

Generative AI and cybersecurity: Exploring opportunities and threats at their intersection

Kunter Orpak

Received 10 February 2025 | Accepted 16 April 2025 | Published 11 September 2025

Abstract

Generative AI, particularly large language models (LLMs), is reshaping the cybersecurity landscape by enabling both innovative defense mechanisms and novel forms of attack. This article explores the dual role of generative AI in both offensive and defensive cybersecurity operations. While GenAI offers significant advancements in defensive capabilities, it is also being leveraged by nation-state actors to enhance the sophistication and success rates of cyberattacks. The article analyzes how LLMs are applied in offensive engagements such as red teaming, penetration testing, and threat intelligence, while also identifying emerging technical, operational, and strategic risks associated with their deployment. Special attention is given to the cybersecurity challenges of generative AI systems themselves, highlighting limitations in conventional frameworks and proposing governance-oriented mitigations such as model evaluation, human-in-the-loop oversight, GenAI-specific red teaming, and the structured dissemination of threat intelligence derived from GenAI-enabled security practices.

Relevance to practice

As generative AI systems rapidly integrate into business and IT environments, internal auditors, internal control specialists, and IT audit professionals should understand their cybersecurity implications. This article explores the intersection of generative AI and cybersecurity, providing insights into both opportunities and risks. By examining AI-driven offensive and defensive security applications, associated threats, and mitigation strategies, the article equips professionals with the knowledge to assess and manage AI-related cyber risks in organizations.

Keywords

Generative AI, cybersecurity, AI risk management, LLM security, AI governance, cyber threat intelligence, adversarial AI attacks, penetration testing, AI in offensive security, AI in cyber defense, AI red teaming, AI compliance in audit

1. Introduction

Generative AI differs from other AI models primarily in its ability to generate novel content rather than just analyzing or acting on existing data. Traditional AI models typically use specific data to solve specific problems and generate specific answers based on input data. In contrast, generative AI models, like large language models (LLMs), are capable of creating new and original content by mapping input information into a high-dimensional latent space and driving stochastic behavior to produce novel outputs even with

the same input stimuli (Corchado et al. 2023). LLM-based agents have demonstrated significant potential in attaining human-like intelligence by leveraging comprehensive training datasets and a substantial number of model parameters. These agents possess more comprehensive internal world knowledge compared to traditional reinforcement learning models, enabling more informed actions without specific domain training. Furthermore, LLM-based agents offer natural language interfaces, providing flexible and explainable

interactions with human operators (Wang et al. 2024). For the purposes of this article, the term ‘LLM’ will be used to refer to any generative AI model that accepts various forms of input and produces new content as output. The scope of this article is primarily focused on the application of generative AI – particularly LLMs – in the domain of cybersecurity. This article is intended to raise awareness of the capabilities, risks, and implications of LLMs in cybersecurity. It does not aim to provide guidance for malicious use but rather to inform cybersecurity, audit, and risk professionals about emerging threats and responsibilities.

In this article, “*offensive cybersecurity*” refers to proactive security testing methods such as penetration testing, red teaming and threat simulation, aimed at identifying vulnerabilities before malicious actors exploit them. In contrast, “*defensive cybersecurity*” encompasses technologies and processes focused on prevention, detection, response, and recovery from cyber threats.

This article is structured as follows: Section 2.1 explores the role of Generative AI in offensive cybersecurity, while Section 2.2 examines its applications in cyber defense. Section 3 highlights key risks associated with AI-driven security practices, followed by Section 4 discussing governance and mitigation strategies. Finally, the conclusion summarizes insights for internal audit, IT audit and internal control professionals.

2. Impact of Generative AI in cybersecurity domain

The relationship between Generative AI and cybersecurity can be outlined across four distinct categories (Gupta et al. 2023): 1.) Using Gen AI in cybersecurity offensive domain, 2.) Using Gen AI in cyber defense operations, 3.) Risks associated with Gen AI and 4.) Cybersecurity of Gen AI models.

2.1. Using Generative AI in cybersecurity offensive domain

Generative AI has significantly impacted offensive cybersecurity by increasing the sophistication and scale of cyber threats by allowing more complex and varied types of cyber-attacks (Palani et al. 2024). This dual nature of Generative AI highlights its role as a double-edged sword in cybersecurity. LLMs can be a valuable tool for cyber security professionals aiding in tasks such as penetration testing and developing security solutions (Al-Hawawreh et al. 2023). The use of LLMs has been shown to automate cyber-attacks effectively, with language models like ChatGPT able to generate executable attack code and script fragments (Iturbe et al. 2024). Generative AI-powered attacks achieve a 67% higher success rate and a 72% reduction in operational complexity, enhancing the effectiveness of cyber offensive operations by simulating sophisticated attack scenarios (Reddem 2024). Microsoft (Karamthulla et al. 2024) observed that threat actors like Forest Blizzard, Emerald Sleet, Crimson Sandstorm, Charcoal Typhoon, and

Salmon Typhoon are increasingly looking to AI, including Generative AI, to enhance their attacks. The analysis revealed that these LLM-enhanced attack campaigns correspond to several tactics described in the MITRE ATT&CK framework. During the reconnaissance phase, threat actors utilized LLMs to scan and collect victim-specific information, thereby tailoring their attacks more effectively. In the resource development phase, LLMs was applied to enhance malware capabilities and develop supporting infrastructure. For initial access, attackers launched targeted spear phishing campaigns, where LLMs generated highly convincing and context-specific content. Defense evasion was observed through the use of AI-generated obfuscation techniques to bypass detection mechanisms on compromised systems. Finally, during the collection phase, adversaries employed generative models to extract information from sensitive data repositories, often targeting high-profile individuals or critical thematic domains.

Traditional cybersecurity offensive testing methods, like red teaming and penetration testing, are often time and resource-intensive, necessitating the adoption of specialized tools and algorithms for improved efficiency. Integrating LLMs into the red team testing process offers new opportunities to enhance efficiency, precision, and cost-effectiveness by automating complex tasks, improving decision-making, and providing real-time insights during engagements (Patil et al. 2024). LLMs can significantly reduce the time required for each testing phase by rapidly processing extensive datasets and proposing tailored actions (Zaydi and Maleh 2024). It allows even inexperienced IT operations staff to execute tests by providing a more efficient and accessible approach to penetration testing, potentially reducing costs and increasing the frequency of security assessments for organizations, particularly small and medium enterprises that may lack the budget for professional security testing services (Valea and Oprişa 2020).

