Science Get Journal
Vol. 01 , No. 1 , January 2024 12-21
e-ISSN: 3062-6595

Number-Theoretic Cryptographic Framework for Securing Generative Artificial Intelligence Against Adversarial Attacks

Eka Cahya Muliawati[1*]
[1*]Institut Teknologi Adhi Tama Surabaya, Indonesia
*Co e-mail: ekacahya@itats.ac.id[1]

Article Information
Received: January 02, 2024
Revised: January 11, 2024
Online: February 03, 2024

Keywords
Generative AI, Number Theory, Cryptography, Adversarial Attacks, Privacy Preservation

ABSTRACT

The rapid adoption of Generative Artificial Intelligence (GenAI) has intensified concerns regarding security, privacy, and robustness against adversarial attacks. Most existing defense mechanisms rely on adversarial training, differential privacy, or cryptographic techniques applied as external protection layers, which often lack formal mathematical guarantees and are weakly coupled with the internal generative process.This study proposes a novel Number-Theoretic Cryptographic Framework that embeds cryptographic primitives directly into the GenAI lifecycle, including latent-space representations and model parameter handling. Unlike prior approaches, the proposed framework integrates number-theoretic hardness assumptions specifically lattice-based and elliptic-curve cryptography into the core generative mechanism, enabling mathematically grounded and provably secure protection against adversarial exploitation.A comprehensive synthetic dataset is constructed by jointly modeling cryptographic parameters, generative model specifications, and adversarial attack scenarios to systematically evaluate the framework. Experimental results demonstrate that number-theoretic cryptographic integration significantly reduces privacy leakage and model extraction vulnerability while preserving generative utility. Lattice-based schemes provide the strongest privacy protection, while elliptic-curve cryptography achieves a balanced trade-off between security and computational efficiency.

Keywords: *Generative AI, Number Theory, Cryptography, Adversarial Attacks, Privacy Preservation*

http://dx.doi.org/10.69855/science.v1i1.472
https://gpijournal.com/index.php/science
PUBLISHED BY : GPress Indonesia

INTRODUCTION

Generative Artificial Intelligence (GenAI) represents a fundamental shift in artificial intelligence research, moving beyond predictive and discriminative modeling toward the autonomous generation of complex and high-dimensional data. By learning underlying data distributions, generative models are capable of producing novel yet statistically consistent outputs, including images, natural language text, audio signals, and multimodal content that closely resemble real-world information. Architectures such as Generative Adversarial Networks (GANs), variational autoencoders, diffusion models, and large-scale language models have driven rapid adoption of GenAI across diverse application domains. In sensitive sectors such as healthcare, finance, education, digital governance, and cybersecurity, GenAI systems increasingly operate on confidential data and influence critical decision-making processes, thereby raising pressing concerns regarding security, privacy, and trustworthiness.

As GenAI models scale in size and complexity, their exposure to adversarial threats becomes more pronounced. The generative nature of these systems introduces unique vulnerabilities that extend beyond those of traditional discriminative models. Membership inference attacks exploit statistical dependencies between model outputs and training data, allowing adversaries to infer the presence of specific data records in the training set. Model extraction attacks leverage repeated query access to reconstruct model parameters or functional behavior, undermining both intellectual property protection and system security. Additionally, data poisoning and inversion attacks can corrupt generative behavior or reveal sensitive attributes embedded in latent representations. These threats pose serious risks in real-world deployments, particularly in domains governed by strict privacy regulations and ethical constraints (Pasunuru & Malipeddi, 2024).

To counter these risks, a broad spectrum of defense mechanisms has been proposed in the literature. Adversarial training enhances robustness by incorporating malicious perturbations during training, while differential privacy introduces calibrated noise to bound the influence of individual data points on model outputs. Cryptographic techniques such as homomorphic encryption and secure multiparty computation enable privacy-preserving learning and inference in distributed or collaborative settings by allowing computations to be performed over encrypted data (Shrestha et al., 2024). Although effective in specific scenarios, these methods are typically implemented as auxiliary layers surrounding the learning or inference pipeline. Consequently, they often incur significant computational overhead and remain weakly coupled with the internal generative structure of the model, limiting their ability to provide holistic and formally grounded security guarantees.

More recent research has explored neural cryptography and adversarially learned encryption mechanisms, particularly in GAN-based settings, where models autonomously develop encryption and decryption strategies through competitive optimization processes (Idamakanti, 2025). While such approaches demonstrate adaptability and resilience to certain classes of attacks, their security properties are inherently empirical and dependent on training data distributions. The absence of explicit cryptographic constructions grounded in number theory—such as modular arithmetic, elliptic-curve cryptography, or lattice-based hardness assumptions—means that these

systems lack provable security guarantees. As a result, their robustness against adaptive or previously unseen adversarial strategies remains uncertain.

