A Formal Framework for Assessing and Mitigating Emergent Security Risks in Generative AI Models: Bridging Theory and Dynamic Risk Mitigation

Aviral Srivastava
College of IST
Pennsylvania State University
State College, PA 16802
aks7873@psu.edu

Sourav Panda
College of IST
Pennsylvania State University
State College, PA 16802
sbp5911@psu.edu

Abstract

As generative AI systems, including large language models (LLMs) and diffusion models, advance rapidly, their growing adoption has led to new and complex security risks often overlooked in traditional AI risk assessment frameworks. This paper introduces a novel formal framework for categorizing and mitigating these emergent security risks by integrating adaptive, real-time monitoring, and dynamic risk mitigation strategies tailored to generative models' unique vulnerabilities. We identify previously under-explored risks, including latent space exploitation, multi-modal cross-attack vectors, and feedback-loop-induced model degradation. Our framework employs a layered approach, incorporating anomaly detection, continuous red-teaming, and real-time adversarial simulation to mitigate these risks. We focus on formal verification methods to ensure model robustness and scalability in the face of evolving threats. Though theoretical, this work sets the stage for future empirical validation by establishing a detailed methodology and metrics for evaluating the performance of risk mitigation strategies in generative AI systems. This framework addresses existing gaps in AI safety, offering a comprehensive road map for future research and implementation.

1 Introduction

Generative AI models, including large language models (LLMs) and multimodal systems like diffusion models, have dramatically transformed a variety of industries. These models enable advanced capabilities in content creation, natural language understanding, and image generation, yet their complex architectures present substantial and often unforeseen security vulnerabilities. The rapid deployment of generative AI systems has far outpaced the development of comprehensive security frameworks, leaving critical gaps in our ability to identify and mitigate risks. Traditional AI safety approaches, while effective in identifying issues like adversarial attacks and data leakage, are insufficient to manage the dynamic, emergent threats specific to generative AI.

Current risk assessment models fall short in addressing the evolving nature of these systems, particularly their susceptibility to latent space exploitation, adversarial prompt manipulation, and feedback-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

loop-induced degradation. Moreover, the integration of multimodal capabilities introduces novel attack surfaces, allowing adversaries to exploit vulnerabilities across text, image, and video simultaneously. These emergent risks pose significant threats to both the functionality of generative AI systems and the broader ecosystem in which they operate, particularly in fields like healthcare, finance, and security. This paper seeks to address these gaps by proposing a formal framework for assessing and mitigating emergent security risks in generative AI models. To guide the development and evaluation of this framework, we pose the following research questions (RQs):

- **RQ1:** How can emergent security risks in generative AI models, such as latent space exploitation and multimodal cross-attack vectors, be systematically identified and categorized across the model lifecycle?
- **RQ2:** What adaptive, real-time monitoring techniques can be integrated into generative models to provide continuous risk assessment and dynamic mitigation in response to evolving threats?
- **RQ3:** How can formal verification methods be utilized to ensure the robustness and scalability of generative AI models against emergent threats while maintaining system efficiency?

This paper presents a layered risk identification approach that spans the entire lifecycle of generative models, from data ingestion to model inference, incorporating continuous adversarial testing, anomaly detection, and dynamic watermarking techniques. While this framework is theoretical, it offers a robust foundation for future empirical validation and sets forth a roadmap for the development of scalable, adaptive security measures in generative AI.

By addressing both the theoretical underpinnings and practical challenges of securing generative AI models, this paper aims to fill critical gaps in AI safety research. Our framework not only enhances existing methods but also introduces novel solutions designed to evolve alongside the generative models they protect, ensuring their secure deployment in increasingly complex environments.

2 Literature Review

Security Risks in Generative AI Generative AI systems, such as large language models (LLMs) and diffusion models, present unique security challenges due to their complex architectures and emergent capabilities. Barrett et al. discuss the dual-use dilemma of generative AI, highlighting both its capabilities and the security risks it poses. They synthesize findings from a workshop held at Google, co-organized by Stanford University and the University of Wisconsin-Madison, emphasizing the need for robust security measures [2]. Similarly, Humphreys et al. examine the ethical obligations companies have when implementing generative AI, focusing on the potential cyber security risks and the moral responsibilities involved [8]. Feffer et al. explore the concept of red-teaming in generative AI, analyzing its effectiveness and limitations in identifying and mitigating security risks [7].

Risk Assessment in AI Risk assessment is a critical component of AI governance. Novelli et al. propose a methodology for assessing AI risk magnitudes, focusing on the construction of real-world risk scenarios and applying a proportionality test to balance competing values [10]. Schmitz et al. provide a systematic overview of risk aggregation schemes used in existing AI risk assessment frameworks, focusing on how trade-offs among risk dimensions are incorporated [13].

Risk Mitigation in AI Effective risk mitigation strategies are essential for securing AI systems. Kaminski describes the growing convergence around risk regulation in AI governance, offering an analytic framework for understanding the use of risk regulation in AI governance [9]. Parimala provides a comprehensive overview of recent advances in anomaly detection techniques, highlighting the role of AI and machine learning in this field [11].

AI Safety AI safety is a crucial aspect of AI development and deployment. Shah and Mishra review the advancements in AI-driven occupational health and safety technologies, offering predictive insights and risk mitigation strategies [14]. Deepak discusses the global enthusiasm around AI safety, highlighting the challenges and potential issues associated with current AI safety measures [5].

Multi-Modal Evaluation in AI Multi-modal evaluation is essential for assessing the performance of AI systems across different modalities. Chen et al. introduce a comprehensive benchmark for

evaluating large vision-language models in medical applications, addressing the limitations of current benchmarks [3].