The use of LLMs in the threat intelligence phases of offensive security tests significantly enhances the accuracy and speed of information extraction and analysis. LLMs can automate the extraction and summarization of important information from large datasets, such as historical cyber incident reports, thereby improving the accuracy of threat intelligence and the ability to forecast future threats (Sufi 2024). Overall, LLMs provide deep insights that help offensive cybersecurity professionals respond more efficiently to emerging threats and risks (Hassanin and Moustafa 2024).

LLMs, particularly ChatGPT, have high potential to enhance cyberattacks done by individuals with entry-level skills (Yigit et al. 2024). Using generative AI in cybersecurity, particularly in Capture the Flag (CTF) exercises, has significant potential (Chamberlain and Casey 2024). LLMs can automate the generation of attack scenarios, provide personalized feedback, and simulate real-world threat actors, enhancing the realism and effectiveness of CTF exercises. When LLMs are properly fine-tuned and combined with prompt engineering techniques like Chain of Thought (CoT) and Optimization by PROMpting (OPRO), they can

effectively automate threat modelling, involving simulating attacks to identify vulnerabilities (Yang et al. 2024).

LLM agents are valuable tools for the reconnaissance phase of penetration tests (Temara 2023). In the reconnaissance phase, LLM agents generate detailed reports, enhancing initial information gathering (Hilario et al. 2024). That means that they provide insightful information, like technology stack, domain names, SSL/TLS configurations, ports and services used, that can be directly used for planning the next phase of a penetration test, offering meaningful insights that previously required multiple tools to obtain. During scanning, it automated test scenario generation, streamlining vulnerability detection. In exploitation, the LLM quickly responded to vulnerabilities, providing strategic exploitation options (Hilario et al. 2024). LLM-generated phishing and spam emails crafted to be more sophisticated and realistic, often bypassing the keyword-based and heuristic approaches that traditional spam detectors rely on. These spam detectors struggle with zero-shot and few-shot rephrased learning scenarios, where the emails are designed to evade detection by mimicking legitimate communication more closely (Afane et al. 2024).

Through neural machine translation, LLMs can effectively generate syntactically and semantically correct software exploits from natural language descriptions, though minor errors prevent full automation, indicating great potential (Liguori et al. 2021). MITRE ATT&CK top tactics, where LLMs are effective in generating successful executable code fragments, are “initial access” (TA0001), “defense evasion” (TA0005), and “discovery” (TA0007). The tactics “persistence” (TA0003), “privilege escalation” (TA0004), and “exfiltration” (TA0010) also showed satisfactory outcomes (Iturbe et al. 2024).

LLMs can effectively facilitate cyber offensive attacks, specifically generating viruses and polymorphic malwares (Gupta et al. 2023). They can be leveraged to generate code that targets CPU vulnerabilities, such as those that allow viruses to read kernel memory, thereby gaining control over the system. Additionally, LLMs can be leveraged to generate polymorphic malware, which is designed to alter its code with each execution to evade detection by traditional antivirus systems.

There are some early successful examples of LLM applications illustrating the potential of LLM-based automation to transform cybersecurity by reducing manual effort, enhancing accuracy, and enabling comprehensive threat assessment. PTHelper (Gracia and Sánchez-Macián 2024) streamlines the penetration testing process by automating transitions between phases, using modules for scanning, exploiting, natural language processing, and reporting, demonstrating effectiveness in both black-box and controlled environments. PentestGPT (Deng et al. 2023a) simplifies testing by guiding LLMs through micro-steps, reducing reliance on domain expertise, though expert oversight remains essential for accuracy. AutoAttacker (Xu et al. 2024) leverages GPT-4 for automating post-breach cyber-attack stages, excelling in lateral movement and credential gathering with modular components for planning and navigation. GAIL-PT (Chen et al. 2022) uses generative adversarial imitation learning to address the challenges of high-dimensional action spaces, integrating expert knowledge to improve decision-making in penetration testing.

Figure 1 illustrates how LLMs can be integrated into different phases of offensive cybersecurity operations, including penetration testing, red teaming, and threat intelligence-based simulations. The figure emphasizes

Figure 1. Key applications of LLMs across phases of cyber offensive security testing.

Penetration test and red teaming automation/guidance

- PTHelper- automates penetration test phases
- PentestGPT- AI guidance through penetration test steps
- Autoattacker- Automates post-breach attack stages
- PenHeal- AI-driven vulnerability remediation
- GAIL-PT - Generative adversarial imitation learning for pentesting

Reconnaissance & exploitation

- Extracting technology stack, domain and port information
- Automating scenario generation and vulnerability detection
- Enhancing phishing and spam sophistication

Adversarial Cyber Offensive Operations

- Generating polymorphic malware
- Creating AI-enhanced phishing e-mail
- Exploiting CPU/memory vulnerabilities

Capture the Flag Exercises

- Automating attack scenario generation
- Providing feedback to red team operators
- Simulating real-world threat actors

how LLMs not only streamline traditional workflows but also enable new capabilities, such as automated scenario generation, enhanced reconnaissance, and dynamic exploitation options. This visual supports the argument that LLMs can significantly improve the efficiency, scalability, and accessibility of offensive security testing.

2.2 Using Gen AI on cyber defense operations

In the context of cyber defense, LLMs excel in tasks such as threat detection, vulnerability analysis, and automated defense mechanisms (Zhou et al. 2024). They offer adaptive and intelligent technologies that can dynamically create and deploy actionable defense mechanisms, thereby increasing the efficiency of security operations. LLMs enhance cybersecurity by analyzing historical and real-time data to accurately predict future threats and vulnerabilities, enabling organizations to implement proactive security measures and strengthen their defenses (Metta et al. 2024). Furthermore, their capability to automate routine cyber operations tasks like threat analysis and incident response, allowing cybersecurity professionals to dedicate more time to strategic decision-making and complex investigations.

