Journal of Electronic Research and Application, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Security Protection Technologies for Generative Artificial Intelligence and Data Privacy

Yiran Zhai*

University of Toronto, Toronto, Canada

**Corresponding author: Yiran Zhai, yokoko1123@163.com*

Abstract: The widespread application of generative artificial intelligence has raised profound concerns about data privacy. This study systematically analyzed the multidimensional privacy and security risks hidden by this technology throughout the entire lifecycle of data, covering key aspects such as data collection, model training, and content generation. The study further sorted out and explored the current mainstream data privacy and security protection technology system, advocating the construction of a multi-level defense plan covering data, models, outputs, and system architecture. The continuous evolution of technology and the synergy of institutional norms have become the key to balancing technological innovation and privacy protection.

Keywords: Generative AI; Data privacy; Security risk; Security protection

Online publication: March 31, 2026

1. Introduction

Generative AI technology has demonstrated unprecedented content creation capabilities, and its development momentum is deeply rooted in the learning and integration of massive data. This data-intensive nature has transformed privacy and security concerns from peripheral concerns to core issues that constrain their healthy development. The complexity of technical architecture also poses a risk of failure for traditional data protection methods. Exploring a new security protection technology framework that is compatible with it has become an urgent task to promote the good development of this technology, which requires both cutting-edge breakthroughs in technology and forward-looking considerations in governance.

2. Theoretical foundations of generative AI and data privacy security protection

2.1. Core concepts, technical architecture, and data flow of generative AI

Generative AI is an AI technology that relies on machine learning techniques to generate new text, images, and

other content based on existing data patterns. Unlike traditional AI which focuses on judgment and classification, its core is to create based on existing data rather than simple recognition. The technical architecture of generative AI mainly consists of a basic model layer, a data processing layer, and an interaction layer. The basic model layer is the core support, commonly including transformer architecture and diffusion models. The data processing layer is responsible for organizing and filtering various types of raw data, while the interaction layer receives user instructions and outputs generated results. The data process begins with data collection. The staff will summarize various types of raw data that are publicly available or authorized, and then perform preprocessing such as denoising, deduplication, and sensitive information removal on these data to avoid low-quality data affecting the performance of the model. The preprocessed high-quality data will be input into the basic model for training. During the training process, the model will autonomously learn the rules in the data, and after training, it will call relevant data to generate corresponding content based on user instructions. The flow and processing of data throughout the entire process are closely related to subsequent data privacy and security protection [1].

2.2. Analysis of the correlation between generative AI and data security

The correlation between generative AI and data security revolves closely around the entire process of the data lifecycle. The theory of data security emphasizes the three core goals of confidentiality, integrity, and availability, and the operational process of generative AI profoundly affects the implementation status of these goals. During the data collection phase, a large amount of personal or sensitive information is included in the training set, directly challenging the principle of data confidentiality. The dependence of the model training process on data quality is related to the reliability of the learning results, that is, the integrity of the data. The usability and controllability of the final generated content are constrained by inherent biases and security risks in the training data. The specific correlation between data security requirements and the characteristics of generative artificial intelligence technology is shown in **Table 1**.

Table 1. Correlation analysis between characteristics of generative AI technology and data security objectives

Core objectives of data security	Corresponding key stages of generative AI	Core focus and potential risks
Confidentiality	Data collection and preprocessing	Sensitive information such as personal privacy and trade secrets contained in raw data are at risk of leakage during the collection and storage process
Integrity	Model training and optimization	The quality, representativeness, and malicious contamination of training data directly determine the accuracy and fairness of the model output
Usability	Generate output and services	The generated content must meet the requirements of trustworthiness, reliability, and ethical standards, and avoid generating harmful or misleading information

