OPEN ACCESS
*World Journal of Multidisciplinary Studies*
ISSN (Online): 3049-0855 | Vol. 3, No. 10, Oct. 2026
Website: https://wasrpublication.com/index.php/wjms

# Generative Artificial Intelligence in Cybersecurity: Applications, Opportunities, Risks, Governance, and Future Directions

**Amna Hassan Ali Al-Shidi**[1*], and **Jasim Alhadhrami**[2]

[1]*IT Instructor, University of Buraimi, Al Buraimi, Oman*

[2]*Project Management Instructor, Al Buraimi Vocational College, Al Buraimi, Oman*.

Received: 17/08/2026 | Revised: 12/09/2026 | Accepted: 06/10/2026 | Published: 27/10/2026

**Abstract —** Generative Artificial Intelligence (GenAI), and especially Large Language Models (LLMs), is changing how cybersecurity work is done, because these systems can produce and reshape natural language, code, and other digital material. In this review, we bring together recent academic and institutional sources, concentrating on the period 2024–2026, to analyze how GenAI is applied in cybersecurity, what opportunities and risks it brings, how it is governed, and where research is still lacking. Reported uses include cyber threat intelligence, threat detection, vulnerability assessment and remediation, defense against phishing and social engineering, penetration testing, deception, security operations, incident response, and software security. GenAI is, however, a dual-use technology. Empirical work indicates that it can make phishing more effective and cheaper to run, help attackers exploit known vulnerabilities, and open new attack surfaces such as indirect prompt injection and the poisoning of training data. Our analysis suggests that GenAI is most valuable when it supports expert judgment instead of substituting for it. Its adoption is held back by prompt injection, hallucination, leakage of sensitive information, data and model poisoning, weak supply chains, excessive agency, and evaluation practices that differ from study to study. When we compare the main governance instruments—NIST guidance, the OWASP Top 10 for LLM applications, ENISA and CISA publications, the EU AI Act, and ISO/IEC 42001—we find that they agree in principle but are only loosely connected in practice. Drawing on these findings, we propose a four-stage Human–AI Collaboration Model (GenAI assistance, human validation, controlled action, and continuous feedback) that turns the evidence into a workable process for security teams. We close with priorities for standardized benchmarking, secure agentic systems, privacy-preserving deployment, and validation in real environments.

**Keywords:** Generative artificial intelligence; Large language models; Cybersecurity; Cyber threat intelligence; Prompt injection; Vulnerability assessment; Security operations; AI governance; Agentic AI; Human–AI collaboration.

## 1. Introduction

Today's security environments generate ever-growing amounts of data, face attack methods that evolve quickly, rely on widely distributed infrastructure, and continue to suffer from a lack of skilled security professionals. Under time pressure, security teams have to make sense of alerts, logs, vulnerability reports, threat-intelligence feeds, source code, and incident records. Rule-based tools and conventional machine-learning systems still play an important role, yet they are mostly built for classification, prediction, or the detection of anomalies. GenAI adds something different: it can create and reshape information, explain technical material, communicate with analysts in everyday language, and help with reasoning that unfolds over several steps.

LLMs sit at the center of this shift. Systematic reviews published recently indicate that research on LLMs for cybersecurity covers vulnerability detection, secure code generation, program repair, threat intelligence, anomaly detection, malware analysis, and tasks related to attacks. To give a sense of the scale, one systematic literature review analyzed over 300 works, 25 LLMs, and more than 10 downstream cybersecurity scenarios (Zhang, J. et al., 2025). Yigit et al. (2025) offer a complementary picture by

*Corresponding Author
**Amna Hassan Ali Al-Shidi***
Email: amna.h@uob.edu.om

This is an open access article under the CC BY-NC license

organizing GenAI methods in cybersecurity into defensive applications, threats, and governance issues, whereas Ferrag et al. (2025) concentrate on LLM applications and their weaknesses, discussing security use cases together with prompt injection, data poisoning, and adversarial instructions.

For security organizations, GenAI can condense threat reports, pull out indicators of compromise, link related pieces of security information, explain vulnerabilities, write defensive code, and support incident-response activities. The benefit is clearest where analysts need to turn large volumes of unstructured material into short, usable knowledge. In the area of cyber threat intelligence (CTI), for instance, Balasubramanian et al. (2025) describe GenAI being applied to security reports, blogs, social media, network traffic, malware samples, dark-web data, and threat-intelligence platforms.