Dynamic Watermarking in AI Dynamic watermarking is a promising technique for protecting the integrity of AI-generated content. Zhong et al. present a comprehensive overview of deep learning-based image watermarking techniques, focusing on robustness and adaptability [16].

Real-Time Anomaly Detection in AI Real-time anomaly detection is critical for identifying and mitigating security threats in AI systems. Dini and Saponara explore the design and assessment of real-time anomaly detection techniques for automotive cybersecurity, leveraging neural networks and fingerprinting techniques [6].

While the existing body of work provides valuable insights into the security risks, risk assessment methods, and mitigation strategies for AI systems, they fall short in addressing the dynamic and evolving nature of generative AI models. Current methods often lack the ability to continuously adapt to new, unforeseen vulnerabilities that arise post-deployment. Our proposed framework addresses these gaps by integrating continuous risk monitoring, real-time adversarial simulations, and adaptive mitigation techniques that evolve alongside the AI models. This dynamic approach ensures that emergent threats are detected and mitigated in real-time, providing a more robust and scalable solution compared to static, pre-deployment assessments.

3 Novel Security Risks in GenAI

3.1 Introduction of New Vulnerabilities:

While current research has identified some key vulnerabilities in generative AI (GenAI) systems, such as hallucinations, data leakage, and adversarial prompt manipulation, emerging threats have not been fully explored. In this section, we analyze a set of novel security risks, focusing on unsupervised learning dynamics, self-improving models, latent space exploitation, and cross-modal attacks in multimodal systems. These vulnerabilities represent critical areas for future research, particularly in terms of adaptive, real-time risk monitoring.

3.2 Unsupervised Learning Vulnerabilities

Risk Overview: Generative AI models, particularly those trained on unsupervised or self-supervised datasets, rely heavily on vast, uncurated data sources scraped from public repositories. Without the explicit labeling present in supervised learning, these systems are prone to absorbing noisy, biased, or harmful data. As models continue to evolve based on this data, they may reinforce incorrect associations or harmful biases, often without detection. This risk is exacerbated when feedback loops emerge between model outputs and training data.

Potential Impact:

- **Bias Amplification:** Generative models trained on unsupervised datasets may inadvertently learn and perpetuate societal biases, including racial, gender, or cultural biases. The absence of real-time oversight makes these biases ingrained, resulting in systemic output biases.
- **Feedback Loops:** When generative models interact with user-generated content on platforms like social media, a feedback loop can develop. In this cycle, the biased or harmful outputs generated by the model are reintroduced as new data, further reinforcing incorrect or problematic patterns. Over time, this can degrade the model's accuracy and increase its vulnerability to adversarial manipulation.

Example Case: A GenAI model trained on uncurated social media data may generate biased outputs when asked to summarize trending topics. If the model then scrapes its own generated content as part of future training data, it perpetuates and amplifies the biases present in the original input [2, 4]

3.3 Self-Improvement Risks

Risk Overview: One of the most exciting but risky developments in GenAI is the emergence of self-improving models. These systems are designed to fine-tune themselves in real time based on feedback from their interactions. While this self-adaptive capability offers efficiency gains, it also introduces the risk of unchecked bias reinforcement and the propagation of security vulnerabilities. Without human oversight, these models may learn from harmful inputs, deepening their vulnerabilities over time.

Potential Impact:

- **Bias Reinforcement:** Self-improving models that adjust based on real-world data without external oversight may unintentionally reinforce harmful biases. Over time, small vulnerabilities could become more severe as the model's weights and parameters shift to accommodate flawed or adversarial inputs
- **Exacerbating Security Issues:** Without structured validation, self-improving models could become increasingly susceptible to adversarial prompts. Malicious users might feed the model adversarial or biased inputs, which the model could learn from, thereby worsening its vulnerabilities in subsequent interactions.

Example Case: A self-improving chatbot that interacts with malicious users may start learning from harmful, biased, or adversarial inputs, causing it to produce increasingly flawed or harmful outputs. As the chatbot fine-tunes itself based on these interactions, it becomes more vulnerable to adversarial attacks, with its security issues compounding over time [4, 15].

3.4 Hidden Information Leakage: Latent Space Vulnerabilities

Risk Overview: One of the less-discussed but critical risks in generative models is latent space exploitation. The latent space represents the internal high-dimensional feature representations learned by the model. While latent spaces are crucial for model functionality, they can inadvertently retain sensitive patterns from the training data. Attackers with knowledge of the model's architecture can exploit these latent spaces to extract sensitive information, leading to significant privacy concerns in sectors like healthcare and finance.

Potential Impact:

- **Latent Space Exploitation:** Attackers with sophisticated knowledge of a model's architecture could manipulate inputs to probe the latent space and retrieve sensitive information embedded in these internal representations. For example, in healthcare models, the latent space might reveal sensitive patient data even when the output is anonymized.
- **Inferred Privacy Leaks:** Through adversarial probing, an attacker could generate outputs that reveal latent correlations, potentially identifying individuals from supposedly anonymized data or uncovering proprietary business strategies from models trained on confidential data.

Example Case: In financial systems, attackers could probe the latent space of a generative model used for stock prediction to uncover sensitive market trends or proprietary investment strategies, leading to significant financial or privacy breaches [2, 4]

3.5 Simultaneous Multimodal Exploitation

Risk Overview: Multimodal generative models that process and generate text, images, audio, and video simultaneously present new attack surfaces. These models integrate multiple types of data, making them vulnerable to cross-modal exploits. An attack targeting one modality (e.g., text) could expose weaknesses in another modality (e.g., images or videos), enabling adversaries to exploit the model across different dimensions simultaneously.