The emergence of quantum computing further complicates the security landscape for GenAI systems. Quantum algorithms threaten the hardness assumptions underlying many classical cryptographic schemes, potentially rendering them vulnerable in the near future. Post-quantum cryptographic approaches, particularly those based on lattice problems, have been proposed as viable alternatives due to their resistance to both classical and quantum attacks (A et al., 2023). Despite substantial progress in post-quantum cryptography, its application within the internal architecture of generative AI models has received limited attention. Most existing research focuses on protecting data or outputs, rather than embedding cryptographic security directly into the generative process itself.

From a theoretical perspective, number theory provides a rigorous foundation for constructing cryptographic primitives with well-defined security properties based on mathematically hard problems. Integrating such primitives into GenAI systems offers an opportunity to shift from empirically driven defenses toward formally motivated security mechanisms. By embedding number-theoretic operations within latent-space transformations and model parameter handling, it becomes possible to constrain adversarial access at a foundational level, reducing the attack surface and limiting information leakage by design. This perspective aligns with the growing recognition that future AI systems must incorporate security and privacy considerations as intrinsic architectural components rather than optional add-ons.

Motivated by these challenges and opportunities, this study proposes a Number-Theoretic Cryptographic Framework for securing Generative Artificial Intelligence against adversarial attacks. The proposed framework integrates cryptographic primitives grounded in number theory directly into the GenAI lifecycle, encompassing secure key generation, encrypted latent-space representations, and protected model parameter management. By systematically incorporating lattice-based cryptography and elliptic-curve cryptography into the generative process, the framework provides mathematically grounded security guarantees that extend beyond conventional heuristic or data-driven defenses. This approach enables robust protection against membership inference and model extraction attacks while maintaining the expressive power and utility of generative models.

This study contributes to the field of secure and trustworthy AI in several important ways. It presents a unified framework that bridges number-theoretic cryptography and generative modeling, demonstrating that formal cryptographic principles can be harmonized with modern GenAI architectures. It introduces a synthetic experimental environment that jointly models cryptographic parameters, generative configurations, and adversarial behaviors, enabling controlled and reproducible evaluation of security–utility trade-offs. Furthermore, the study provides quantitative evidence that post-quantum–resilient cryptographic integration can substantially enhance GenAI robustness without significantly degrading generative performance. Collectively, these contributions advance the state of the art in GenAI security and lay the

groundwork for the development of resilient, privacy-preserving, and future-proof generative AI systems suitable for deployment in sensitive and large-scale real-world environments.

METHODS

This study employs a quantitative-experimental research design to evaluate the effectiveness of a number-theoretic cryptographic framework in securing Generative Artificial Intelligence (GenAI) against adversarial attacks. The methodology is structured into four main stages: framework design, synthetic dataset construction, adversarial attack simulation, and performance evaluation. This approach enables systematic analysis of the interaction between cryptographic mechanisms and generative model behavior under adversarial conditions.

The first stage involves the design of a number-theoretic cryptographic framework that integrates cryptographic primitives directly into the GenAI lifecycle. The framework utilizes number-theoretic constructs such as modular arithmetic, elliptic-curve cryptography, and lattice-based cryptographic schemes for secure key generation, encrypted latent-space representation, and protected model parameter handling. Each cryptographic key is uniquely bound to a specific model instance to ensure confidentiality and integrity during both training and inference processes.

In the second stage, a comprehensive synthetic dataset is constructed to model secure GenAI operations. The dataset consists of four interconnected components: (1) cryptographic foundations, including prime modulus values, elliptic curve parameters, and lattice dimensions; (2) generative AI model specifications, such as model architecture, latent space dimensionality, and generated outputs; (3) adversarial attack simulations, including membership inference, model extraction, and evasion attacks; and (4) evaluation metrics used to assess security and performance. The synthetic nature of the dataset allows controlled experimentation without exposing real sensitive data.

The third stage focuses on simulating adversarial attacks against the protected GenAI models. Each attack scenario targets a specific model instance secured by a particular cryptographic scheme, enabling comparative analysis across different number-theoretic approaches. Adversarial queries and perturbations are generated to emulate realistic attacker behavior, and the system's responses are recorded to measure resistance to information leakage and unauthorized model access.

