



Review Article

The Dual Use of Generative Artificial Intelligence in Cybersecurity: A Survey from Deepfake Attacks to Intelligent Defense

Dalin Xiang^{1,*} , Xinkai Fu²

¹Department of General Mechanics and Foundations of Mechanics, Hainan International Commercial Aerospace Launch Co., Ltd, Wenchang, China

²Department of Electronic and Communication Engineering, Hainan International Commercial Aerospace Launch Co., Ltd, Wenchang, China

Abstract

The rapid iteration of generative Artificial Intelligence (AI) is reshaping the cybersecurity offense-defense landscape. This paper systematically reviews research progress in this field from 2021 to 2026, focusing on the dual use of generative AI in cybersecurity. On the one hand, malicious actors exploit generative AI to create deepfake content, automate phishing emails, and conduct social engineering attacks, significantly improving the personalization and scalability of attacks. On the other hand, the same technological system demonstrates significant effectiveness in defense scenarios such as threat detection, vulnerability remediation, data augmentation, and security knowledge management. Existing reviews often treat attack and defense separately, lacking an integrated analytical framework. This paper analyzes the field from four dimensions: attack surface expansion, defense technology evolution, governance frameworks, and evaluation benchmarks. We first review generative AI-driven cyberattacks, covering deepfake attacks, automated phishing, automated malware development, and intelligent social engineering. We then examine defense applications, including intelligent threat detection, data augmentation and class imbalance handling, vulnerability detection and automated remediation, security knowledge management, and security education. Next, we identify key shortcomings in current research regarding adversarial robustness, privacy protection, interpretability, and deployment constraints. We further outline five future research directions: adaptive defense and adversarial robustness, federated learning and privacy protection, interpretability and human-machine collaboration, interdisciplinary governance frameworks, and multimodal security evaluation benchmarks. This review aims to provide a structured reference for building adaptive defense systems and interdisciplinary governance paths in the era of generative AI.

Keywords

Generative Artificial Intelligence, Cybersecurity, Deepfake, Adversarial Attacks, Intelligent Defense

*Correspondence: Dalin Xiang (709366545@qq.com)

Received: 11 September 2026; **Accepted:** 20 September 2026; **Published:** 30 September 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

With the rapid iteration of large language models and generative AI technologies such as ChatGPT, DALL-E, and GPT-4, the cybersecurity offense-defense landscape is undergoing structural changes. Radanliev et al. [1] pointed out that generative AI not only expands the technical means of security defense but also gives rise to new forms of attack. In their systematic review of LLM security applications, Ferrag et al. [2] emphasized that generative AI has penetrated multiple sub-fields such as hardware design security, intrusion detection, and malware analysis, with an impact far exceeding early expectations. Ugale and Sinha [3] summarized this phenomenon as the parallel trends of “democratization of attacks” and “intelligent defense”: the lowering of technical barriers enables attackers to launch high-quality attacks at lower cost, while defenders also gain automated threat analysis and response capabilities. This paper starts from four dimensions—attack surface expansion, defense technology evolution, governance frameworks, and evaluation benchmarks—and systematically reviews the dual use of generative AI in cybersecurity from 2021 to 2026. The literature search covers major academic platforms such as PubMed, IEEE Xplore, ScienceDirect, ACL Anthology, and NeurIPS.

2. Generative AI-Driven Cyberattacks

2.1. Deepfake Attacks

Deepfakes are the most intuitive application form of generative AI in the attack domain. Radanliev et al. [1] pointed out that generative models can synthesize highly realistic video, audio, and text content, which is widely used in social engineering attacks for identity impersonation and reputation damage. Ugale and Sinha [3] classified deepfake attacks into three categories—video forgery, audio forgery, and text forgery—corresponding to the technical paths of GANs, speech synthesis, and large language models, respectively. Yigit et al. [4] further found that generative AI can customize attack content based on the target victim's social profile, significantly improving attack success rates.

In recent years, the technical focus of deepfake attacks has been extending from “content generation” to “detection evasion.” Surveys at IEEE PuneCon and IEEE Recent Trends conferences [5, 6] indicate that adversarial perturbations and anti-forensic attacks against deepfake detection systems constitute a new threat dimension. By injecting adversarial noise or obfuscating generative model fingerprints, attackers can effectively reduce detector accuracy, forming an adversarial co-evolution of “generation—detection—evasion.”