LLMs hold transformative potential in the field of cybersecurity defensive operations, offering significant advancements across various applications including threat intelligence, cybersecurity risk monitoring, vulnerability management, static malware analysis, dynamic debugging, anomaly detection and behavior analysis, web content security, phishing and spam detection, digital forensic, fuzz testing, program repairing, secure code generation, honeypots, and incident response and recovery (Zhang et al. 2025). Additionally, by being trained on frameworks like MITRE ATT&CK and D3FEND, LLMs can provide comprehensive insights into both attack techniques and corresponding defense procedures, facilitating a more robust cybersecurity posture. This dual capability of LLMs not only accelerates the detection and response to cyber threats but also empowers cybersecurity professionals to develop more sophisticated defense strategies (Alotaibi et al. 2024).

A recent comprehensive review by Ding et al. (2025) highlights that the integration of LLMs into cyber defense operations offers significant advancements in managing and enhancing cybersecurity posture. By analyzing extensive datasets, LLMs can extract valuable features and information, providing strategic recommendations to mitigate cyberattacks and effectively detect threats. Their application in security datasets allows for the generation of human-like text, which aids in threat and risk detection, facilitating rapid responses to potential threats. However, the successful deployment of LLMs necessitates access to comprehensive datasets, including security data, network traffic, and log files, which are crucial for accurate threat detection and risk mitigation. Organizations must exercise caution in several areas

before implementing LLMs in their cyber operations. Ensuring data privacy and protection is critical, given the large-scale datasets involved. Awareness of the potential vulnerabilities and threats that LLMs might introduce is essential to maintaining a robust cybersecurity posture. Compliance with regulatory requirements is also vital to avoid legal and operational challenges. Furthermore, LLMs should be seamlessly integrated with existing cybersecurity systems to maximize their effectiveness. Lastly, ethical and responsible use of LLMs is crucial to prevent misuse and maintain trust in their outputs, ensuring that these advanced tools contribute positively to cybersecurity efforts (Ding et al. 2025).

In the current cyber security landscape various real-world cybersecurity products are being used in the cybersecurity operations that leverage Generative AI to enhance security measures (Sai et al. 2024). Google Cloud Security AI Workbench and Microsoft Security Copilot are designed to enhance threat detection and response. SentinelOne Purple AI focuses on addressing emerging threats with advanced AI techniques. Talon Enterprise Browser integrates with Microsoft Azure OpenAI Service to provide enterprise-grade access to Generative AI tools like ChatGPT, enhancing data protection and productivity. SlashNext Generative Human AI defends against advanced threats such as business email compromise and financial fraud by mimicking human threat researchers. Recorded Future AI leverages over a decade of threat analysis data to provide real-time threat landscape analysis and improve analyst efficiency. SecurityScorecard integrates with OpenAI's GPT-4 to enhance its cybersecurity assessments, providing more comprehensive insights into potential vulnerabilities.

Figure 2 illustrates how LLMs and LLM integrated products can support cyber security operations. This figure illustrates also the increasing role of generative AI in operational cyber defense, showing how LLM-integrated tools support threat identification, detection, protection, response and recovery.

3. Risks associated with Gen AI

The risks associated with Generative AI can be classified under 3 categories particularly operational, technical and lastly systemic and strategic risk. These categories are derived from a synthesis of academic literature in this article.

As shown in Figure 3, the risks associated with generative AI span operational, technical, and systemic & strategic categories. Each type of risk requires a different set of mitigation strategies, as discussed in the following sections. While many of the risks associated with generative AI are systemic in nature, operational and technical risks, such as hallucinations, jailbreak vulnerabilities, or model theft, can manifest in both offensive and defensive cybersecurity operations contexts.

Figure 2. Gen AI on cyber defense operations.

