accuracy by manipulating the models, trained data and samples. These kinds of attacks are much more impactful mostly on basic and customized models than those of large models which have at least basic security being positioned to address these attacks. [14], [15].

IV.2 ADVERSARIAL PROMPTS AND INPUT MANIPULATION

Another notable threat on GenAI systems are the prompt injection and input manipulation attacks, where the prompt itself tricks the users to input key essential inputs which in turn will be used to manipulate the original prompt to cause damage as well as to some extent to steal sensitive information. These kinds of injections can either be found in documents and attachments which will skip the security checks of GenAI models and lead to security concerns and prompt manipulations. These are most common nowadays due to the flexibility of easy sharing, API integration ability and browser based. Simplest way to explain this GenAI hole is that the prompts and inputs hide within certain forms of emails and documents to make it look legitimate but once triggered or trusted, it does its manipulation in many ways. [13], [6]

IV.3 MODEL EXTRACTION, MEMBERSHIP INFERENCE, AND INVERSION

These concepts are few indirect ways of counterfeiting, copying and predicting identities of well-trained private AI models. Model extraction is something that matches the results score of a well-trained and private AI with the one created by a publicly available GenAI. This can be achieved by training the public or free AI models with the same answers and the results provided by the private well-trained AI. By this way, the AI brain will get stolen without stealing the model itself but by just mimicking the results and training your own AI models with those results. Membership inference is like data fingerprinting, where based on the results and confident score, it's possible to detect the existence of data record and the identity of input in the other private trained model.

Model inversion is a kind of Reverse engineering secrets, where the input is been checked for its confident score and then based on the high score, the private information can be rebuilt or synthesized. [14], [5].

IV.4 DEEPFAKES, MISINFORMATION, AND SOCIAL ENGINEERING

Deepfakes are another technique of collecting raw materials, information via data harvesting and use it as a source of truth to fake by impersonating the exact attributes of those collected information and sources. The original identity was then used to synthesize misinformation and to draw social engineering victims. A forger AI can be used to compete the detective AI by fooling it with high percentage of pass rate against its results. For example, an user identities from images and videos can be collected from multiple sources, a forger AI will use it to exactly impersonate the person and to train the detective AIs to prove it was right. It will then be used in social engineering to fake information. Nowadays Deepfakes are more challenging as the forged information is nearly perfect with the actual information and identities. These are newly evolving dangers in AI deployments as new projects quietly reshape the entire security landscape. Once known as a productive tool, it is now forcing security / risk teams to rethink it. [12], [17].

IV.5 SYSTEM-LEVEL AND OPERATIONAL RISKS

Due to the over dependencies and adaptations of GenAI's, Operational, system level and organizational risks are growing due to various reasons such as prompt exploitation, Insufficient compliance and logging monitoring, Unidentified security checks and lack of forensic evaluation against those GenAI tools. The outputs and outcomes generated by GenAI's are mostly used in real time without properly getting the results evaluated or validated, which cause it more attractive for attackers and makes it a more important subject of discussion as well as pain point for Security, compliance, risk and governance teams.

Table 1: Types of Attacks in Generative AI.

Attack Type	Description	Example	Impact / Risk
Data poisoning	Malicious samples corrupt training behavior	Backdoor trigger in training set	Model produces manipulated or biased outputs; long-term integrity compromise
Prompt injection	Adversarial instructions override system policies	Hidden prompt in document	Unauthorized actions, data leakage, bypass of safety controls
Model extraction	Queries used to reconstruct a surrogate model	Copying a hosted LLM API	Intellectual property theft; loss of competitive advantage
Membership inference	Determining if a record was in training data	Patient record presence check	Privacy breach; regulatory non-compliance (e.g., HIPAA risk)
Deepfake synthesis	Fabricated audio/video content	Impersonated CEO voice call	Fraud, reputational damage, social engineering attacks

Source: Authors (2026).

V. PRIVACY AND SECURITY CONCERNS IN GENERATIVE AI

Safety issues and data protection concerns are emerging in multiple ways and into deeper level of many layers of Generative AI which makes it difficult to resolve completely via technology security measures alone. Those concerns need to be addressed at all the different parallels such as users, ethical, social, technological, legal and organization perspectives. There are complex and many layered safety challenges across all these mentioned parallels exist.

V.1 USER AND ETHICAL PERSPECTIVES

It quite common to deepfake someone's identity not only from productivity viewpoint but these arises in social environments as well where people impersonate another person's look, voice, style etc. Content makers more frequently use prompt and reference data such as images and videos to synthesize innovative outputs for the purpose of popularity and monetary gains. In addition, curiosity goes up to higher levels and draws users to generate unethical content such as harassment, racism, hatred, restrictive and spreading false perceptions. Content-making AI stirs up controversies as it spits biased information spreading across society and social environment which highlights moral dilemma from the perspective of ethics. [12], [17].