2.2. Automated Phishing Attacks

The reshaping of phishing attacks by generative AI is reflected at three levels: personalization, automation, and

multilingualization. The Fraud-R1 benchmark proposed by Yang et al. [7] at ACL 2025 shows that large language models exhibit strong deceptive capabilities in multi-turn conversational phishing inducement, and traditional detection rules based on keyword matching are difficult to cope with. Research by Cheng et al. [8] at NeurIPS 2025 further revealed that adversarial rewriting techniques can “humanize” machine-generated phishing text, thereby bypassing AI text detectors. This evasion capability significantly increases the difficulty of identifying phishing attacks.

2.3. Automated Malware Development

Generative AI is changing the malware development paradigm. Ugale and Sinha [3] pointed out that generative AI can automatically generate malicious code based on attack targets and achieve polymorphic attacks by rapidly generating code variants. Research by Ayyaz and Malik [9] shows that the combination of GAN and GPT technologies enables malware to possess self-evolution capabilities, allowing it to adjust attack strategies according to environmental changes. Although the DCodeBERT model developed by Bensaoud and Kalita [10] is used for vulnerability detection and remediation, its technical principles similarly reveal the dual-use potential of code generation models in the security domain.

2.4. Intelligent Upgrading of Social Engineering Attacks

The core of social engineering attacks lies in exploiting human weaknesses, and generative AI happens to provide powerful tools in this dimension. Yigit et al. [4] analyzed the application of models such as GPT-4 and Gemini in social engineering attacks, pointing out that they can simulate real conversations, automatically collect target information, and adjust attack strategies in real time. The Best-of-N jailbreak method proposed by Hughes et al. [11] at NeurIPS 2025 further demonstrated that through multiple sampling and screening, attackers can bypass model safety alignment mechanisms and generate more persuasive social engineering content.

3. Applications of Generative AI in Cybersecurity Defense

3.1. Intelligent Threat Detection

Generative AI has achieved quantifiable results in the field of threat detection. The GAIA-FL framework proposed by Aouedi and Boissel [12] combines generative AI with federated learning, using an adaptive generation control mechanism to synthesize minority-class attack samples, significantly improving intrusion detection performance in data-scarce

environments. A study in the IEEE Internet of Things Journal [13] similarly confirmed that the combination of federated learning and generative data augmentation can improve detection accuracy while protecting privacy. In their survey of LLM attack and defense technologies, Liao et al. [14] identified key shortcomings in current defense systems: the lack of adaptive and scalable defenses, insufficient detection of adversarial attacks, and the scarcity of general defense mechanisms.

3.2. Data Augmentation and Class Imbalance Handling

Class imbalance is a common problem in cybersecurity data. Experiments by Aouedi and Boissel [12] showed that minority-class samples synthesized by generative AI can effectively improve the performance of detection models on imbalanced datasets. Research in the IEEE IoT Journal [13] further showed that under a federated learning framework, generative data augmentation can maintain system functionality even when 50% of nodes are compromised, providing a feasible path for balancing privacy protection and security effectiveness.

3.3. Vulnerability Detection and Automated Remediation

The application of generative AI in the field of software security is evolving from assisted detection to automated remediation. The DCodeBERT model proposed by Bensaoud and Kalita [10] demonstrated vulnerability detection and remediation performance superior to baselines across multiple programming languages. Research by Jiang et al. [15] at IEEE S&P 2025 revealed another side: LLM-based vulnerability detectors are themselves vulnerable to black-box adversarial attacks, where attackers can disable the detector through minor perturbations. This finding suggests that the reliability boundaries of automated remediation capabilities need to be carefully evaluated.

3.4. Security Knowledge Management and Decision Support

Piemonti et al. [16] explored methods for enhancing cybersecurity knowledge bases using multiple LLMs, constructing a structured security knowledge system by generating attack-mitigation pairs and their prioritized execution order. Experiments showed that over 70% of responses generated by models such as Claude 3.5 and DeepSeek R1 were rated as high quality, demonstrating the effectiveness of generative AI in professional security knowledge processing. This direction elevates generative AI from a "tool" to a component of "knowledge infrastructure."

3.5. Security Education and Skill Development

The generative AI-based virtual mentoring system

proposed by Yamanaka et al. [17] promotes learners' critical thinking by encouraging reasoning, asking "why" questions, and providing credible hints. The system uses retrieval-augmented generation technology to obtain accurate information from authoritative sources, significantly improving mentoring quality. From a broader perspective, generative AI faces an inherent tension in security education: learners need to understand AI's security risks, while AI itself can serve as a teaching tool for understanding these risks.