The same technology also reshapes the threat landscape. Adversaries can rely on generative systems to write convincing social-engineering material, help build malicious code, automate reconnaissance and exploitation, and run deceptive campaigns at a larger scale. Schmitt and Flechais (2024) discuss what GenAI means for social engineering and phishing, and in a controlled experiment, Heiding et al. (2024) found that phishing emails written by LLMs drew far more clicks than generic phishing messages. Further empirical work shows that LLMs can automatically generate attack techniques for breach-and-attack simulation (Iturbe et al., 2024) and that LLM agents are able to exploit known vulnerabilities once they are given the public vulnerability descriptions (Fang et al., 2024). For these reasons, recent reviews describe GenAI as a double-edged technology that boosts defensive capability while also widening the options available to attackers (Ibrar et al., 2025).

A further set of concerns comes from AI applications as targets in their own right. Successive editions of the OWASP Top 10 for LLM applications list risks such as prompt injection, disclosure of sensitive information, supply-chain weaknesses, data and model poisoning, improper output handling, excessive agency, and misinformation (OWASP Foundation, 2023; OWASP GenAI Security Project, 2025, 2026). Greshake et al. (2023) showed that indirect prompt injection, where malicious instructions are hidden in content that an LLM-integrated application retrieves later, works against real-world systems, and Carlini et al. (2024) demonstrated that poisoning web-scale training data can be done cheaply. On the guidance side, the NIST Generative AI Profile stresses risk management across the AI lifecycle and the principles of trustworthy AI (NIST, 2024a), and it is accompanied by NIST recommendations on secure software development for generative AI and dual-use foundation models (NIST, 2024b).

For this reason, we approach GenAI in two ways: as a cybersecurity *capability* and as a security *object*. Our goal goes beyond listing applications; we aim to bring the evidence on benefits, limitations, threats, governance, and future research together in one analytical framework.

Interest in this topic has risen steeply over the last few years. To show this, we counted the publications on generative AI in cybersecurity indexed in Google Scholar for each period between 2014 and 2026 (Figure 1). The numbers stayed low, under roughly 100 publications per period, until 2021–2022. They then climbed to 612 in 2022–2023, after conversational LLMs became publicly available, and reached about 3,180, 8,000, and 13,500 publications in the following three periods. The 2025–2026 figure is incomplete, as it only covers publications indexed up to the search date.