V.2 REGULATORY AND LEGAL PERSPECTIVES

Regulatory authorities and laws are worrying around the content creating AIs due to its flexibility across doing almost anything as it's been instructed or trained to. Some concerns include GDPR compliance issue where usage of Intellectual property is restricted in boundaries, HIPPA compliance issue which states not to use non pseudonymized direct PII (Personal Identifiable information) or PHI (Personal Health information) data for training or stimulatory purposes. There are large number of increases in the number of legal cases being filed around accountability, disclosure, privacy and other ethical controversies. However, there are privacy protection frameworks proposed to mitigate these risks but it in turn compromises its utility and benefits to a fuller extent. There, a balanced and optimal model to accompany both security and utility are highly in need.

V.3 TECHNOLOGICAL AND INSTITUTIONAL PERSPECTIVES

Content generative AIs memorize the models, interactions and conversations from inputs time to time and sometimes it leaks out sensitive information without applying pseudonymization or becomes vulnerable due to its identifiable patterns on even anonymized data.

To address this concern, technological controls are in place such as input validation and adversarial training but those also being challenged by the evolving security threats and attacks. Many organizations often skip building proper oversight systems and lacks consideration of AI security policies, guidelines and controls which lead to improper building of customized productivity AIs. Prompt engineers often create Generative AIs with stable images with clear details, for example, images of a hotel room. Those details are again misused to identify the actual identity of the source (Hotel rooms, name and other details can be exposed via synthesis and reverse engineering).

VI. AI CYBERSECURITY AND REGULATION

AI Cybersecurity Regulations are in place across different contexts and one of which is EU AI act which highlights the first comprehensive framework on AI regulation. The framework includes various acts such as differentiating different levels of risks into Unacceptable risk, High risk, limited and minimal. It also mentions the transparency requirements and encourages evolution of AI innovation and use cases in Europe. The framework primarily focuses on multiple aspects like diversity and complexity of AI technologies with its difficulty assessing those risks and the urge to develop secure-by-design mode of AI systems. The AI regulation framework emphasizes high quality training data, detailed documentation, record management, human oversight, risk management and accuracy. This act has its own tiered structure of fines and penalties, based on violations levels such as prohibited practices, high risk obligations, Improper management of training data, Incorrect information to authorities etc. In a granular level, this act bans acts such as creating or using manipulating techniques, social scoring, biometric exposure, facial database scraping, predictive policing etc. Also, the act promotes cross functional collaboration between security, compliance, risk, privacy, integrity and confidentiality. The framework is planned to be enforced by August 2027 and hopefully it may add some value and key principles to secure AI models and their use.

VII. AI AND SECURITY FROM AN INFORMATION-SECURITY AND RISK-MANAGER STANDPOINT

From a data protection perspective by an information security professional, the AI systems and its associated data and integrations should be treated same as the other critical systems in place. The key considerations include securing Assets & data, Maintaining the risk register for new and emerging risks, implementing necessary controls to comply with regulatory and organizational policies, Evaluation and continuous monitoring of anomalies and deviations. It requires the maturity of shifting from reactive security into proactive security which means that the built-in security and protocols are not only sufficient but also the human, technological, legal and organizational controls should also be in place to avoid AI risk concerns. There is a necessity of considering security concepts in AI in different aspects such as i. Classification of assets, ii. Risk register management, iii. Mapping with appropriate control frameworks, iv. Threat modelling, v. vulnerability management, vi. Logging and monitoring, vii. Identity and access etc., AI adoption can be business driven but should be accompanied by laws and regulations.

VIII. AI'S ROLE IN CYBERSECURITY AND ITS INTERACTION WITH GENERATIVE AI SECURITY

AI plays key roles in cybersecurity which includes cyber defense systems such as phishing filters, endpoint security, anomaly detection, malware analysis, threat hunting and threat modelling. GenAI's plays vital role in reducing human efforts and improvising security and accuracy in blue defense especially situations like automating reported incidents, automating playbook & runbooks in SIEM tools, Deepfake and prompt injection detection to an extent, social engineering simulations, code reviews and so on. AI's feeds detective AIs to defend and secure against forger AI's and false promptings. However, there are associated risks which need to be further improvised such as RAG pulling integrations which may lead to data exfiltration, not updated HR models for hiring use cases where there may be situations that lack adaptive compliance pulling due to lack of real time updates. [12], [17], [13], [6], [11].