4. Technical Challenges, Risks, and Governance Frameworks

4.1. Adversarial Attacks and Model Vulnerability

The adversarial robustness of generative AI models themselves constitutes a recursive security challenge. Radanliev et al. [1] pointed out that model poisoning, data leakage, and output manipulation are the three core risks. Liao et al. [14] listed adversarial attack detection, general defense mechanisms, and adaptive scalable defense as open problems urgently in need of resolution, and emphasized the importance of interdisciplinary cooperation and ethical considerations in risk mitigation. The increase in defense system complexity may itself introduce new attack surfaces.

4.2. Privacy Protection and Ethical Frameworks

Generative AI faces multidimensional privacy challenges in security applications. Ugale and Sinha [3] emphasized the need to establish governance mechanisms at three levels: data privacy protection, algorithmic transparency, and ethical frameworks. Research by Aouedi and Boissel [12] showed that the combination of federated learning and differential privacy can alleviate privacy dilemmas to a certain extent, but the trade-off between privacy budget and detection performance still requires systematic study.

4.3. Technical Complexity and Deployment Constraints

The technical complexity of generative AI imposes constraints on practical deployment. Ferrag et al. [2] pointed out that increases in model parameters can lead to degraded performance on hardware with limited capabilities, sometimes even performing worse than smaller models. Piemonti et al. [16] also showed that in such cases, advanced prompt engineering techniques are needed to maintain output quality. Therefore, the deployment of generative AI in the security domain requires finding a balance among model capability, inference efficiency, and operational cost.

5. Future Research Directions

5.1. Adaptive Defense and Adversarial Robustness

The core limitation of current defense systems lies in their static nature. Liao et al. [14] listed adaptive scalable defense as one of the most urgent open problems. Future defense systems need to shift from "robust at training time" to "adaptive at runtime," maintaining detection capability against emerging threats through continuous learning and online updates.

5.2. Federated Learning and Privacy Protection

Federated learning provides a path for generative AI security applications that balances privacy and effectiveness. Experiments by Aouedi and Boissel [12] showed that the combination of differential privacy and generative data augmentation can maintain system functionality when some nodes are compromised, but the trade-off relationship between privacy and performance still requires in-depth exploration.

5.3. Interpretability and Human-Machine Collaboration

The "black-box" nature of generative AI in security decision-making limits its credibility. Explainable AI technologies need to be deeply integrated with security scenarios, enabling analysts to understand the basis and confidence level of AI judgments, while also considering cognitive load and trust calibration in human-machine collaboration.

5.4. Interdisciplinary Governance Frameworks

The cybersecurity governance of generative AI involves law, ethics, organizational behavior, and international relations. Current governance measures in organizations deploying generative AI security tools clearly lag behind technology adoption. In the future, it is necessary to integrate technical standards, legal norms, and ethical guidelines to establish dynamic governance mechanisms.

5.5. Multimodal Security Evaluation Benchmarks

Existing security evaluation benchmarks mostly target a single modality, while actual attacks often involve coordinated generation of multimodal content. Surveys on deepfake detection at IEEE conferences [5, 6] pointed out that audio-visual joint detection is an important direction for multimodal security evaluation, but there is currently a lack of unified evaluation frameworks and standard datasets. Building multimodal security evaluation benchmarks covering text, images, audio, and video is foundational work for promoting the systematic development of this field.

6. Conclusion

Generative artificial intelligence exhibits distinct dual-use characteristics in cybersecurity. On the attack side, it lowers the technical barriers to deepfakes, automated phishing, and malware development, bringing the personalization and scalability of attacks to new levels; on the defense side, it demonstrates considerable potential in threat detection, data augmentation, vulnerability remediation, and security knowledge management. These two forces are tightly coupled through the adversarial co-evolution mechanism of "generation—detection—evasion." Future research needs to focus on explainable AI, federated learning, adaptive defense, and multimodal evaluation benchmarks, while strengthening the construction of interdisciplinary governance frameworks. Only through the coordinated evolution of technological innovation and institutional norms can we fully leverage the positive role of generative AI in cybersecurity and minimize its potential risks.