3. Analysis of data privacy security risks in generative AI applications

3.1. Privacy leakage risks in data collection and preprocessing

In the data collection process of generative AI, developers often excessively crawl Internet data. Some developers failed to verify the legitimacy of the data source in order to rapidly expand the amount of training data, and even crawled unpublished personal social information, contact information and other sensitive content. After these privacy data that should be protected were collected at will, it was difficult to fully identify and eliminate them even when they entered the preprocessing stage. The privacy leakage risk in the preprocessing stage is

mainly due to the negligence of developers. Some developers, in order to save time and costs, did not perform standardized desensitization on the collected data, only simply deleting obvious sensitive identifiers, but ignoring the privacy-related information hidden in the data. Some developers temporarily store sensitive data that has not been processed on unencrypted servers, which can easily be obtained by attackers through vulnerabilities. Some companies have even experienced situations where employees mistakenly input confidential data into generative AI tools, resulting in privacy breaches. These actual problems make the collection and preprocessing stages a high-risk area for privacy breaches [2].

3.2. Memory and inference attack risks during model training

The partially sensitive data collected and preprocessed, even after simple desensitization, will still be unintentionally remembered by the model during the training process. The model will not distinguish the sensitivity of the data and will retain all the input training data features intact. When encountering similar user instructions in the future, the sensitive information in these memories may be directly output. For example, training data containing personal identity information may be unintentionally leaked by the model when generating content. Attackers will use this feature of the model to carry out inference attacks. They will construct specific prompt words to hide malicious instructions in seemingly harmless inputs, inducing the model to output sensitive data memorized during the training process. Some attackers may even activate malicious instructions after specific user operations through delayed triggering, tampering with the model's long-term memory and obtaining private information. Some AI models with long-term memory functions have been used by attackers to inject false information and steal privacy. The indiscriminate retention of memory during the model training process further amplifies the privacy leakage risks brought by this inference attack.

3.3. Content traceability and information leakage risks in generation output

When the model generates output content after training, it is often difficult to achieve accurate content traceability and clarify the original training data source corresponding to the output content. Once the output content contains sensitive privacy information, both staff and users find it difficult to trace the source of the information and quickly identify the leaking nodes. In many office scenarios, AI-generated content contains non-desensitized privacy information but cannot be traced. The risk of information leakage in the output content comes from both the non-standard output of the model itself and improper use by users. Sometimes, the model unintentionally mixes personal information and internal data memorized during the training process when generating content, and some users lack privacy protection awareness and may copy, forward, or even upload AI-generated content containing sensitive information to public network platforms. Additionally, there are individual users who generate and output confidential content through AI and spread it illegally without authorization, resulting in widespread leakage of private information and internal data. These common problems in practical scenarios make the generation and output stage another key link in privacy leakage, and also highlight the subsequent hidden dangers caused by the lack of content traceability.

3.4. Emerging attack surfaces in system deployment and interaction

The widespread deployment and open interaction of generative AI systems have exposed a new type of attack surface that is completely different from pure research and development environments. The attack surface first extends to its integration interface with various external tools and databases. Through the method of "prompt

word injection", attackers can disguise malicious instructions as ordinary user input, deceive AI assistants to perform unauthorized data access or leakage operations. The multimodal capability expands this risk, and cleverly incorporating semantic instructions of image pixels or audio can effectively avoid the scrutiny of traditional text filtering mechanisms. More complex threats come from intelligent agents with autonomous planning and execution capabilities, which, once maliciously induced, may launch chain attacks between multiple systems at machine speed. The root cause of these attacks being effective is that artificial intelligence systems are inherently difficult to distinguish between user data requests and executable operational instructions. When they have the ability to interpret natural language, access critical resources, and perform operations simultaneously, it constitutes a high-risk attack path [3].