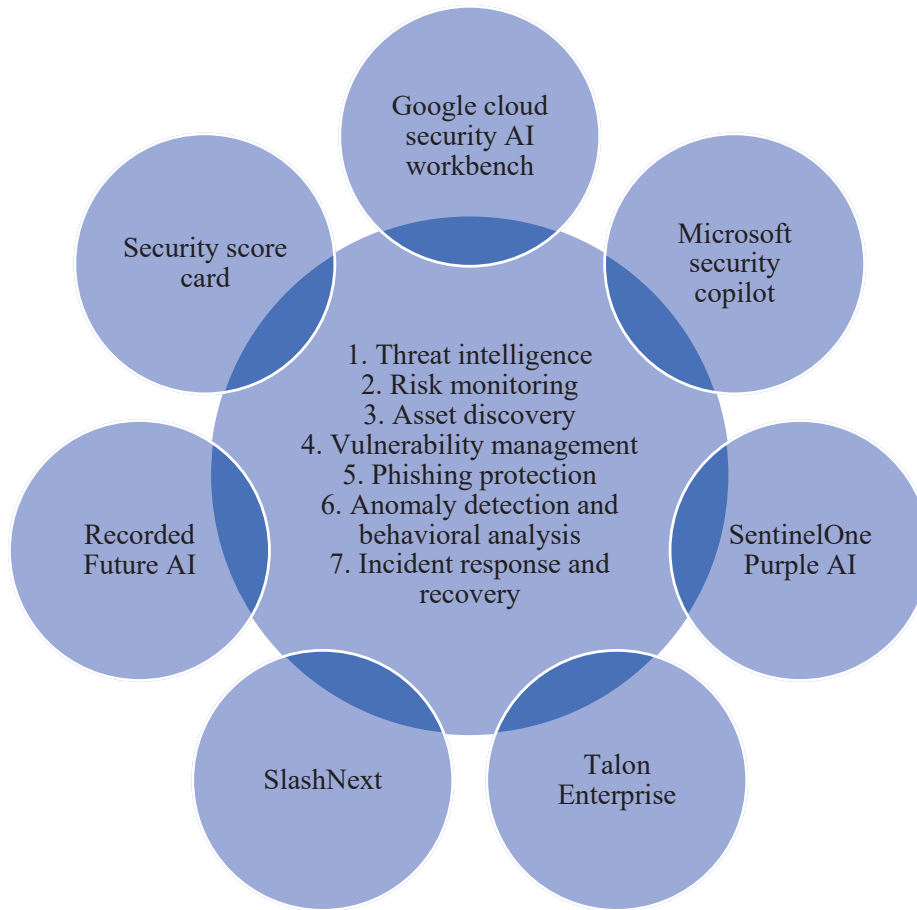
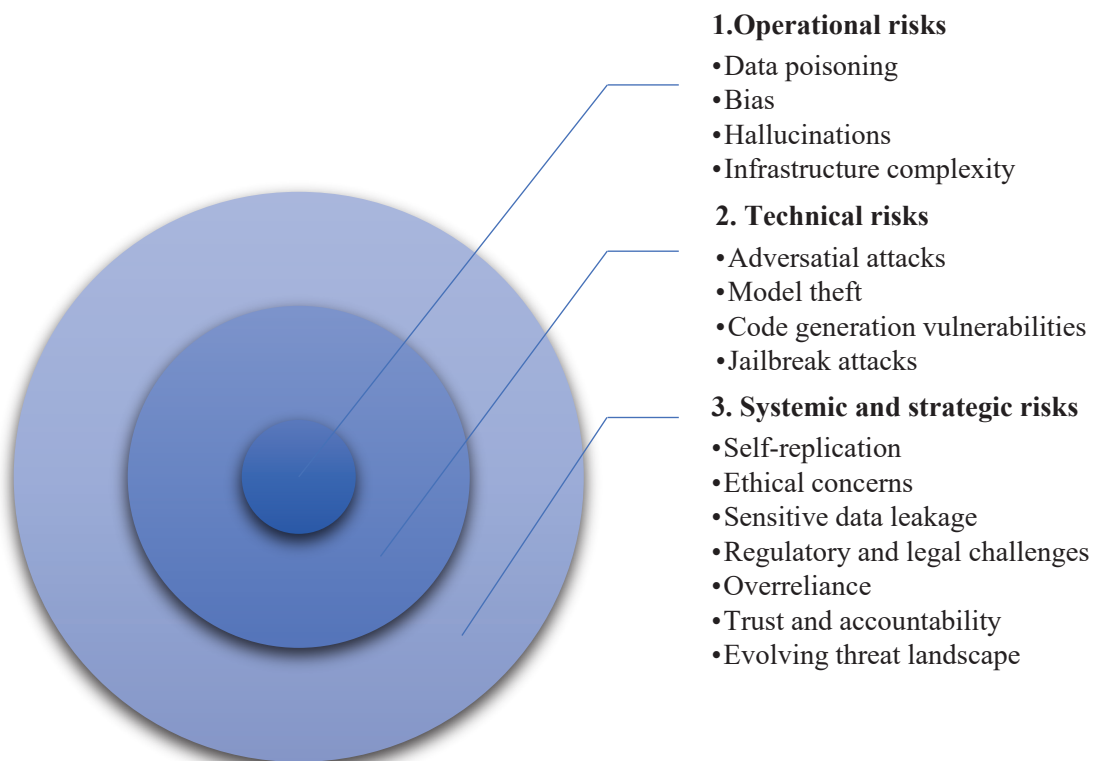


Figure 3. Risks associated with Generative AI.



3.1. Operational risks

While LLMs have significant potential in cybersecurity, particularly in threat intelligence process, they are not yet perfectly accurate due to hallucination where LLMs generate false information (Patsakis et al. 2024). LLM models can produce biased and unreliable content, and their increased consistency might make such content more credible and potentially more dangerous (Corchado et al. 2023). They may exhibit biases due to their training datasets, potentially leading to inaccurate or skewed recommendations (Zaydi and Maleh 2024). Another common operational concern is data poisoning, where malicious actors can corrupt the training datasets, leading to compromised AI performance and decision-making (Maryam et al. 2024). Attackers can manipulate the training data to cause the algorithm to make incorrect decisions, like misclassified cyber threats, misleading events/alerts and incorrect mitigations. The complexity of integrating AI systems with existing cybersecurity infrastructure poses a significant hurdle, as it requires substantial skills, resources and expertise (Jana et al. 2024).

3.2. Technical risks

First major challenge is the potential for adversarial attacks, where malicious actors can exploit LLM models by feeding them deceptive inputs to manipulate their outputs (Jana et al. 2024). Adversarial attacks, particularly prompt injection attacks, pose a significant risk, as they involve manipulating LLM inputs to generate unauthorized or harmful outputs, which can be exploited to simulate adversarial tactics and test system defenses against such manipulations (Taghavi and Feyzi 2024). A well-known example is Morris II, the first worm specifically designed to target Generative AI ecosystems using adversarial self-replicating prompts, highlighting a novel attack vector that exploits the interconnected nature of Generative AI-powered applications (Cohen et al. 2024). This example revealed the potential for adversarial attacks to compromise Generative AI systems, leading to malicious activities such as spamming, data exfiltration, and phishing.

Model theft is another critical risk, where unauthorized entities gain access to and replicate AI models, undermining proprietary technologies and security protocols (Maryam et al. 2024). Model theft may pose significant losses on organizations, reputational risks, including using the stolen model for malicious purposes (Taghavi and Feyzi 2024).

Although many developers have adopted AI technology in their workflows and generally find the code provided by AI to be usable and fairly accurate, there is caution regarding AI-generated code being insecure and inaccurate in coding and scripting practices (Sergeyuk et al. 2024). The use of AI-powered tools in cyber security operations may lead to the production of insecure code, posing significant risks (Oh et al. 2023). This can result in cyber security practitioners unknowingly incorporating insecure code into their systems, potentially compromising the integrity, effectiveness and security of their operations.

Despite the deployment of undisclosed defenses by service providers, LLM agents are vulnerable to jailbreak attacks, where malicious prompts can manipulate these models to bypass their safeguards and generate harmful or sensitive content (Deng et al. 2023b). Users could manipulate the LLM by crafting jailbreak prompts that bypass internal controls, leading the model to generate unauthorized or harmful information.