IX. DEFENSIVE STRATEGIES AND MITIGATION TECHNIQUES

There are already defense strategies in place against its corresponding threat vector. These strategies are applied to defend AI at different levels of its life cycle from model building, training data to monitoring and evaluation.

Table 2: Mapping Defensive Strategies with Threat Vectors.

Defense Strategy	Primary Threat Mitigated	Implementation Layer	Key Techniques	Residual Risk
Input Sanitization	Prompt Injection, Jailbreaking	Pre-processing / API Gateway	Input validation, allowlists, context filtering, prompt templating	Advanced prompt injections and obfuscation may still bypass controls
RLHF / Red Teaming	Harmful Content Generation	Model Training / Fine-tuning	Human feedback loops, adversarial testing, safety tuning	Model may still fail in unseen or edge-case scenarios
Retrieval Augmented Generation (RAG)	Hallucinations, Poisoning	System Architecture	Grounding with trusted data, vector DB filtering, source attribution	Reliance on external data sources may introduce poisoned or unverified content
Watermarking	Deepfakes, IP Theft	Post-processing / Output	Digital watermarking, content fingerprinting, cryptographic signatures	Watermarks can be removed, altered, or degraded
Differential Privacy	Membership Inference, Privacy Leakage	Training Data Curation	Noise injection, privacy budgets, anonymization techniques	Trade-off between model accuracy and privacy protection

Source: Authors (2026).

IX.1 INPUT FILTERING AND SANITIZATION

This is a Cyber defense technique where it blocks any unwanted inputs to change the application logic and cause corruption to the AI model. It works by blocking specific malicious input patterns by simply rejecting blacklisted inputs which may contain any commands or files. This technique also mitigates buffer overflow but imposing length restriction on sessions connected with AIs. However, these are all static security measures which can be bypassed by more adversarial jail break attempts. This proves the need for additional or secondary AI models to evaluate the inputs of the primary AI model [18].

IX.2 ADVERSARIAL TRAINING AND ROBUSTNESS

Adversarial training and robustness are another cyber defense strategy where the AI model will be trained with details of malicious inputs and threats to create resistance. This instructs the AI to behave correctly, irrespective of the malicious inputs and allows the model to degrade gracefully in case of adversarial attacks. The model was trained, tested and validated for its robustness to avoid hallucinations and unintended behaviors. This layer of defense primarily focuses on mitigating evasion or inference attacks where attackers modify the inputs rationally to fool the model and make it learn from those modifications. It also protects against data poisoning and prompt injections to avoid manipulatory commands and inputs been sent to the model for training. [13], [6], [14], [15], [19].

IX.3 WATERMARKING AND PROVENANCE TRACKING

This strategy is like version control by embedding hidden signal in the source model to verify authenticity. The contents applied in the model are tracked for its integrity by maintaining the metadata history of details like creator, created by, modification history. Without this control deepfakes may remain unchecked and unidentified. This is applied in the post processing or output phase of the implementation layer. There are multiple types of watermarking such as text watermarking, image watermarking, and audio watermarking to ensure fake claims are being avoided. [12], [17].

IX.4 ARCHITECTURAL GUARDRAILS

Architectural Guardrails are system level controls outside the AI model itself which ensure that validate, monitor and isolate the model irrespective of how it behaves. This ensures that principles of limiting authority, enforcement of business rules and prevention of unsafe actions by securing the model based on its behavior. AI models need to be limited to whitelisted API and external tools based on principle of least privilege to avoid lateral or privilege escalation to execute harmful action in the database of backbone system of AI models. Some of the examples are Input guardrails, context guardrails and output guardrails which specifically monitor the AI model by the way it behaves.

X. TOWARD AN ADAPTIVE GOVERNANCE FRAMEWORK FOR SECURITY IN GEN AI

Bringing these strands together, security in Generative AI can be organized around a lifecycle-centric, AI-RMF-aligned governance framework. A typical GenAI system can be described through stages of problem framing and context, data acquisition and curation, model design and training, deployment and integration, operation and monitoring, and retirement and decommissioning. Security and governance measures must be applied consistently across all stages. [16], [8].