Potential Impact:

- **Cross-Modal Consistency Exploits:** Attackers can inject adversarial inputs into one modality (e.g., an image) that lead to harmful outputs in another modality (e.g., misleading or biased text). For instance, an adversarial image might be paired with a deceptive or biased textual description, which reinforces false or harmful narratives.
- **Coordinated Multimodal Attacks:** Attackers could simultaneously manipulate multiple modalities to generate realistic yet harmful content, such as deepfakes. This is particularly concerning in scenarios like news generation or multimedia content creation, where both visual and audio elements can be manipulated to spread misinformation or discredit individuals.

Example Case: In a news generation scenario, an adversary could generate fake articles supported by AI-generated deepfake images or videos. These multimodal attacks make it more challenging for users to distinguish between legitimate and fake content, amplifying the potential for misinformation [2, 4, 15].

4 Parameters for Risk Categorization

Based on the insights from NIST's Risk Management Framework (RMF) and the ISO 27001 cybersecurity standards [12, 1], we can classify risks using the following parameters

Potential Impact

- Low: Minimal consequences if exploited, affecting a limited number of users or non-critical processes. For instance, generating biased but non-malicious text in an isolated scenario.
- Medium: Significant impact on some operations or users, such as privacy breaches affecting identifiable groups.
- High: Severe consequences, such as critical system failures or widespread misinformation, impacting infrastructure, security, or public trust.

Exploitability

- Low: Requires a sophisticated attack vector or rare conditions, making exploitation unlikely.
- Medium: Moderate effort or specific conditions required for an exploit, such as known vulnerabilities in handling adversarial inputs
- High: Easily exploitable due to well-known vulnerabilities or lack of adequate safeguards, such as unsupervised learning models prone to bias without regular updates.

Scope of Effect

- Low: Localized effects with minimal spread, possibly confined to a single user or output.
- Medium: Broader reach, affecting multiple systems, users, or outputs.
- High: Extensive impact, potentially leading to cascading failures across multiple systems, modalities, or platforms.

By combining these parameters, we can systematically assess the severity of each security risk in generative AI models. Refer to the Appendix for a table to categorize the key risks outlined in Section 3 of the paper.

5 Novel Risk Mitigation Strategies

We propose a series of novel risk mitigation strategies for generative AI models, focusing on adaptive mechanisms that evolve in response to new attack patterns. These strategies include real-time filtering, dynamic watermarking, multi-layered mitigation, and human-AI collaboration.

Contextual Real-Time Filtering We introduce a real-time filtering system that evolves based on new attack patterns. For example, a filtering mechanism could detect and block problematic prompts (e.g., adversarial input) before they are processed, learning from each failed attack.

Dynamic Watermarking We propose the development of a self-evolving watermarking system, where watermark techniques adapt as adversaries develop new removal methods. The watermarking algorithm can evolve based on detection failures [2].

Multi-Layered Mitigation We propose a multi-tier system for mitigating risks at various stages of the model lifecycle:

- **Tier 1:** Pre-deployment filtering, applied during model training to ensure that adversarial inputs are detected and mitigated before deployment.
- **Tier 2:** Inference-time defense, focusing on real-time output filtering to detect and block potentially harmful outputs.
- **Tier 3:** Post-inference validation, where high-risk interactions are validated after generation to reduce false positives and mitigate any remaining risks.

Human-AI Collaboration In decision-critical environments, we suggest mechanisms for human oversight. Feedback loops can be incorporated where human analysts intervene when GenAI outputs are flagged as potentially harmful. This collaboration between human oversight and AI-driven analysis will ensure that higher-risk outputs receive necessary validation and correction when required.

Refer to Appendix Figure 1 for a visualization of the Working Structure of the Formal Framework for Risk Assessment in Generative AI

6 Formal Framework for Risk Assessment

In this section, we present a comprehensive formal framework for systematically assessing and mitigating security risks in generative AI models. This framework leverages formal verification, continuous monitoring, and adaptive risk mitigation to address both known and emergent threats. By adopting a layered approach, we ensure that risks are identified and mitigated at critical stages of the GenAI lifecycle, from data ingestion to model inference.

6.1 Theoretical Grounding

The framework is grounded in formal verification methods, which have been widely used in software engineering and AI safety to mathematically guarantee that a system behaves according to its specifications [12, 1]. Formal verification is particularly suitable for assessing generative AI models because of their complexity and the emergent behaviors they exhibit. These methods allow us to model expected behavior and prove that deviations—such as biases, adversarial attacks, or data leakage—will be detected and mitigated before deployment.

6.2 Risk Identification Layers

Our framework operates through a layered approach, designed to identify and mitigate risks at three key stages of the generative AI lifecycle: the data layer, model training layer, and model inference layer. Each stage involves distinct methodologies for risk identification, leveraging formal verification, adversarial testing, and real-time anomaly detection.

6.2.1 Data Layer

At the data layer, risks primarily involve biases, noisy data, and privacy violations that could influence model behavior later in the lifecycle. This layer applies formal verification techniques to ensure that the data used for training is both consistent and free from harmful biases or inaccuracies.

Formal Verification Algorithm: We deploy a bias filtering algorithm that uses mathematical models to verify that the training data does not reinforce societal biases, such as gender or racial biases. This approach detects anomalies within the dataset, ensuring that harmful patterns are flagged before they can affect model outputs. This algorithm functions by analyzing patterns in the data that deviate from established baselines of fairness and accuracy [1].