3.5. Technical deficiencies in data lifecycle management

In the application practice of generative AI, technical deficiencies in data lifecycle management run through every link from collection to destruction, becoming amplifiers of systemic privacy risks. Data governance often lacks a unified technical framework that covers the entire process, resulting in fragmented toolchains and inconsistent standards at different stages. This technological decentralization makes it difficult to continuously track and verify the quality and security status of data as it flows across systems. Especially in the process of data cleaning and annotation, automated tools have limitations in recognizing and processing sensitive information, while manual operations inevitably introduce omissions and subjective biases, which may result in harmful or biased low-quality data entering the training process without effective purification. The more tricker part is the lack of effective technical means to ensure that the data is completely destroyed or irreversibly desensitized after completing its training mission. These "dormant" copies of data in backup systems or legacy environments constitute long-term sources of leakage that are difficult to detect. These isolated technological shortcomings are interconnected in the chain of data lifecycle, which may ultimately lead to systemic security crises far beyond single point issues [4].

4. Core technologies for generative AI data privacy protection

4.1. Data-level privacy protection: Differential privacy, federated learning, data de-identification

The privacy protection technology at the data level constitutes the first line of defense against data risks in generative artificial intelligence. Among them, differential privacy technology provides quantifiable strong privacy guarantees for data publishing by adding carefully designed mathematical noise to data or query results. The United Nations High Commissioner for Refugees (UNHCR) uses this technology to synthesize refugee datasets that can be used for public research. Federated learning provides another approach, allowing multiple participants to collaboratively train models without exchanging raw data, only sharing encrypted model parameter updates, effectively addressing the risk of model reverse attacks. Some cutting-edge research attempts to incorporate fully homomorphic encryption into it to provide stronger security. As a more direct and widely used technology, data anonymization involves transforming, replacing, or generalizing sensitive fields according to specific scenarios. In practice, various modes such as static anonymization (permanent replacement of stored data) and dynamic anonymization (real-time deformation display based on access permissions) have emerged to serve the entire process from data development to production queries. These technologies each focus on their own strengths and work together to establish a strong foundation of privacy from the very beginning of data utilization [5].

4.2. Model-level security: Adversarial training, model watermarking, secure alignment

The security technology at the model level enhances the security of generative AI by directly intervening in its training mechanism and internal structure. Adversarial training actively adds imperceptible small perturbations (adversarial samples) to the training data, enhancing the robustness of the model to malicious induced inputs while repeatedly learning to recognize and resist these interferences, thereby reducing the risk of the model being deceived by attack methods such as prompt word injection. Model watermarking technology involves embedding unique and detectable signals in the weights or specific activation patterns of the model. These signals can be extracted and verified through specific algorithms when the model is illegally copied or stolen, providing a technical basis for intellectual property protection and leak tracing of the model. Security alignment technology mainly relies on methods such as reinforcement learning based on human feedback. By allowing the model to learn and internalize a complex reward function that complies with security and ethical standards, it constrains its output from the decision-making mechanism, enabling it to actively refuse to execute instructions that may leak training data privacy or produce harmful content [6]. These technologies work together from different dimensions to build generative models that are inherently more robust and have more controllable behavior.

4.3. Output-level control technologies: Content filtering, traceability techniques, output review

Output-level control technologies serve as the final security gate before generative AI applications, designed to manage generated content effectively. Content filtering systems typically combine rule-based pattern recognition (e.g., keyword blacklists) with machine learning classification models, conducting multi-stage scans of output text to identify and block sensitive content containing personal identifiers or specific biases. Tracing technologies embed imperceptible unique identifiers into generated content. For instance, implanting visually undetectable digital watermarks in AI-generated images via specialized algorithms, or embedding character patterns adhering to specific statistical distributions in text, enabling source model or specific generation session identification in case of unauthorized content leakage [7]. Output review mechanisms further introduce manual or semi-automated verification processes to conduct secondary checks on high-risk scenarios, with their decision logic abstracted into a simplified risk assessment function: risk level = f (content sensitivity, application scenario criticality). The system then determines whether to approve directly, log alerts, or escalate to manual review based on this evaluation. These technologies work in concert to ensure the final information delivered to users is secure and compliant.