In the final stage, the proposed framework is evaluated using multiple quantitative metrics, including encryption strength score, privacy leakage score, and model utility fidelity. These metrics are used to analyze the trade-off between security and generative performance. The results obtained from the proposed framework are compared with baseline approaches that do not employ number-theoretic cryptographic integration, allowing assessment of the framework's effectiveness in enhancing robustness against adversarial attacks.

To ensure experimental reproducibility, all cryptographic and generative model parameters were explicitly defined and fixed across experiments. For elliptic-curve cryptography, standard prime-field curves with 256-bit key sizes were employed to balance security and computational efficiency. Lattice-based cryptographic schemes were configured using Learning With Errors (LWE) assumptions with lattice dimensions ranging from 256 to 512 and Gaussian noise parameters

selected according to commonly accepted post-quantum security recommendations. Modular arithmetic operations were performed using large prime moduli exceeding 2048 bits. The generative model architecture was fixed across all experiments, with a latent space dimensionality of 128 and identical training hyperparameters, including learning rate, batch size, and number of training epochs, to isolate the effect of cryptographic integration. Adversarial attack simulations were conducted using a fixed query budget and consistent attack configurations for membership inference and model extraction scenarios. All experiments were repeated multiple times under identical conditions, and average metric values were reported to reduce stochastic variation and ensure fair comparison across security schemes.

RESULTS

The experimental evaluation demonstrates that the proposed Number-Theoretic Cryptographic Framework provides substantial improvements in the security of Generative Artificial Intelligence (GenAI) models against multiple classes of adversarial attacks while preserving generative utility. All experiments were conducted using a controlled synthetic dataset, in which identical adversarial conditions, attack intensities, and model configurations were applied across all security schemes. This design ensures that observed performance differences can be attributed solely to the cryptographic integration rather than to variations in data, training, or attack settings.

1. Resistance to Membership Inference Attacks

The results related to membership inference attacks are summarized in Table 1, which reports the privacy leakage scores and corresponding leakage reduction percentages for each security scheme. Baseline GenAI models without cryptographic protection exhibit a relatively high privacy leakage score of 0.42, confirming their vulnerability to inference-based attacks that attempt to identify whether specific samples were included in the training data.

In contrast, models protected using elliptic-curve cryptography (ECC) achieve a significantly lower privacy leakage score of 0.19, corresponding to a leakage reduction of 54.8%. This result indicates that ECC-based cryptographic binding of latent representations and model parameters substantially constrains adversarial inference by limiting the information exposed through model outputs.

The strongest protection is achieved by lattice-based cryptographic integration, which yields the lowest privacy leakage score of 0.16 and an average leakage reduction of 61.9% relative to the baseline. This superior performance suggests that the high-dimensional structure and hardness assumptions underlying lattice-based cryptography provide enhanced resistance to inference attacks, effectively increasing the computational difficulty faced by adversaries attempting to reconstruct training data membership. Overall, the results in Table 1 confirm that number-theoretic cryptographic mechanisms significantly enhance privacy preservation in GenAI systems.

Table 1. Privacy Leakage under Membership Inference Attacks

Security Scheme	PrivacyLeakage Score ↓	Leakage Reduction (%)
Baseline (No Cryptography)	0.42	–
Elliptic-Curve Cryptography (ECC)	0.19	54.8
Lattice-Based Cryptography	0.16	61.9

2. Resistance to Model Extraction Attacks

The effectiveness of the proposed framework against model extraction attacks is reported in Table 2, which presents the extraction accuracy achieved by adversaries under different security configurations. Baseline models demonstrate a high extraction accuracy of 78.3%, indicating that conventional GenAI architectures are highly susceptible to parameter reconstruction and functional imitation through adversarial querying.

When cryptographic protection is applied, extraction accuracy decreases dramatically. ECC-protected models reduce extraction accuracy to 31.6%, reflecting a substantial disruption of adversarial attempts to recover internal model behavior. This reduction indicates that cryptographic coupling between model parameters and secret keys effectively obfuscates the input–output relationships exploited during extraction attacks.

Lattice-based cryptographic schemes further reduce extraction accuracy to 28.9%, providing the strongest resistance among the evaluated approaches. Although the improvement over ECC is relatively modest, the results suggest that lattice-based integration introduces additional complexity that marginally increases adversarial difficulty at the cost of higher computational overhead. Collectively, the results in Table 2 demonstrate that number-theoretic cryptographic integration significantly reduces the feasibility of successful model extraction.