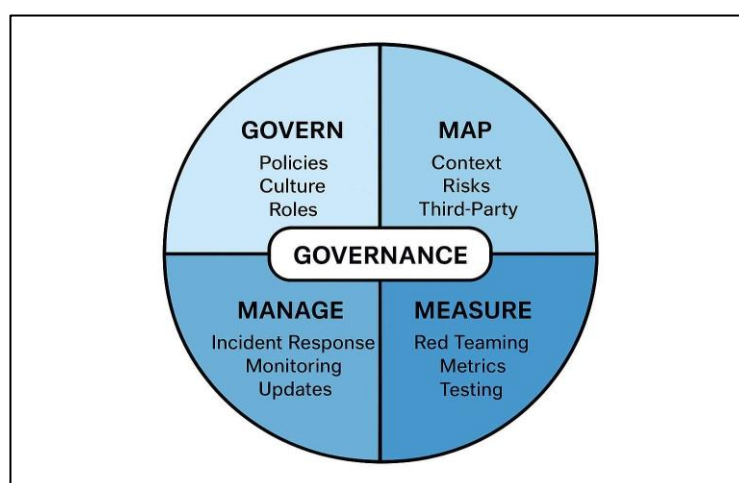


Figure 2: Adaptation of the NIST AI Risk Management Framework (AI-RMF) for the Generative AI Lifecycle.
Source: Authors (2026).

X.1 MAPPING GENERATIVE AI LIFECYCLE TO NIST AI-RMF FUNCTIONS

The NIST AI Risk Management Framework defines four core functions—GOVERN, MAP, MEASURE, and MANAGE—that can be operationalised for GenAI. GOVERN establishes roles, responsibilities, and organisational culture; MAP characterises context, stakeholders, and risk scenarios; MEASURE assesses risks, harm, and control effectiveness; and MANAGE implements and adjusts risk responses over time.

Table 3: Mapping NIST AI-RMF Functions to Generative AI Lifecycle Stages.

Lifecycle Stage	Govern	Measure	Manage
Data acquisition & curation	Data governance policies; data owners and steward	Data quality, bias, and privacy assessments	Access control; sanitization; incident response for data breaches
Model Training & evaluation	Model risk policies; secure ML Ops standards	Robustness (poisoning, backdoors, leakage)	Adversarial training; differential privacy; peer review
Deployment & integration	Acceptable use and integration standards	System level threat modelling (API abuse, tool exploitation)	Rate limiting; content filtering; segmentation; logging
Operation & monitoring	AI risk committee; dashboards and reporting	Incident, drift, and misuse metrics	Policy updates; control tuning; patching; retraining

Source: Authors (2026).

X.2 GOVERNANCE ARTEFACTS AND DECISION GATES

To make this framework operational, organizations can define governance artefacts such as AI system profiles, data-protection impact assessments, model and system cards, adversarial-testing and red teaming reports, incident logs, and decommissioning documentation. Decision gates—go/no-go checkpoints—can require multi-stakeholder approval before moving between lifecycle phases, ensuring that security and governance concerns are addressed proactively rather than retrospectively.

XI. CROSS-CUTTING CHALLENGES AND RESEARCH GAPS

Irrespective of the progress reflected in the reviewed literature, there are several open and dynamic problems remain. There is no widely accepted standard for measuring the safety of AI models and neither adversarial law in place to be followed. Generative AI systems depend heavily on third-party foundation models, external datasets, open-source libraries, customized prompts and cloud infrastructure and the current frameworks provide limited guidance on how to evaluate and manage these dependencies. Regulatory frameworks also face tensions between ambitious expectations and current scientific capabilities, especially for cybersecurity and robustness requirements. Many organizations lack structured AI-governance bodies, formal policies for acceptable use of GenAI, and personnel who understand both AI and security. There is no well-defined framework for AI security which is been widely accepted and due to which there are gaps in designating the research proposal to secure AI models.

XII. CONCLUSION

This literature review has examined the characteristics of AI models, their applications and the systems used. The survey is handled from the perspective of information security, security for AI, and the regulatory constraints. We proposed an RMF centric and lifecycle-oriented security approach to structure security in AI models and its applications. Understanding from the perspective of lack of framework and promising regulations, ever evolving cyber-attacks and misuse of AI applications demands more and broader research in Gen AI security. Addressing these gaps is essential to deploy dynamic and adaptive security for AI systems which makes applications more trustworthy and usable in many real time scenarios.

XIII. AUTHOR'S CONTRIBUTION

Conceptualization: Mohammed Basheer.

Methodology: Mohammed Basheer.

Investigation: Mohammed Basheer.

Discussion of results: Mohammed Basheer and Ebby Darney.

Writing – Original Draft: Mohammed Basheer.

Writing – Review and Editing: Mohammed Basheer and Ebby Darney.

Resources: Ebby Darney.

Supervision: Ebby Darney.

Approval of the final text: Mohammed Basheer and Ebby Darney.