3.3. Systemic and strategic risks

An important concern regarding the systemic risks of LLMs is about their self-replication capability. AI systems driven by LLMs such as Meta's Llama31-70B-Instruct and Alibaba's Qwen25-72B-Instruct were demonstrated to be able to autonomously create separate copies of themselves, which could lead to the uncontrolled proliferation of AI systems, potentially forming independent networks that might act against its usage purpose (Pan et al. 2024). There are also concerns about the ethical aspects and potential misuse of AI-driven cyber offensive applications (Raman et al. 2024) and AI-generated content (Jana et al. 2024), which can be used to create convincing cyber-attacks or deepfakes, complicating the detection of genuine threats. Furthermore, the use of LLMs in cybersecurity processes poses significant data leakage risks due to their need for accessing sensitive system information, which can lead to unauthorized access, especially in cloud-hosted environments (Zaydi and Maleh 2024). Lastly, some regulatory challenges have been defined in using LLMs for cybersecurity (Sai et al. 2024), including the risk of intellectual property violations due to content generation similar to proprietary research, and the need for quality control and standardization to ensure consistent AI-generated advice. Additional challenges include defining data ownership, ensuring continuous monitoring and validation of AI performance, obtaining informed consent from users, maintaining interpretability and transparency of AI decision-making processes, and preventing over-reliance on LLM models, which could diminish human expertise.

The inherent black-box nature of LLMs presents significant challenges in understanding and controlling their operations, which raises critical concerns about transparency and accountability (Barman et al. 2024). Due to their complex and opaque internal workings, it is difficult for users and developers to predict or explain the outputs generated by these models. The lack of explainability in LLMs limits their utility as coding and scripting tools, which could hinder the understanding and mitigation of security risks during cyber operations (Khoury et al. 2023). This lack of explainability and transparency can make it also difficult to assess the reliability and accuracy of the AI's threat detection and exploit capabilities (Mohammed 2024).

Overreliance on content generated by LLMs poses significant risks, particularly due to the difficulty in detecting incorrect or misleading information produced by these models (Yao et al. 2024). LLMs are capable of

generating highly convincing text that can easily be mistaken for accurate cyber intelligence information, leading to incorrect actions and conclusions.

4. Cybersecurity of Gen AI models

The wide use of LLM based agents in the IT landscape has widened the exploit surface available to attackers. This expansion is driven by several risk factors outlined in Section 3 of this article, including vulnerabilities inherent to generative AI models, the increased feasibility of adversarial and jailbreak attacks, and the misuse of LLMs for generating potentially harmful or exploitable code. It is necessary to address cybersecurity risks specific to generative AI systems, on the top of traditional cybersecurity practices.

LLMs can amplify existing security risks and introduce new ones, emphasizing the need for a thorough understanding of the system's capabilities and applications. About the amplified cybersecurity risks, Microsoft has published its early lessons learned from its red teaming of 100 generative AI products (Bullwinkel et al. 2025). Cyber security professionals and LLM practitioners are advised to implement system-level mitigations, such as input sanitization, and model-level improvements, like instruction hierarchies, to manage and prioritize instructions effectively. Model-level evaluation is a critical AI governance infrastructure, providing insights into the safety and alignment of models, particularly about responsible training, responsible deployment, transparency and security (Shevlane et al. 2023). Microsoft's publication also warns that LLMs exposed to untrusted inputs may produce arbitrary outputs, including private information, emphasizing the necessity for robust input validation and data handling protocols. Additionally, the involvement of subject matter experts has been considered crucial for evaluating LLM outputs in specialized domains, including cyber security, where LLMs may not be reliable. Regular AI red teaming practices is recommended to enhance the communication of methods and findings, thereby improving the overall security posture.

A critical enabler of this is AI governance, which plays a key role in ensuring secure and ethical AI deployment. It addresses systemic concerns such as algorithmic bias, data privacy, transparency, and responsible use of AI technologies (Mohammed 2024). Within this context, model-level evaluation is essential for limiting the creation, deployment, and proliferation of generative AI systems that may pose risks to organizations (Shevlane et al. 2023). Such evaluations help identify whether a Gen AI model possesses potentially harmful capabilities or a tendency to apply these capabilities inappropriately.

Human oversight in Gen AI operations remains equally vital, since it involves assessing Gen AI safety questions that require emotional intelligence and understanding the full range of interactions users might have with Gen AI systems (Bullwinkel et al. 2025). Only human subject matter experts can evaluate model responses within specific domains and judge whether outputs are inappropriate, misleading, or harmful. Existing cybersecurity frameworks, such as NIST CSF 2.0, COBIT 2019, ISO 27001:2022, and the latest ISO 42001:2023, still exhibit significant gaps in addressing the multifaceted risks associated with LLMs, necessitating enhancements and the integration of human-expert-in-the-loop validation processes to ensure secure and compliant LLM integration (McIntosh et al. 2024). In this context, cybersecurity professionals in the fields of internal audit and IT audit can play a pivotal role by reviewing and validating LLM-generated outputs, particularly in high-risk or regulated environments, thereby reinforcing trust, accuracy, and accountability in AI-driven cybersecurity operations.

Furthermore, given that LLMs can amplify existing security risks and introduce new ones, Gen AI red teaming is a crucial practice for assessing the safety and security of Gen AI systems, as it pushes beyond model-level safety benchmarks by emulating real-world attacks against end-to-end systems (Bullwinkel et al. 2025).

Finally, aligning various cybersecurity efforts, including Gen AI red teaming, with real-world risks is indispensable, which necessitates the dissemination of insights and threat intelligence gathered from extensive cybersecurity practices (Bullwinkel et al. 2025).

Figure 4. Cybersecurity approach for Gen AI Models.

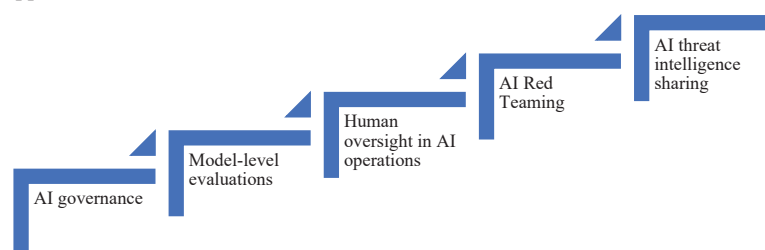


Figure 4 provides a structured overview of key cybersecurity practices tailored to generative AI systems, particularly LLMs. This approach reflects the broader message of this section: securing generative AI requires controls beyond conventional cybersecurity frameworks.