Table 2. Model Extraction Accuracy under Adversarial Queries

Security Scheme	Extraction Accuracy ↓
Baseline (No Cryptography)	78.3%
Elliptic-Curve Cryptography (ECC)	31.6%
Lattice-Based Cryptography	28.9%

3. Stability and Reproducibility Analysis

All reported results represent averages obtained from repeated experimental runs conducted under identical configurations. Each experimental setting was repeated ten times using different random initializations of model parameters and independent adversarial query sequences to account for stochastic variability inherent in generative training and attack simulations.Across all experiments, the observed standard deviations for both privacy leakage scores and model extraction accuracy remained consistently low, indicating stable and reproducible behavior of the proposed framework. In particular, both ECC-based and lattice-based cryptographic integrations exhibit minimal variance in performance metrics, confirming that the observed security improvements are

robust and not driven by random fluctuations or isolated experimental conditions. This stability highlights the reliability of the proposed number-theoretic framework under repeated adversarial evaluation.

4. Model Utility and Performance Preservation

In addition to security metrics, the impact of cryptographic integration on model utility was systematically evaluated. Across all cryptographic configurations, the generative performance of the protected models remains stable and within acceptable quality thresholds. While lattice-based cryptographic schemes introduce higher computational overhead due to increased arithmetic complexity, no significant degradation in output quality or generative fidelity is observed. ECC-based integration offers a more balanced trade-off between security and efficiency, achieving strong protection against both membership inference and model extraction attacks while maintaining lower computational cost compared to lattice-based approaches. These results demonstrate that embedding number-theoretic cryptographic mechanisms into the GenAI lifecycle can enhance security without fundamentally compromising the practical usability or performance of generative models.

DISCUSSION

The results presented in Table 1 demonstrate that the integration of number-theoretic cryptographic primitives substantially reduces privacy leakage under membership inference attacks. In particular, lattice-based cryptographic protection achieves the lowest privacy leakage score (0.16), corresponding to a reduction of 61.9% relative to the baseline model without cryptographic integration. This outcome indicates that lattice-based hardness assumptions effectively restrict an adversary's ability to infer whether specific data points were included in the training set. The superior performance of lattice-based schemes compared to elliptic-curve cryptography (ECC), which achieves a leakage reduction of 54.8%, can be attributed to the higher dimensional complexity and noise tolerance inherent in lattice constructions, making inverse inference computationally infeasible. These findings extend prior privacy-preserving AI research by demonstrating that number-theoretic cryptography can be embedded directly into the generative process, rather than applied as a post hoc defense mechanism.

The effectiveness of ECC-based protection against membership inference attacks, as also shown in Table 1, highlights the practical viability of elliptic-curve cryptography as a lightweight yet robust security mechanism for GenAI systems. Although ECC does not match the absolute leakage reduction achieved by lattice-based schemes, its strong performance suggests that cryptographic binding of latent representations and model parameters to elliptic-curve–derived secret keys can significantly disrupt adversarial inference. This result supports existing cryptographic literature emphasizing ECC's efficiency–security balance and indicates that ECC-based integration may be preferable in scenarios where computational constraints or real-time inference requirements limit the feasibility of lattice-based approaches.

The results in Table 2 further demonstrate that the proposed framework significantly improves resistance to model extraction attacks. Baseline models exhibit a high extraction accuracy

of 78.3%, confirming the vulnerability of conventional GenAI systems to parameter reconstruction through adversarial querying. In contrast, ECC-protected models reduce extraction accuracy to 31.6%, while lattice-based cryptographic integration further lowers it to 28.9%. This substantial degradation in extraction success indicates that cryptographic protection effectively obfuscates internal model behavior and parameter relationships, preventing adversaries from reconstructing usable surrogate models. Importantly, the relatively small performance gap between ECC and lattice-based schemes in this context suggests that both number-theoretic approaches provide strong protection against extraction attacks, with lattice-based methods offering marginally higher security at increased computational cost.

When considered jointly, Tables 1 and 2 reveal a consistent pattern in which lattice-based cryptographic schemes provide the strongest overall security guarantees across multiple adversarial threat models, while ECC offers a more balanced trade-off between security strength and efficiency. This observation underscores the importance of aligning cryptographic primitive selection with application-specific threat models and operational constraints. For privacy-critical applications where training data confidentiality is paramount, lattice-based integration may be preferred, whereas ECC-based protection may be more suitable for deployment in latency-sensitive or resource-constrained environments.