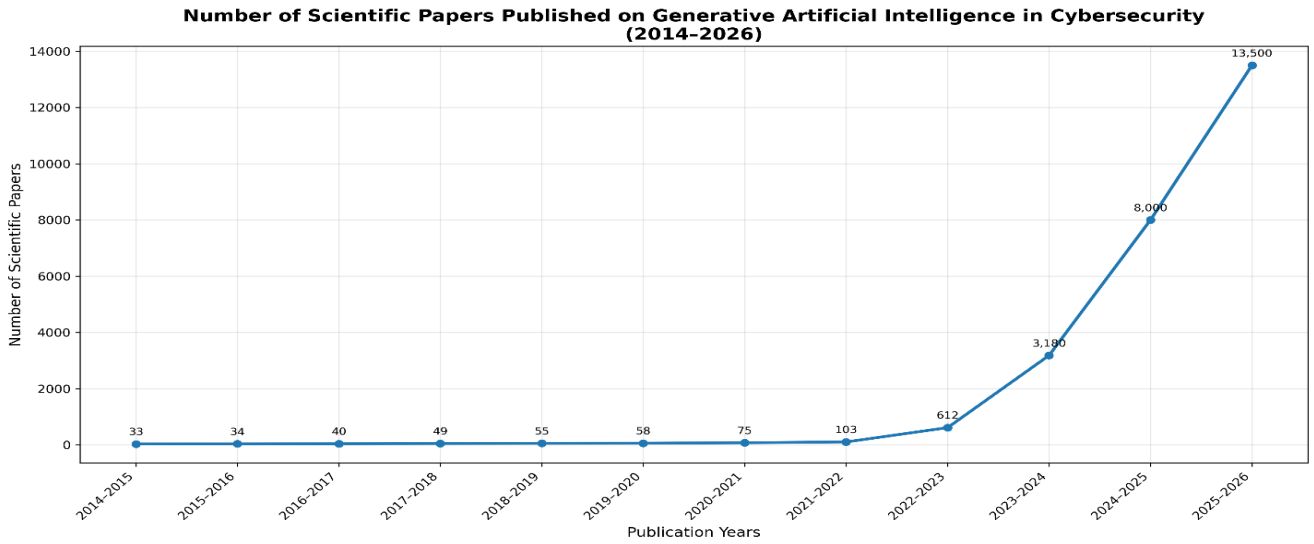



**Figure 1. Number of scientific publications on generative artificial intelligence in cybersecurity, 2014–2026.**

Source: Compiled by the authors from Google Scholar. The 2025–2026 value is partial.

### *1.1 Problem Statement*

Research on AI and cybersecurity has grown quickly, and the arrival of GenAI and LLM-based systems has produced a new body of evidence that is scattered and fragmented. Some studies examine defensive uses, some look at attacks against LLMs, and others investigate how attackers use LLMs. A large part of this literature is made up of secondary reviews, whereas primary empirical findings are spread across security, software-engineering, and machine-learning venues. Because models are developing so quickly, older results are also hard to generalize. What is missing is a consolidated review that links applications, opportunities, risks, governance requirements, and future research in a single analytical framework, and that separates claims made at the survey level from findings that have been demonstrated empirically.

### *1.2 Aim and Objectives*

This review aims to critically analyze the role of GenAI, and of LLMs in particular, in cybersecurity by drawing together recent academic and institutional literature on applications, opportunities, risks, governance requirements, and directions for future research. We seek to offer a joined-up understanding of GenAI both as a cybersecurity capability and as a possible source of security risk, with a focus on human oversight, secure-by-design development, and trustworthy adoption. To reach this aim, we set the following objectives:

- To analyze the main uses of GenAI in cybersecurity, namely cyber threat intelligence, threat detection, vulnerability assessment and remediation, malware analysis, defense against phishing and social engineering, penetration testing, deception, security operations, incident response, and software security.
- To determine the key opportunities that GenAI adoption offers, especially in terms of automation, speed, scalability, contextual support, knowledge transfer, and collaboration between humans and AI.
- To critically evaluate the risks linked to GenAI, such as prompt injection, hallucination, disclosure of sensitive information, data and model poisoning, improper output handling, supply-chain risks, excessive agency, misinformation, and malicious use.
- To compare the main governance instruments that apply to GenAI security, including NIST, OWASP, ENISA, CISA, the EU AI Act, and ISO/IEC 42001, and to highlight where they complement each other and where they leave gaps.
- To pinpoint weaknesses in current research with respect to evaluation, real-world validation, reliability, the security of agentic AI, privacy, explainability, benchmarking, governance, and regulatory harmonization.
- To put forward an integrated Human–AI Collaboration Model and a research agenda that encourage trustworthy, human-centered adoption of GenAI in cybersecurity.

### *1.3 Research Questions*

In line with these objectives, we address four research questions:

- **RQ1.** Which cybersecurity functions has GenAI been used for, and how mature is the evidence behind these uses?
- **RQ2.** Which security and reliability risks result from defenders' and attackers' use of GenAI, and from GenAI systems themselves?
- **RQ3.** To what extent do current governance frameworks and standards deal with these risks, and where are they insufficient?
- **RQ4.** Which research gaps are still open, and which directions should future studies prioritize?

### *1.4 Contributions*

The review offers five contributions. First, it combines secondary reviews with influential primary empirical studies and benchmarks, which partly overcomes the “review of reviews” problem that is common in this area. Second, it links each application area to representative evidence and rates how mature that evidence is (Table 4). Third, it compares the governance instruments that matter for GenAI security (Table 5). Fourth, it introduces a four-stage Human–AI Collaboration Model that puts the findings into practice (Section 4.7). Fifth, it sets out a research agenda in which every future direction is tied directly to a gap identified in the review.

### *1.5 Organization of the Paper*

The rest of the paper is structured as follows. Section 2 presents the literature review, covering secondary reviews as well as primary empirical studies. Section 3 explains the review methodology. Section 4 reports and discusses the results and presents the Human–AI Collaboration Model. Section 5 sets out the research gaps, Section 6 describes directions for future research, and Section 7 provides the conclusion.