Abbreviations

AI	Artificial Intelligence
LLM	Large Language Model
GPT	Generative Pre-trained Transformer

Author Contributions

Dalin Xiang: Conceptualization, Resources, Supervision, Writing – original draft, Writing – review & editing

Xinkai Fu: Data curation, Investigation, Methodology, Validation, Visualization

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Radanliev, P., Santos, O., & Ani, U. D. (2025). Generative AI cybersecurity and resilience. *Frontiers in Artificial Intelligence*, 8, 1568360. <https://doi.org/10.3389/frai.2025.1568360>
- [2] Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, 5, 1–18. <https://doi.org/10.1016/j.iotcps.2025.01.001>
- [3] Ugale, S. P., & Sinha, A. (2025). Generative AI in cybersecurity: Threats, vulnerabilities, and protective measures. *2025 International Conference on Future Technologies (ICFT)*, 1–8. <https://doi.org/10.1109/ICFT63334.2025.10915338>

- [4] Yigit, Y., Buchanan, W. J., Tehrani, M. G., & Maglaras, L. (2025). Review of generative AI methods in cybersecurity. *Internet of Things and Cyber-Physical Systems*, 5, 241–261. <https://doi.org/10.1016/j.iotcps.2026.04.001>
- [5] Deepfake survey: Detection methods, challenges and ethical implications. (2025). 2025 IEEE International Conference on Recent Trends in Engineering and Technology, 1–6. <https://doi.org/10.1109/ICRTET63335.2025.11123903>
- [6] A comprehensive survey of deepfake detection methods, datasets, and future directions. (2026). 2025 IEEE Pune Section International Conference (PuneCon), 1–8. <https://doi.org/10.1109/PuneCon66778.2025.11273051>
- [7] Yang, S., Zhu, S., Wu, Z., Wang, K., Yao, J., Wu, J., Hu, L., Li, M., Wong, D. F., & Wang, D. (2025). Fraud-R1: A multi-round benchmark for assessing the robustness of LLM against augmented fraud and phishing inducements. *Findings of the Association for Computational Linguistics: ACL 2025*, 4374–4420. <https://doi.org/10.18653/v1/2025.findings-acl.226>
- [8] Cheng, Y., Sadasivan, V. S., Saberi, M., Saha, S., & Feizi, S. (2025). Adversarial paraphrasing: A universal attack for humanizing AI-generated text. *Advances in Neural Information Processing Systems (NeurIPS 2025)*. <https://doi.org/10.48550/arXiv.2506.07001>
- [9] Ayyaz, S., & Malik, S. M. (2024). A comprehensive study of generative adversarial networks (GAN) and generative pre-trained transformers (GPT) in cybersecurity. 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), 1–8. <https://doi.org/10.1109/ICDS62089.2024.10756505>
- [10] Bensaoud, A., & Kalita, J. (2025). Advancing software security: DCodeBERT for automatic vulnerability detection and repair. *Journal of Industrial Information Integration*, 45, 100834. <https://doi.org/10.1016/j.jii.2025.100834>
- [11] Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Somani, A., Koyejo, S., Sleight, H., Jones, E., Perez, E., & Sharma, M. (2025). Best-of-N jailbreaking. *Advances in Neural Information Processing Systems (NeurIPS 2025)*. <https://doi.org/10.48550/arXiv.2412.03556>
- [12] Aouedi, O., & Boissel, A. (2025). GAIA-FL: Generative AI-augmented federated learning for intrusion detection system. 2025 IEEE Conference on Network Security, 1–9. <https://doi.org/10.1109/TWCMC65282.2025.11059687>
- [13] Efficient and privacy-preserving network intrusion detection based on federated learning in SDN-enabled IIoT network. (2025). *IEEE Internet of Things Journal*, 12(20), 1–15. <https://doi.org/10.1109/JIOT.2025.3555202>
- [14] Liao, Z., Chen, K., Lin, Y.-C. A., Li, K., Liu, Y., Chen, H., Huang, X., & Yu, Y. (2025). Attack and defense techniques in large language models: A survey and new perspectives. *Neural Networks*, 196, 108388. <https://doi.org/10.1016/j.neunet.2025.108388>
- [15] Jiang, Y., Huang, S., Treude, C., Su, X., & Wang, T. (2026). Shield broken: Black-box adversarial attacks on LLM-based vulnerability detectors. *IEEE Transactions on Software Engineering*, 52(1), 246–265. <https://doi.org/10.1109/TSE.2025.3638998>
- [16] Piemonti, A., Cianchini, V., Danousis, M., & Skianis, C. (2025). Organizing and augmenting cybersecurity knowledge using generative AI. 2025 IEEE 11th International Conference on Network Softwarization (NetSoft), 585–590. <https://doi.org/10.1109/NetSoft64993.2025.11080630>
- [17] Yamanaka, M., Watanabe, K., & Hasegawa, H. (2025). A generative AI mentoring system for cultivating cybersecurity skills. 2025 IEEE Global Conference on Consumer Electronics (GCCE), 1–5. <https://doi.org/10.1109/GCCE63974.2025.10949695>