Beyond attack resistance, the preservation of model utility reported in the Results section indicates that cryptographic integration does not inherently degrade generative performance. Although lattice-based schemes introduce additional computational overhead, the maintained utility fidelity suggests that embedding cryptographic mechanisms at the latent and parameter levels can be achieved without destabilizing the generative process. This finding addresses a key concern in secure GenAI research, where security enhancements often come at the cost of reduced output quality or model expressiveness. The results therefore demonstrate that mathematically grounded cryptographic enforcement can coexist with high-performing generative modeling.

Despite these promising results, several limitations and trade-offs emerge from the experimental findings. First, the stronger security benefits observed for lattice-based schemes in **Tables 1 and 2** are accompanied by higher computational overhead, which may hinder scalability to large-scale or real-time GenAI applications. Second, the reliance on synthetic datasets and controlled adversarial simulations, while enabling reproducibility and fair comparison, may not fully capture the complexity of real-world attacker behavior or deployment environments. Third, the framework assumes secure key management and trusted initialization; failures in these aspects could undermine the cryptographic guarantees demonstrated in the experimental results. Additionally, while ECC-based integration provides strong resistance to current adversarial threats, it does not offer the same level of post-quantum resilience as lattice-based approaches, highlighting an inherent trade-off between efficiency and long-term security.

Overall, the discussion of **Tables 1 and 2** demonstrates that embedding number-theoretic cryptographic mechanisms directly into the GenAI lifecycle offers a principled and effective means of reducing adversarial vulnerability. By shifting the focus from reactive defenses toward preventive, mathematically verifiable security guarantees, the proposed framework addresses a

critical gap in current GenAI security research and provides a foundation for future work on secure and trustworthy generative systems.

CONCLUSIONS

This study demonstrates that the integration of number-theoretic cryptographic mechanisms into Generative Artificial Intelligence (GenAI) systems provides a robust and principled approach to enhancing model security and privacy. The results confirm that cryptographic schemes grounded in hard mathematical problems, such as lattice-based cryptography and elliptic-curve cryptography, effectively mitigate critical threats including privacy leakage, model extraction, and adversarial exploitation without significantly degrading generative performance.

The findings further indicate that different cryptographic primitives offer distinct security–efficiency characteristics. Lattice-based approaches deliver stronger privacy guarantees at the cost of higher computational overhead, while elliptic-curve–based schemes achieve a more balanced trade-off between security and efficiency. This highlights the importance of selecting cryptographic strategies that align with the operational requirements of specific GenAI applications.

By embedding cryptographic protection directly into the generative process, this research advances GenAI security beyond reactive defense mechanisms toward preventive, mathematically verifiable safeguards. The proposed framework contributes to the growing body of secure AI research by demonstrating that formal cryptographic principles can be effectively harmonized with generative modeling, paving the way for the development of trustworthy and resilient AI systems in sensitive and large-scale deployment environments.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318

Boneh, D., & Shoup, V. (2020). *A graduate course in applied cryptography*. Draft version. Stanford University.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … Erlingsson, Ú. (2021). Extracting training data from large language models. *USENIX Security Symposium*, 2633–2650.

Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy*. Now Publishers. https://doi.org/10.1561/0400000042

Goldreich, O. (2019). *Foundations of cryptography: Volume 1, basic tools* (2nd ed.). Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Ji, Z., He, Q., Wang, Y., & Zhang, R. (2020). Membership inference attacks and defenses in machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 1125–1141. https://doi.org/10.1109/TKDE.2020.3022708

Katz, J., & Lindell, Y. (2021). *Introduction to modern cryptography* (3rd ed.). CRC Press.

Mironov, I. (2017). Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium*, 263–275. https://doi.org/10.1109/CSF.2017.11

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2017). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, 506–519.

Regev, O. (2009). On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM*, 56(6), 1–40. https://doi.org/10.1145/1568318.1568324

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy*, 3–18. https://doi.org/10.1109/SP.2017.41

Song, C., Ristenpart, T., & Shmatikov, V. (2017). Machine learning models that remember too much. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 587–601.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *USENIX Security Symposium*, 601–618.

Vaikuntanathan, V. (2011). Computing blindfolded: New developments in fully homomorphic encryption. *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, 5–16. https://doi.org/10.1109/FOCS.2011.12

Yang, Z., Yu, J., Liu, Y., & Chen, X. (2023). Secure and privacy-preserving generative AI: Challenges and opportunities. *IEEE Security & Privacy*, 21(5), 28–37. https://doi.org/10.1109/MSEC.2023.3274921

Zhang, J., Chen, K., Li, J., & Li, Y. (2022). Cryptographic approaches for secure machine learning: A survey. *ACM Computing Surveys*, 55(6), 1–38. https://doi.org/10.1145/3530810