Dynamic Data Monitoring: In addition to formal verification, continuous monitoring is deployed to detect new patterns of bias that may emerge in real time. A real-time anomaly detection system

monitors the data for unexpected shifts in distribution or content. If a bias or noisy data is detected, the data can be flagged, and the model training process is halted for further review.

6.2.2 Model Training Layer

During model training, adversarial manipulation becomes a significant threat, where attackers may attempt to perturb the training data to manipulate the model's learning. Our framework integrates automated adversarial testing directly into the training process, enabling the system to continuously probe for weaknesses before the model is deployed.

Adversarial Testing Module: This module employs real-time adversarial simulation that introduces adversarial inputs at various points during the training process to test the model's robustness. The simulation leverages adversarial perturbations—slight modifications to the input data that are designed to mislead the model. These perturbations mimic real-world attacks, allowing the framework to detect vulnerabilities before deployment. The module monitors how effectively the model resists adversarial inputs and adjusts accordingly.

Metrics for Robustness: Metrics used to evaluate the effectiveness of adversarial testing include the model's accuracy under adversarial conditions, the degree of perturbation required to mislead the model, and the model's ability to maintain consistency in outputs across multiple adversarial inputs. These metrics help define what constitutes a "high-risk" versus "low-risk" vulnerability. For example:

- **High-risk:** If a small perturbation results in significantly misleading outputs, indicating that the model is highly susceptible to adversarial attacks.
- **Low-risk:** If the model remains robust even under large adversarial perturbations, demonstrating strong resistance to manipulation.

6.2.3 Model Inference Layer

At the inference stage, risks are primarily associated with real-time attacks and unforeseen vulnerabilities that manifest when the model interacts with external environments. This is where dynamic risk monitoring becomes critical.

Real-time Adversarial Simulation and Response: During inference, each input-output pair is subjected to real-time adversarial simulation. This system continuously analyzes the model's interactions with external inputs, looking for any signs of adversarial activity. For instance, when the model encounters a prompt, the adversarial simulation generates variations of the input designed to probe for weaknesses (e.g., modifying a prompt to generate biased or harmful outputs). These adversarial tests run concurrently with the model's normal operation, ensuring that vulnerabilities can be detected and responded to in real-time [12].

Latent Space Monitoring: Anomaly detection in the latent space is employed to track patterns that emerge deep within the model's architecture. By establishing a statistical baseline of what constitutes "normal" latent space activity, the framework can detect deviations that suggest potential risks, such as information leakage or bias amplification. The system continuously monitors latent space activity for anomalies, which could indicate an ongoing attack or emergent bias that was not detected during training [1].

Dynamic Watermarking: To further secure the inference layer, the framework implements dynamic watermarking, which involves embedding invisible markers in the outputs generated by the model. These watermarks evolve to resist adversarial removal techniques. If the watermark is compromised or missing, it indicates potential tampering or misuse of the model outputs.

6.3 Dynamic Risk Monitoring

To address the continuously evolving threat landscape, our framework incorporates dynamic risk monitoring mechanisms that operate at all stages of the model lifecycle. These include:

- **Real-time Adversarial Simulations:** These simulations are continuously applied during both training and inference to detect potential vulnerabilities early and predict the likelihood of exploitation. The feedback loop generated by adversarial testing informs the model's security posture, dynamically adjusting its architecture as new threats are identified.

- **Anomaly Detection Algorithms:** Anomalies in the model's performance or latent space activity are flagged in real-time, triggering preemptive countermeasures. These algorithms are particularly focused on detecting latent space leaks—high-dimensional spaces where information from the training data can leak into the model's predictions.
- **Continuous Model Refinement:** When vulnerabilities are detected, the model is fine-tuned in real time. This process leverages reinforcement learning techniques, where the model autonomously learns from adversarial inputs and updates its parameters to resist similar attacks in the future. This continuous adjustment ensures that the model remains secure even as new threats emerge [1].

6.4 Metrics for Evaluating the Framework

The success of the proposed framework is evaluated through a series of well-defined metrics:

- **Attack Detection Rate:** The percentage of adversarial attacks successfully detected during training and inference.
- **Response Time to Threats:** The time it takes for the system to detect and respond to an identified vulnerability.
- **Model Robustness:** The model's accuracy and consistency when subjected to adversarial inputs or anomalies.
- **False Positive Rate:** The frequency of incorrectly flagged inputs or behaviors that do not represent actual security risks.
- **Latency Impact:** The effect of continuous monitoring and adversarial simulation on the model's real-time performance, ensuring that security measures do not compromise system efficiency.

These metrics are essential for distinguishing between "high-risk" and "low-risk" scenarios, enabling the model to prioritize mitigation efforts where they are most needed.

Refer to Appendix Figure 2 for a visualization of the Formal Framework for Risk Assessment in Generative AI.

By integrating formal verification, continuous adversarial simulation, and real-time anomaly detection, this framework provides a comprehensive, adaptive approach to securing generative AI models. The framework's layered architecture ensures that vulnerabilities are identified and mitigated throughout the entire model lifecycle, from deployment training. Though theoretical, this framework lays the foundation for empirical validation and future development in securing generative AI systems against evolving threats.

7 Limitations

This paper presents a theoretical framework for assessing and mitigating emergent security risks in generative AI models. While the framework is rooted in existing risk management methodologies such as those from NIST, ISO, and MITRE, there are several inherent limitations:

Absence of Empirical Validation As this work is purely theoretical, none of the proposed mitigation strategies have been tested in real-world environments. The effectiveness of dynamic risk monitoring, adversarial simulations, and anomaly detection systems must be empirically validated before they can be widely adopted. The lack of empirical testing means the practical feasibility of the framework remains unproven, especially in diverse AI use cases such as healthcare, finance, and large-scale media generation.