XIV. REFERENCES

- [1] A. Al Siam, M. Alazab, A. Awajan, and N. Faruqi, "A comprehensive review of AI's current impact and future prospects in cybersecurity," IEEE Access, vol. 13, pp. 14029–14050, 2025, doi: 10.1109/ACCESS.2025.3528114.
- [2] M. Trigka and E. Dritsas, "The evolution of generative AI: Trends and applications," IEEE Access, vol. 13, pp. 98504–98529, 2025, doi: 10.1109/ACCESS.2025.3574660.
- [3] M. Choudhury, Z. Elyoseph, N. J. Fast, et al., "The promise and pitfalls of generative AI," Nat. Rev. Psychol., vol. 4, pp. 75–80, 2025, doi: 10.1038/s44159-024-00402-0.
- [4] A. Golda et al., "Privacy and security concerns in generative AI: A comprehensive survey," IEEE Access, vol. 12, pp. 48126–48144, 2024, doi: 10.1109/ACCESS.2024.3381611.

- [5] Y. Liu, J. Huang, Y. Li, et al., "Generative AI model privacy: A survey," *Artif. Intell. Rev.*, vol. 58, Art. no. 33, 2025, doi: 10.1007/s10462-024-11024-6.
- [6] R. Pedro, M. E. Coimbra, D. Castro, P. Carreira, and N. Santos, "Prompt-to-SQL injections in LLM-integrated web applications: Risks and defenses," in *Proc. IEEE/ACM 47th Int. Conf. Softw. Eng. (ICSE)*, Ottawa, ON, Canada, 2025, pp. 1768–1780, doi: 10.1109/ICSE55347.2025.00007.
- [7] P. Shetty, "AI and security, from an information security and risk manager standpoint," *IEEE Access*, vol. 12, pp. 77468–77474, 2024, doi: 10.1109/ACCESS.2024.3408144.
- [8] R. Hamon, H. Junklewitz, J. Soler Garrido, and I. Sanchez, "Three challenges to secure AI systems in the context of AI regulations," *IEEE Access*, vol. 12, pp. 61022–61035, 2024, doi: 10.1109/ACCESS.2024.3391021.
- [9] D. Humphreys, A. Koay, D. Desmond, et al., "AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business," *AI Ethics*, vol. 4, pp. 791–804, 2024, doi: 10.1007/s43681-024-00443-4.
- [10] M. Bethany et al., "Lateral phishing with large language models: A large organization comparative study," *IEEE Access*, vol. 13, pp. 60684–60701, 2025, doi: 10.1109/ACCESS.2025.3555500.
- [11] J. Zhang, P. Wu, J. London, and D. Tenney, "Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: A comprehensive analysis," *IEEE Access*, vol. 13, pp. 28335–28352, 2025, doi: 10.1109/ACCESS.2025.3540075.
- [12] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023, doi: 10.1109/ACCESS.2023.3300381.
- [13] E. Derner, K. Batistič, J. Zahálka, and R. Babuška, "A security risk taxonomy for prompt-based interaction with large language models," *IEEE Access*, vol. 12, pp. 126176–126187, 2024, doi: 10.1109/ACCESS.2024.3450388.
- [14] J. Malik, R. Muthalagu, and P. M. Pawar, "A systematic review of adversarial machine learning attacks, defensive controls, and technologies," *IEEE Access*, vol. 12, pp. 99382–99421, 2024, doi: 10.1109/ACCESS.2024.3423323.
- [15] Y. Zheng, C.-H. Chang, S.-H. Huang, P.-Y. Chen, and S. Picek, "An overview of trustworthy AI: Advances in IP protection, privacy-preserving federated learning, security verification, and GAI safety alignment," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 4, pp. 582–607, Dec. 2024, doi: 10.1109/JETCAS.2024.3477348.
- [16] A. O. Almagrabi and R. A. Khan, "Optimizing secure AI lifecycle model management with innovative generative AI strategies," *IEEE Access*, vol. 13, pp. 12889–12920, 2025, doi: 10.1109/ACCESS.2024.3491373.
- [17] T. C. King, N. Aggarwal, M. Taddeo, et al., "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Sci. Eng. Ethics*, vol. 26, pp. 89–120, 2020, doi: 10.1007/s11948-018-00081-0.
- [18] T. Tsmindashvili et al., "Improving LLM outputs against jailbreak attacks with expert model integration," *IEEE Access*, vol. 13, pp. 134976–134988, 2025, doi: 10.1109/ACCESS.2025.3592458.
- [19] A. Adel and N. Alani, "Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT," *AI & Society*, vol. 40, pp. 6799–6812, 2025, doi: 10.1007/s00146-025-02406-7.