A new dimension in securing LLMs is the integration of LLMs into red teaming practices, such as an automated red teaming LLM agent that simulates adversarial conversations with LLMs, leveraging multiple adversarial prompting techniques, allowing for scalable

and efficient stress-testing of known vulnerabilities, thus freeing human testers to explore new risk areas (Pavlova et al. 2024). This approach enhances productivity and efficiency by automating prompt generation, conversion, and response scoring, allowing for extensive coverage of potential risks (Haider et al. 2024). However, manual red-teaming on Generative AI systems remains still crucial for capturing issues that automated methods might miss, particularly in complex, nuanced interactions (Bengio et al. 2025).

5. Conclusions

Generative AI, particularly LLMs, has rapidly emerged as a powerful tool in cybersecurity, benefiting both cyber defenders and adversaries. On one hand, cybersecurity professionals leverage LLMs to enhance penetration testing, red teaming, and threat intelligence-driven security tests, enabling faster, more sophisticated, and cost-effective offensive security operations. On the other hand, malicious actors exploit the same technology to automate cyberattacks, craft advanced phishing campaigns, and develop polymorphic malware, expanding the cyber threat landscape. The cybersecurity community maintains a balanced perspective on the adoption of LLMs, recognizing both their value in strengthening defense operations and the significant challenges, risks, and potential for misuse they introduce. By openly addressing both the offensive and defensive capabilities of generative AI, this article aims to equip professionals with the knowledge to anticipate threats, not to support their misuse. Responsible innovation and risk-informed governance remain essential.

As AI-driven cyber security applications evolve, so do the risks and regulatory challenges associated with their use. Generative AI introduces vulnerabilities such as model exploitation, adversarial attacks, jailbreak exploits,

and biased or unreliable outputs, which could undermine security efforts if not properly managed. While AI provides remarkable efficiencies, it also increases organizations' exploit surfaces, requiring new control frameworks and continuous risk assessment.

For internal auditors, IT auditors, and internal control professionals, generative AI is not just an IT concern but a governance and risk management issue. To mitigate the risks associated with generative AI, these professionals can play a key role by assessing whether appropriate AI governance frameworks are in place and integrating AI-specific risks into enterprise risk management and audit plans. These professionals should understand the implications of generative AI in cybersecurity, ensuring that organizations harness AI's benefits while mitigating its risks. By balancing innovation with security, they can contribute to the responsible adoption of AI, strengthen ethical AI governance, and ensure compliance with evolving regulatory standards. As generative AI continues to shape the cybersecurity domain, the key challenge will be ensuring AI remains an asset rather than a liability. By proactively addressing the risks and opportunities of AI in security, professionals across cybersecurity, audit, and internal control fields can play a pivotal role in securing the AI-driven future.

Future research opportunities

This article aimed to highlight the intersection between generative AI and cybersecurity, focusing on both opportunities and associated risks. Future research could further explore how audit, risk, and internal control functions can enhance the cybersecurity assurance of GenAI systems. This includes examining control frameworks, audit methodologies, and regulatory compliance strategies tailored to the unique characteristics of AI-based technologies.

■ **K. Orpak RE CISSP CCSP CISA CIA ISO27001LA CSX-F CDPO CFSA CCSA – Kunter**, Senior Supervision Officer – DORA TLPT / TIBER-EU Test Manager, Dutch Authority for the Financial Markets (AFM). PhD Researcher, Faculty of Economics and Business, University of Amsterdam. *This article has been written within the scope of his academic affiliation with the University of Amsterdam.*

The author confirms having no financial interests or conflicts of interest related to the subject matter or materials discussed in this article.

Acknowledgements

The author acknowledges the use of ChatGPT-4o, an advanced language model, to assist in the linguistic refinement and structural improvements of this manuscript. The tool was used solely for linguistic and structural refinement; all conceptual contributions, critical analysis, and findings are entirely the author's own.