Dynamic Threat Landscape The constantly evolving nature of generative AI means that new, unforeseen vulnerabilities could emerge that are beyond the scope of this framework. Models like GPT and other multimodal systems are rapidly changing, and novel forms of attacks or biases may arise that our current framework may not fully address. As these systems scale, managing real-time risk also becomes more complex, and our framework may need to be adapted over time to incorporate future security trends.

Resource and Scalability Constraints The continuous risk monitoring and real-time adversarial simulation components of this framework could be resource-intensive. Ensuring that these defenses operate effectively without slowing down model performance or requiring significant computational power poses a real challenge. Small- to medium-sized organizations may find it difficult to implement such resource-heavy solutions.

Latent Space Monitoring Monitoring the latent space in real time presents significant technical challenges. While we propose anomaly detection for latent space monitoring, the ability to interpret the high-dimensional features in latent spaces and identify security risks remains an under-explored area. Future research will need to refine this approach and improve the interpretability of latent space representations.

8 Future Directions

Despite these limitations, the theoretical framework presented in this paper provides a solid foundation for further research and development in generative AI security. Key avenues for future work include:

Empirical Testing and Validation The next logical step is to empirically validate this framework by applying it to real-world generative models across different domains, such as NLP, image generation, and multi-modal applications. This includes testing the effectiveness of adversarial simulations, dynamic watermarking, and latent space anomaly detection in identifying and mitigating risks. Future research could focus on benchmarking the proposed methods against existing mitigation strategies to determine their practical viability.

Adapting to Evolving AI Models As generative models continue to evolve, particularly with advances in self-improving AI and reinforcement learning, the framework will need to be continuously refined. Research should focus on adapting the framework to handle more sophisticated AI architectures and emergent behaviors. For example, future work could explore how the framework might extend to self-supervised learning or more advanced multi-modal AI systems.

Resource Optimization While the current framework is resource-intensive, future work should explore ways to optimize real-time monitoring without compromising performance. Research could focus on developing more efficient adversarial testing methods or scalable anomaly detection algorithms to minimize computational overhead while maintaining robust security. This will be crucial for making the framework accessible to a wider range of organizations with varying computational resources.

Ethical and Societal Implications Further exploration is needed regarding the ethical implications of deploying dynamic risk assessment frameworks. Future research should assess how these systems align with data privacy laws, intellectual property rights, and societal norms, particularly when used in high-stakes domains like healthcare, education, and finance. Incorporating human oversight and ensuring compliance with ethical AI standards will be critical as these models become more integrated into decision-making processes.

9 Conclusion

As generative AI models become increasingly integrated into various industries and applications, the need for robust security frameworks is paramount. This paper presents a theoretical framework for assessing and mitigating emergent security risks in generative AI, focusing on both existing vulnerabilities and novel threats such as unsupervised learning dynamics, self-improving models, and latent space exploitation. By leveraging formal verification techniques, dynamic risk monitoring, and continuous real-time adversarial simulations, we aim to provide a comprehensive solution for addressing both known and unforeseen security threats in generative models. We introduced a layered risk assessment approach that addresses risks at the data, training, and inference stages, emphasizing the importance of continuous monitoring and adaptation as models evolve. While the framework proposed is theoretical, it provides a critical foundation for future research, with clear pathways for empirical validation, scalability improvements, and integration of ethical and societal considerations.

In addressing the limitations of the proposed framework, we recognize that empirical validation is essential, and future work should focus on testing the effectiveness of the mitigation strategies across real-world generative AI applications. Additionally, the dynamic and rapidly evolving nature of AI will require ongoing refinement of this framework to ensure that it remains relevant and effective against new and emerging threats. Ultimately, this work contributes to the broader field of AI safety by offering a scalable and adaptable framework designed to ensure the safe deployment of generative AI models in increasingly complex and high-stakes environments.

References

[1] Nist cyber risk scoring (crs) - program overview. https://csrc.nist.gov/CSRC/media/Presentations/nist-cyber-risk-scoring-crs-program-overview/images-media/NIST%20Cyber%20Risk%20Scoring%20(CRS)%20-%20Program%20Overview.pdf, 2024. Accessed: 2024-09-20.

[2] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.

[3] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024.

[4] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*, 2024.

[5] P Deepak. Ai safety: necessary, but insufficient and possibly problematic. *arXiv preprint arXiv:2403.17419*, 2024.

[6] Pierpaolo Dini and Sergio Saponara. Design and experimental assessment of real-time anomaly detection techniques for automotive cybersecurity. *Sensors*, 23(22):9231, 2023.

[7] Michael Feffer, Anusha Sinha, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? *arXiv preprint arXiv:2401.15897*, 2024.

[8] Declan Humphreys, Abigail Koay, Dennis Desmond, and Erica Mealy. Ai hype as a cyber security risk: the moral responsibility of implementing generative ai in business. *AI and Ethics*, pages 1–14, 2024.

[9] Margot E. Kaminski. Regulating the risks of ai. *Boston University Law Review*, 103:1347, 2023.

[10] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. Ai risk assessment: A scenario-based, proportional methodology for the ai act. *Digital Society*, 3(1):13, 2024.

[11] Venkata Krishna Parimala. Introductory chapter: Anomaly detection–recent advances, ai and ml perspectives and applications. *Anomaly Detection-Recent Advances, AI and ML Perspectives and Applications*, 2024.