4.4. Architecture and system-level protection: Secure multi-party computation, trusted execution environments, encrypted inference

Architectural and system-level protection technologies aim to establish a trusted foundation for safeguarding data privacy in underlying computing environments. Secure multi-party computation enables two or more parties to collaboratively execute a computational task and obtain results without revealing their original input data, with its mathematical foundation ensuring no party can infer the private information of others from the interaction process. Trusted execution environments create a protected, isolated “enclave” within the CPU through hardware isolation techniques, guaranteeing that code and data loaded inside remain confidential and intact even if the operating system is compromised. Current mainstream implementations include Intel’s SGX and ARM’s TrustZone. Encryption inference technology goes a step further, allowing AI models to directly perform inference operations on encrypted user input data and return encrypted results without decryption, relying on cryptographic schemes such as fully homomorphic encryption. Although it still faces significant computational performance challenges,

it has begun exploratory applications in highly sensitive scenarios like finance. These foundational technologies provide robust security support for upper-layer applications [8].

4.5. Privacy risk assessment and continuous monitoring technologies

Privacy risk assessment and continuous monitoring technologies form a management loop for dynamically addressing data privacy challenges posed by generative AI. This technical system first employs automated tools to systematically scan and map data flows, model interfaces, and system configurations, identifying potential vulnerabilities and compliance gaps across data processing stages. The continuous monitoring module relies on sensors deployed at critical nodes to collect real-time model query logs, data access patterns, and system behavior metrics, comparing them against predefined normal behavior baselines. When deviations from these baselines occur, such as high-frequency access to specific sensitive data or unexpected patterns in model outputs, the system triggers alerts. More importantly, the technical system can iteratively update its assessment models and monitoring rules based on feedback data, enabling it to adapt to evolving external attack tactics and internal system changes, thereby achieving a transition from static evaluation to dynamic self-adaptive security capabilities [9].

5. Conclusion

The data privacy and security protection of generative artificial intelligence is a complex system engineering. Relying solely on a certain technology to break through is difficult to build a solid defense line. It calls for a collaborative evolution from algorithm optimization to system design, from technical standards to governance principles. The future exploration direction will pay more attention to seeking a delicate balance between ensuring data utility and protecting personal privacy, and promoting security mechanisms to be endogenous in the process of technological development. Only by finding a dynamic balance between innovation and protection can this transformative technology truly and reliably integrate into every corner of social development.

Disclosure statement

The author declares no conflict of interest.

References

[1] Liu Y, Hua Y, Chen Q, 2023, Research on Key Technologies of Data Security and Privacy Protection in Internet of Things Group Intelligence. Optical and Quantum Electronics, 56(1): 8.

[2] Liu Y, 2024, Research on Privacy Protection and Security Technology of New Media Big Data Platforms. Academic Journal of Management and Social Sciences, 9(1): 11–15.

[3] Zhang Z, 2025, Research on Security Protection Technologies for Generative Artificial Intelligence and Data Privacy. Cybersecurity Technology and Application, 2025(4): 75–77.

[4] Liu Y, Hua Y, Chen Q, 2024, Retraction Note: Research on Key Technologies of Data Security and Privacy Protection in Internet of Things Group Intelligence. Optical and Quantum Electronics, 56(10): 1754.

[5] Shin H, Ryu K, Kim Y, et al., 2024, Application of Privacy Protection Technology to Healthcare Big Data. Digital Health, 2024(10): 20552076241282242.

[6] Gao Y, Guo L, Zhang T, 2023, Exploring and Envisioning the Application of Blockchain Technology for Privacy

Data Protection, Proceedings of the 5th International Conference on Computing and Data Science (Part 4), 152–160.

[7] Wu H, Bhola J, Neware R, et al., 2023, Key Technologies of Data Security and Privacy Protection in the Internet of Things Group Intelligence Perception. Recent Advances in Electrical & Electronic Engineering, 16(2): 135–142.

[8] Han L, Jia M, 2025, Generative AI Empowering Cybersecurity: Challenges and Opportunities. Research in Scientific Management, 43(6): 93–99.

[9] Xu S, Lü J, Jiang D, 2025, Privacy Data Protection Methods for Generative Artificial Intelligence. Computer & Modernization, 2025(11): 10–15+31.