References

- Afane K, Wei W, Mao Y, Farooq J, Chen J (2024) Next-generation phishing: how LLM agents empower cyber attackers. arXiv. <https://doi.org/10.1109/BigData62323.2024.10825018>
- Al-Hawawreh M, Aljuhani A, Jararweh Y (2023) ChatGPT for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing* 26: 3421–3436. <https://doi.org/10.1007/s10586-023-04124-5>
- Alotaibi L, Seher S, Mohammad N (2024) Cyberattacks using ChatGPT: exploring malicious content generation through prompt engineering. 2024 ASU Int Conf Emerg Technol Sustain Intell Syst (ICETISIS) 00: 1304–1311 (ICETISIS) 00: 1304–1311 (2024). <https://doi.org/10.1109/ICETISIS61505.2024.10459698>
- Barman D, Guo Z, Conlan O (2024) The dark side of language models: exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications* 16: 100545. <https://doi.org/10.1016/j.mlwa.2024.100545>
- Bengio Y, Mindermann S, Privitera D, Besiroglu T, Bommasani R, Casper S, Choi Y, Fox P, Garfinkel B, Goldfarb D, Heidari H, Ho A, Kapoor S, Khalatbari L, Longpre S, Manning S, Mavroudis V, Mazeika M, Michael J, Newman J, Ng KY, Okolo CT, Raji D, Sastry G, Seger E, Sheadas T, South T, Strubell E, Tramèr F, Velasco L, Wheeler N, Acemoglu D, Adekanmbi O, Dalrymple D, Dietterich TG, Felten EW, Fung P, Gourinchas P-O, Heintz F, Hinton G, Jennings N, Krause A, Leavy S, Liang P, Ludermir T, Marda V, Margetts H, McDermid J, Munga J, Narayanan A, Nelson A, Neppel C, Oh A, Ramchurn G, Russell S, Schaake M, Schölkopf B, Song D, Soto A, Tiedrich L, Varoquaux G, Yao A, Zhang Y-Q, Albalawi F, Alserkal M, Ajala O, Avrin G, Busch C, de Leon Ferreira de Carvalho ACP, Fox B, Gill AS, Hatip AH, Heikkilä J, Jolly G, Katzir Z, Kitano H, Krüger A, Johnson C, Khan SM, Lee KM, Ligt DV, Molchanovskiy O, Monti A, Mwamanzi N, Nemer M, Oliver N, Portillo JRL, Ravindran B, Rivera RP, Riza H, Rugege C, Seoighe C, Sheehan J, Sheikh H, Wong D, Zeng Y (2025) International AI safety report. arXiv. <https://doi.org/10.48550/arXiv.2501.17805>
- Bullwinkel B, Minnich A, Chawla S, Lopez G, Pouliot M, Maxwell W, de Gruyter J, Pratt K, Qi S, Chikanov N, Lutz R, Dheekonda RSR, Jagdagdorj B-E, Kim E, Song J, Hines K, Jones D, Severi G, Lundeen R, Vaughan S, Westerhoff V, Bryan P, Kumar RSS, Zunger Y, Kawaguchi C, Russinovich M (2025) Lessons from red teaming 100 generative AI products. arXiv. <https://doi.org/10.48550/arXiv.2501.07238>
- Chamberlain D, Casey E (2024) Capture the flag with ChatGPT: security testing with AI chatbots. *International Conference on Cyber Warfare and Security* 19: 43–54. <https://doi.org/10.34190/icwcs.19.1.2171>
- Chen J, Hu S, Zheng H, Xing C, Zhang G (2022) GAIL-PT: a generic intelligent penetration testing framework with generative adversarial imitation learning. arXiv. <https://doi.org/10.1016/j.cose.2022.103055>
- Cohen S, Bitton R, Nassi B (2024) Here comes the AI worm: unleashing zero-click worms that target GenAI-powered applications. arXiv. <https://doi.org/10.48550/arxiv.2403.02817>
- Corchado JM, Garcia SR, Núñez VJM, López FS, Chamoso P (2023) Generative artificial intelligence: fundamentals. *Advances in Distributed Computing and Artificial Intelligence Journal* 12(1): e31704. <https://doi.org/10.14201/adcaij.31704>
- De Gracia JC, Sánchez-Macián A (2024) PTHelper: an open source tool to support the penetration testing process. arXiv. <https://doi.org/10.48550/arxiv.2406.08242>
- Deng G, Liu Y, Mayoral-Vilches V, Liu P, Li Y, Xu Y, Zhang T, Liu Y, Pinzger M, Rass S (2023a) PentestGPT: an LLM-empowered automatic penetration testing tool. arXiv. <https://doi.org/10.48550/arxiv.2308.06782>
- Deng G, Liu Y, Li Y, Wang K, Zhang Y, Li Z, Wang H, Zhang T, Liu Y (2023b) Jailbreaker: automated jailbreak across multiple large language model chatbots. arXiv. <https://doi.org/10.14722/ndss.2024.24188>
- Ding W, Abdel-Basset M, Ali AM, Moustafa N (2025) Large language models for cyber resilience: a comprehensive review, challenges, and future perspectives. *Applied Soft Computing* 170: 112663. <https://doi.org/10.1016/j.asoc.2024.112663>
- Gupta M, Akiri C, Aryal K, Parker E, Praharaj L (2023) From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. *IEEE Access* 11: 80218–80245. <https://doi.org/10.1109/ACCESS.2023.3300381>
- Haider E, Perez-Becker D, Portet T, Madan P, Garg A, Ashfaq A, Majercak D, Wen W, Kim D, Yang Z, Zhang J, Sharma H, Bullwinkel B, Pouliot M, Minnich A, Chawla S, Herrera S, Warreth S, Engler M, Lopez G, Chikanov N, Dheekonda RSR, Jagdagdorj B-E, Lutz R, Lundeen R, Westerhoff T, Bryan P, Seifert C, Kumar RSS, Berkley A, Kessler A (2024) Phi-3 safety post-training: aligning language models with a “break-fix” cycle. arXiv. <https://doi.org/10.48550/arxiv.2407.13833>
- Hassanin M, Moustafa N (2024) A comprehensive overview of large language models (LLMs) for cyber defences: opportunities and directions. arXiv. <https://doi.org/10.48550/arxiv.2405.14487>
- Hilario E, Azam S, Sundaram J, Mohammed KI, Shanmugam B (2024) Generative AI for pentesting: the good, the bad, the ugly. *International Journal of Information Security* 1–23. <https://doi.org/10.1007/s10207-024-00835-x>
- Huang J, Zhu Q (2024) PenHeal: a two-stage LLM framework for automated pentesting and optimal remediation. arXiv. <https://doi.org/10.2139/ssrn.4941478>
- Iturbe E, Llorente-Vazquez O, Rego A, Rios E, Toledo N (2024) Unleashing offensive artificial intelligence: automated attack technique code generation. *Computers and Security* 147: 104077. <https://doi.org/10.1016/j.cose.2024.104077>
- Jana S, Biswas R, Banerjee C, Patra T, Pal M, Pal K (2024) Leveraging artificial intelligence for enhancing cybersecurity: a comprehensive review and analysis. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*: 173–183. <https://doi.org/10.48175/IJARSCT-19030>
- Karamthulla MJ, Tadamarr A, Tillu R, Muthusubramanian M (2024) Navigating the future: AI-driven project management in the digital era. *International Journal For Multidisciplinary Research* 6(2). <https://doi.org/10.36948/ijfmr.2024.v06i02.15295>