[12] Privacy Engine. Iso 27001 vs. nist cybersecurity framework comparison, 2023. URL https://www.privacyengine.io/blog/iso-27001-vs-nist-cybersecurity-framework/.

[13] Anna Schmitz, Michael Mock, Rebekka Görge, Armin B Cremers, and Maximilian Poretschkin. A global scale comparison of risk aggregation in ai assessment frameworks. *AI and Ethics*, pages 1–26, 2024.

[14] Immad A Shah and SukhDev Mishra. Artificial intelligence in advancing occupational health and safety: an encapsulation of developments. *Journal of Occupational Health*, 66(1):uiad017, 2024.

[15] Alexander Sisto and K.C. Halm. Latest nist guidance identifies generative ai risks and corresponding mitigation strategies, 2024. URL https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2024/08/new-nist-guidance-on-generative-ai-risks.

[16] Xin Zhong, Arjon Das, Fahad Alrasheedi, and Abdullah Tanvir. A brief, in-depth survey of deep learning-based image watermarking. *Applied Sciences*, 13(21):11852, 2023.

Appendix

This appendix provides a set of algorithms and pseudocode to illustrate the key components of the framework proposed in this paper for assessing and mitigating emergent security risks in generative AI models. The included algorithms cover various aspects such as dynamic risk monitoring, adversarial simulations, anomaly detection in latent spaces, and multi-layered risk mitigation strategies.

The primary purpose of these algorithms is to offer conceptual clarity and a deeper understanding of the mechanisms underpinning our framework. While these pseudocodes represent the core logic and processes, they are intended to serve as illustrative examples and are not fully optimized for direct implementation in production environments. Further empirical testing, validation, and refinement are required before they can be applied in real-world applications.

Below, we present the key algorithms that reflect the dynamic, real-time nature of our proposed approach to securing generative AI systems.

Algorithm 1 Dynamic Risk Monitoring for Generative AI Models

```
Input: Model M, Real-time input data D_real, Monitoring frequency f
Output: Detected threats and triggered countermeasures
Initialize threat detection system T_detect
Initialize countermeasure system C_response
while Model M is active do
    Retrieve real-time input x ∈ D_real
    Evaluate x using T_detect for potential adversarial behavior or anomalies
    if T_detect flags an input as suspicious then
        Trigger appropriate countermeasure from C_response
        Log detected threat and corresponding countermeasure
    else
        Continue normal model operation
    end if
    Wait for next monitoring interval 1/f
end while
```

Algorithm 2 Adversarial Simulation for Model Robustness Testing

```
Input: Training data D_train, Model M, Adversarial perturbation function P_adv
Output: Adjusted model M_robust
Initialize adversarial example generator G_adv
for each training sample x ∈ D_train do
    Generate adversarial sample x_adv = P_adv(x) using G_adv
    Train model M on x_adv to update parameters
    Evaluate the robustness of M to adversarial examples
end for
return updated model M_robust
```

Algorithm 3 Latent Space Anomaly Detection for Generative AI

```
Input: Latent space representations L, Baseline statistics S_baseline, Anomaly threshold τ
Output: Detected latent space anomalies
Initialize anomaly detection model A_detect
Compute statistical baseline S_baseline for latent space L
for each latent representation l ∈ L do
    Compare l to S_baseline
    if deviation from S_baseline > τ then
        Flag l as an anomaly
        Log and alert potential anomaly in latent space
    end if
end for
```

Algorithm 4 Dynamic Watermarking for Generative AI Outputs

Input: Model output O, Watermark embedding function W_{embed}, Adversarial detection system A_{detect}
Output: Watermarked output $O_{watermarked}$
Initialize watermark embedding process
$O_{watermarked} = W_{embed}(O)$ ▷ Embed watermark into output
while Model is deployed **do**
    Monitor $O_{watermarked}$ for tampering or adversarial removal
    if A_{detect} flags an adversarial removal attempt **then**
        Adjust watermark W_{embed} to prevent future removal
        Re-embed watermark and update $O_{watermarked}$
    end if
end while
return $O_{watermarked}$

Algorithm 5 Multi-Layered Risk Mitigation Strategy for Generative AI

Input: Model M, Training data D_{train}, Real-time inputs D_{real}
Output: Mitigated security risks across all layers
Tier 1: Pre-deployment Filtering
for each sample $x \in D_{train}$ **do**
    Apply pre-deployment filtering to detect adversarial inputs
    if adversarial input detected **then**
        Remove or correct input x
    end if
end for
Tier 2: Inference-Time Defense
while Model M is deployed **do**
    Monitor real-time input $x \in D_{real}$
    Apply real-time defense mechanisms to filter harmful outputs
end while
Tier 3: Post-Inference Validation
for each high-risk interaction O **do**
    Apply post-inference validation to check for tampering or risk
    if risk detected **then**
        Flag and mitigate the risk, and revalidate O
    end if
end for

Algorithm 6 Latent Space Monitoring and Continuous Feedback Loop

Input: Latent space representations L, Baseline statistics $S_{baseline}$, Model M
Output: Updated model M with fine-tuning
Compute statistical baseline $S_{baseline}$ for latent space L
while Model M is active **do**
    for each latent representation $l \in L$ **do**
        Compare l to $S_{baseline}$
        if deviation from $S_{baseline}$ exceeds threshold τ **then**
            Flag l as anomalous
            Adjust model parameters to mitigate anomaly
        end if
    end for
    Fine-tune M based on flagged anomalies
end while
return updated model M

Algorithm 7 Real-Time Filtering of Adversarial and Harmful Prompts

Input: Real-time prompt P, Adversarial detection system A_{detect}, Threshold τ
Output: Filtered prompt or blocked request
Initialize filtering system F_{system}
while Model M is deployed **do**
    Receive real-time prompt P
    Evaluate P using A_{detect} to assess risk level
    if Risk score $R(P) > \tau$ **then**
        Block prompt and log the event
        Provide feedback to user or system
    else
        Allow P to pass through to the model
    end if
end while

Algorithm 8 Reinforcement Learning-Based Model Fine-Tuning

Input: Model M, Adversarial input set A, Feedback loop F
Output: Updated model $M_{updated}$
Initialize reinforcement learning agent RL_{agent}
while Model M is deployed **do**
    for each adversarial input $x_{adv} \in A$ **do**
        Process x_{adv} through model M
        Receive feedback F from RL_{agent} based on model's response
        Update model parameters based on F to resist similar attacks
    end for
    Continuously fine-tune M with each adversarial input
end while
return updated model $M_{updated}$

Algorithm 9 Cross-Modal Consistency Check for Text-Image-Video Alignment

Input: Text T, Image I, Video V
Output: Verified consistency or flagged inconsistency
Initialize consistency checking function C_{check}
while Model is deployed for multimodal tasks **do**
    Receive text T, image I, and video V
    Compare T and I using $C_{check}(T, I)$
    Compare I and V using $C_{check}(I, V)$
    Compare T and V using $C_{check}(T, V)$
    if Inconsistency found between any pair (e.g., $C_{check}(T, I)$ fails) **then**
        Flag the inconsistency and log it for further analysis
    else
        Confirm consistency and proceed with the output generation
    end if
end while

Algorithm 10 Risk Classification Based on Thresholds for High, Medium, and Low Risk

Input: Risk score R, Thresholds τ_{high}, τ_{medium}
Output: Classified risk level (High, Medium, Low)
if $R > \tau_{high}$ **then**
    Classify as "High Risk"
else if $\tau_{medium} < R \leq \tau_{high}$ **then**
    Classify as "Medium Risk"
else
    Classify as "Low Risk"
end if
Log the risk classification for further analysis

Risk Name	Severity	Potential Impact	Exploitability	Scope of Effect
Hallucinations	Medium	Misleading or incorrect outputs affecting non-critical tasks	Moderate	Localized, affecting isolated outputs
Data Leakage	High	Privacy violations, especially with sensitive data like PII	High	Broad, particularly in models trained on uncurated data sources
Adversarial Prompt Manipulation	High	Harmful or biased content generated through manipulated prompts	Moderate to High	Extensive, affecting multiple outputs across systems
Bias Reinforcement	Medium to High	Reinforces societal biases in outputs (e.g., race, gender)	High	Broad, impacts a large range of outputs in various domains
Unsupervised Learning Vulnerabilities	Medium	Bias amplification and feedback loops	Moderate to High	Broad, affecting large datasets
Self-Improvement Risks	High	Reinforcement of harmful behaviors	High	Wide, with potential cascading effects on model behavior
Hidden Information Leakage (Latent Space)	High	Privacy breaches in sensitive sectors (e.g., healthcare, finance)	High	Broad, with the potential for significant privacy violations
Simultaneous Multimodal Exploitation	High	Cross-modal misinformation or deepfakes	Moderate to High	Extensive, across multiple content formats
Training Data Poisoning	High	Compromised outputs, leading to unreliable or harmful results	Moderate	Broad, affecting model integrity and user trust
Model Confabulation	Medium	Creation of plausible but incorrect outputs, potentially misleading users	Low to Moderate	Localized but recurring, particularly in text-based models