- Khoury R, Avila AR, Brunelle J, Camara BM (2023) How secure is code generated by ChatGPT? arXiv. <https://doi.org/10.1109/SMC53992.2023.10394237>
- Lanka P, Gupta K, Varol C (2024) Intelligent threat detection – AI-driven analysis of honeypot data to counter cyber threats. *Electronics* 13: 2465. <https://doi.org/10.3390/electronics13132465>
- Liguori P, Al-Hossami E, Orbinato V, Natella R, Shaikh S, Cotroneo D, Cukic B (2021) EVIL: exploiting software via natural language. arXiv. <https://doi.org/10.1109/ISSRE52982.2021.00042>
- Maryam R, Mahir RK, Natalie NS (2024) Navigating AI cybersecurity: evolving landscape and challenges. *Journal of Intelligent Learning Systems and Applications* 16: 155–174. <https://doi.org/10.4236/jilsa.2024.163010>
- McIntosh TR, Susnjak T, Liu T, Watters P, Nowrozy R, Halgamuge MN (2024) From COBIT to ISO 42001: evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *Computers and Security* 144: 103964. <https://doi.org/10.1016/j.cose.2024.103964>
- Metta S, Chang I, Parker J, Roman MP, Ehuang AF (2024) Generative AI in cybersecurity. arXiv. <https://doi.org/10.48550/arxiv.2405.01674>
- Mohammed B (2024) The impact of Artificial Intelligence on cyberspace security and market dynamics. *Brazilian Journal of Technology* 7(4): e74677. <https://doi.org/10.38152/bjtv7n4-019>
- Oh S, Lee K, Park S, Kim D, Kim H (2023) Poisoned ChatGPT finds work for idle hands: exploring developers' coding practices with insecure suggestions from poisoned AI models. arXiv. <https://doi.org/10.1109/SP54263.2024.00046>
- Palani K, Kethar J, Prasad S, Torremocha V (2024) Impact of AI and Generative AI in transforming Cybersecurity. *Journal of Student Research* 13(2). <https://doi.org/10.47611/jsrhs.v13i2.6710>
- Pan X, Dai J, Fan Y, Yang M (2024) Frontier AI systems have surpassed the self-replicating red line. arXiv. <https://doi.org/10.48550/arxiv.2412.12140>
- Patil M, Thakare D, Bhure A, Kaundanyapure S, Mune DA (2024) An AI-based approach for automating penetration testing. *International Journal For Research in Applied Science and Engineering Technology* 12: 5019–5028. <https://doi.org/10.22214/ijraset.2024.61113>
- Patsakis C, Casino F, Lykousas N (2024) Assessing LLMs in malicious code deobfuscation of real-world malware campaigns. *Expert Systems with Applications* 256: 124912. <https://doi.org/10.1016/j.eswa.2024.124912>
- Pavlova M, Brinkman E, Iyer K, Albiero V, Bitton J, Nguyen H, Li J, Ferrer CC, Evtimov I, Grattafiori A (2024) Automated red teaming with GOAT: the generative offensive agent tester. arXiv. <https://doi.org/10.48550/arxiv.2410.01606>
- Raman R, Calyam P, Achuthan K (2024) ChatGPT or Bard: who is a better certified ethical hacker? *Computers & Security* 140: 103804. <https://doi.org/10.1016/j.cose.2024.103804>
- Reddem P (2024) The rise of AI-powered cybercrime: A data-driven analysis of emerging threats. *IJFMR* 2582–2160. <https://doi.org/10.36948/ijfmr.2024.v06i06.30744>
- Sai S, Yashvardhan U, Chamola V, Sikdar B (2024) Generative AI for cyber security: analyzing the potential of ChatGPT, DALL-E and other models for enhancing the security space. *IEEE Access* PP (99): 1–1. <https://doi.org/10.1109/ACCESS.2024.3385107>
- Sergeyuk A, Golubev Y, Bryksin T, Ahmed I (2024) Using AI-based coding assistants in practice: state of affairs, perceptions, and ways forward. arXiv. <https://doi.org/10.2139/ssrn.4900362>
- Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J, Kokotajlo D, Marchal N, Anderljung M, Kolt N, Ho L, Siddarth D, Avin S, Hawkins W, Kim B, Gabriel I, Bolina V, Clark J, Bengio Y, Christiano P, Dafoe A (2023) Model evaluation for extreme risks. arXiv. <https://doi.org/10.48550/arxiv.2305.15324>
- Sufi F (2024) An innovative GPT-based open-source intelligence using historical cyber incident reports. *Natural Language Processing Journal* 7: 100074. <https://doi.org/10.1016/j.nlp.2024.100074>
- Taghavi SM, Feyzi F (2024) Using large language models to better detect and handle software vulnerabilities and cyber security threats. <https://doi.org/10.21203/rs.3.rs-4387414/v1>
- Temara S (2023) Maximizing penetration testing success with effective reconnaissance techniques using ChatGPT. arXiv. <https://doi.org/10.22541/au.167947026.68710739/v1>
- Valea O, Oprea C (2020) Towards pentesting automation using the Metasploit framework. 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP): 171–178. <https://doi.org/10.1109/ICCP51029.2020.9266234>
- Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J (2024) A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18: 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- Xu J, Stokes JW, McDonald G, Bai X, Marshall D, Wang S, Swaminathan A, Li Z (2024) AutoAttacker: a large language model guided system to implement automatic cyber-attacks. arXiv. <https://doi.org/10.48550/arxiv.2403.01038>
- Yang S, Yang S, Liu S, Nguyen D, Jang S, Abuadba A (2024) ThreatModeling-LLM: automating threat modeling using large language models for banking system. arXiv. <https://doi.org/10.48550/arxiv.2411.17058>
- Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y (2024) A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing* 4: 100211. <https://doi.org/10.1016/j.hcc.2024.100211>
- Yigit Y, Buchanan WJ, Tehrani MG, Maglaras L (2024) Review of generative AI methods in cybersecurity. arXiv. <https://doi.org/10.48550/arxiv.2403.08701>
- Zaydi M, Maleh Y (2024) Empowering red teams with generative AI: transforming penetration testing through adaptive intelligence. *EDPACS ahead-of-print*: 1–26. <https://doi.org/10.1080/07366981.2024.2439628>
- Zhang J, Bu H, Wen H, Liu Y (2025) When LLMs meet cybersecurity: a systematic literature review. *Cybersecurity* 8: 55. <https://doi.org/10.1186/s42400-025-00361-w>
- Zhou Y, Cheng G, Du K, Chen Z (2024) Toward intelligent and secure cloud: large language model empowered proactive defense. arXiv. <https://doi.org/10.48550/arxiv.2412.21051>