Table 1: Risk Categorization for Generative AI Models

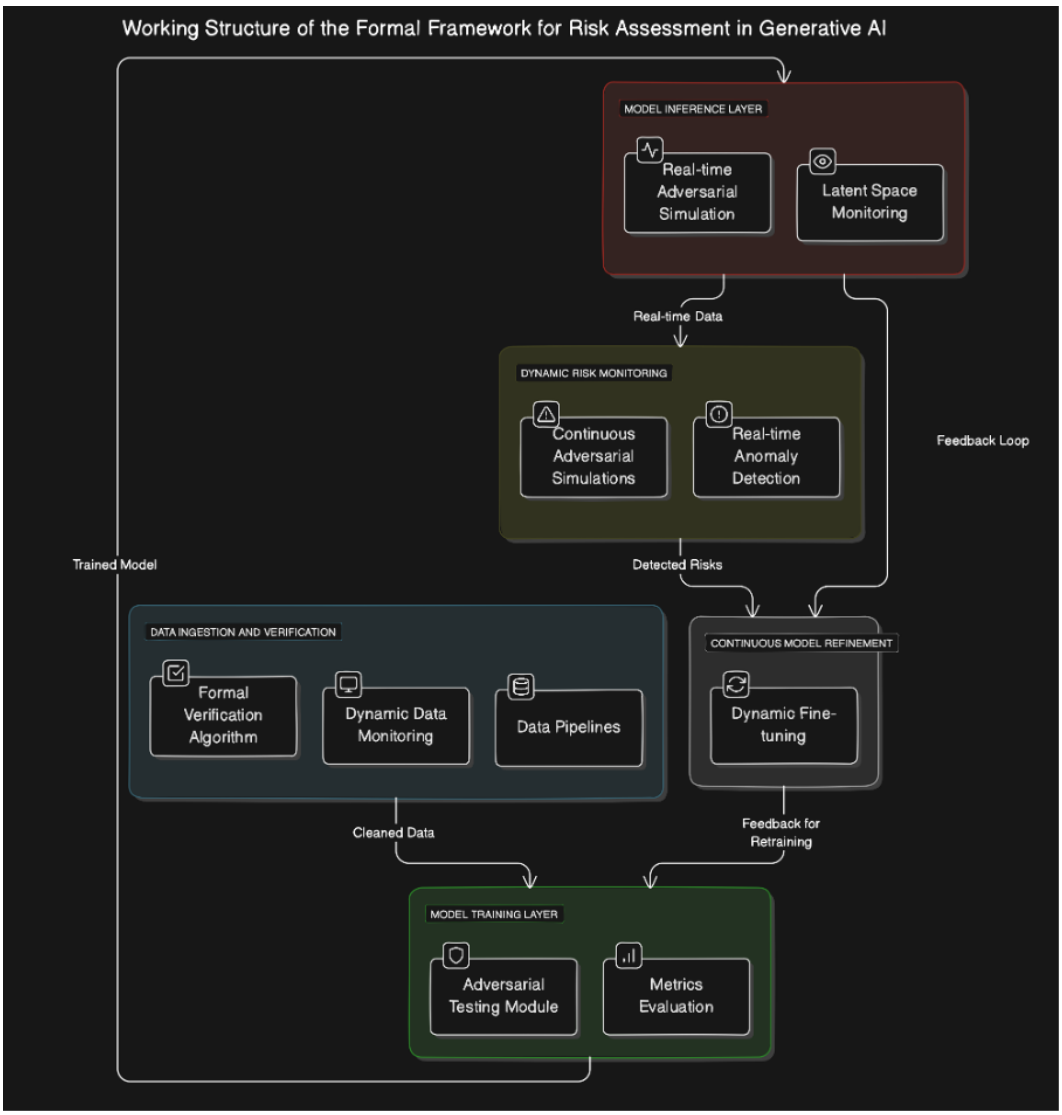

Figure 1: Working structure of the formal framework for risk assessment in Generative AI

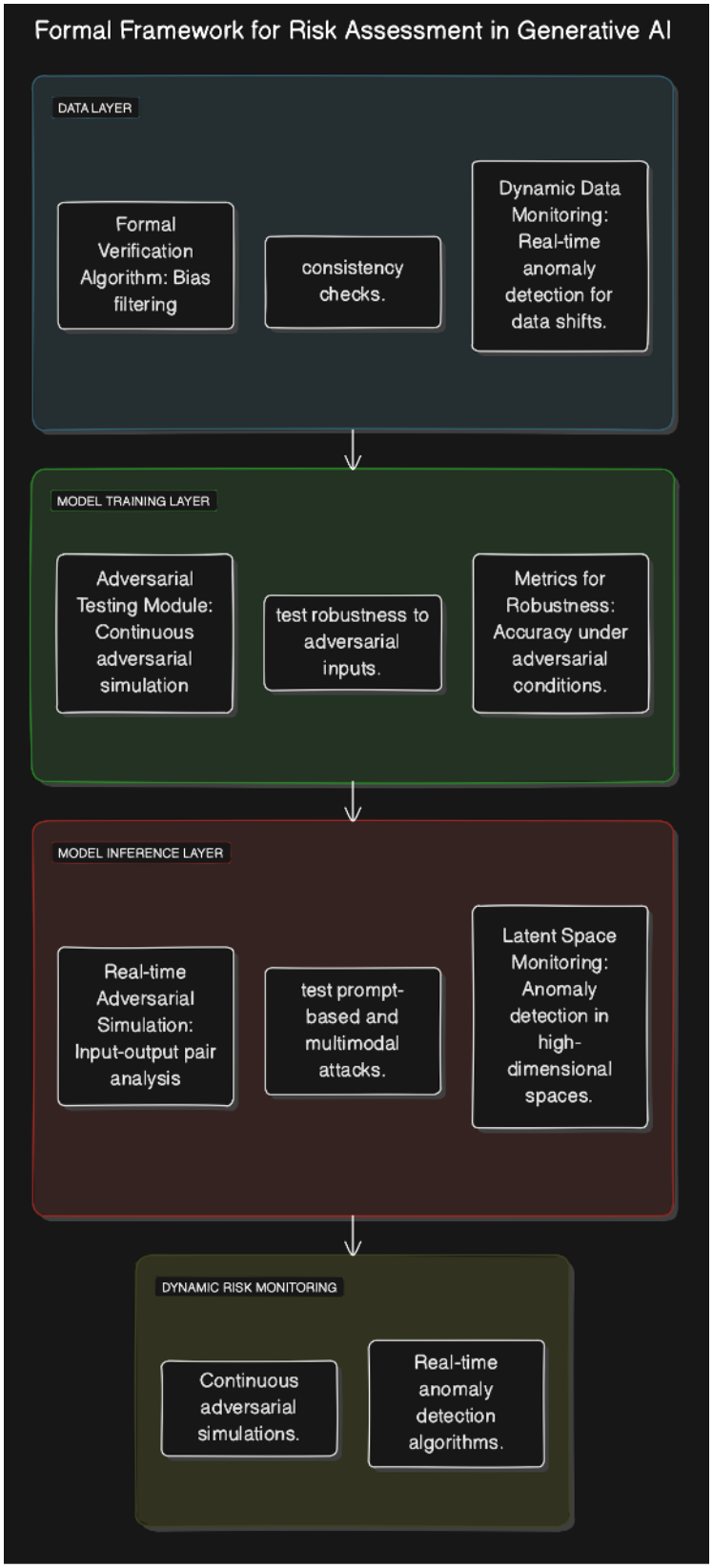

Figure 2: Formal Framework for Risk Assessment in Generative AI