## 2. Literature Review

### *2.1 GenAI Applications and LLMs in Cybersecurity*

The recent literature indicates that GenAI research in cybersecurity has grown along several connected lines: GenAI applications and vulnerabilities, security operations, threat intelligence, security and privacy, vulnerability detection, governance, and autonomous and agentic systems. Taken together, the studies point to the increasing defensive value of GenAI, but also to technical, operational, and governance constraints that still slow down its adoption.

Ferrag et al. (2025) carried out a comprehensive review of GenAI in cybersecurity that addressed applications, vulnerabilities, datasets, model evaluation, and mitigation

strategies; they noted, however, that evaluation practices are inconsistent, which makes it hard to compare studies. Through a systematic literature review of more than 300 works, 25 LLMs, and over 10 downstream scenarios, Zhang, J. et al. (2025) mapped how wide LLM research in cybersecurity has become, yet they also observed that many systems are still prototypes and have not been validated through long-term operational use. Another systematic literature review confirmed this breadth but pointed out that most work still centers on detection and analysis tasks instead of complete operational systems (Xu, H. et al., 2025). Uddin et al. (2025) concentrated on threat intelligence and security operations, reviewing uses in automation, anomaly detection, and incident response, and concluded that reliability and hallucination remain open obstacles for safety-critical use.

### *2.2 Threat Intelligence and the Dual Use of GenAI*

A number of studies deal specifically with threat intelligence and with the dual-use character of GenAI. Using a review combined with case-study analysis of security reports, blogs, social media, and dark-web data, Balasubramanian et al. (2025) investigated GenAI for CTI and found that depending heavily on noisy, unstructured open-source data makes the extracted intelligence less consistent. Ibrar et al. (2025) described GenAI as a double-edged sword in the cyber threat landscape and weighed its defensive benefits against its offensive misuse, though they did not go as far as proposing specific technical countermeasures. Zhou et al. (2025) offered a survey of LLM security risks covering bias, disinformation, ethics, misuse, attacks, and privacy; given this wide coverage, each attack vector is discussed only briefly. Das et al. (2025) reviewed security and privacy challenges alongside new attack and defense mechanisms and observed that the suggested defenses have not been widely tested in real-world settings. Zhang, R. et al. (2025) reached comparable conclusions in their survey of LLM safety, security, and privacy.

### *2.3 Large-Scale Cybersecurity and Vulnerability Management*

Another strand of work looks at LLMs in large-scale environments and in vulnerability management. In their review of LLM applications and challenges for Big Data cybersecurity, Karras et al. (2025) pointed to the computational cost of deploying such systems at scale as a major obstacle. Vadisetty and Polamarasetti (2024) put forward a GenAI-driven distributed cybersecurity framework for AI-integrated global Big Data systems, which again underlines the importance of scalability. Sheng et al. (2025) surveyed LLM-based vulnerability detection in terms of techniques, datasets, benchmarks, and evaluation metrics, and stressed that the lack of standardized benchmarks makes results difficult to reproduce. Mirtaheri et al. (2026) built on this work by developing a systematic taxonomy of AI-powered vulnerability assessment and risk management, grouping methods by the stage of the vulnerability-management lifecycle they serve. A more practical perspective comes from Wang et al. (2025), who studied a team-based, LLM-supported vulnerability remediation process in a large organization. The approach encouraged user engagement, but the generic guidance produced by the LLM was less helpful for simple issues and raised worries that users might depend too much on automated suggestions.

### *2.4 Social Engineering, Threat Detection, Governance, and Knowledge Representation*

Schmitt and Flechais (2024) explored how GenAI is used in social engineering and phishing, but their work was mostly conceptual and offered little empirical measurement of how successful such attacks are. In a survey on the roles of LLMs in cyber threat detection, Chen et al. (2024) reported that these models can still produce false positives and are not well grounded in production data. McIntosh et al. (2024) assessed frameworks from COBIT to ISO/IEC 42001 for governing the commercialization of LLMs and concluded that current frameworks are mostly qualitative and are not linked to measurable technical controls.

Hu et al. (2024) introduced an LLM-assisted method for building cyber threat intelligence knowledge graphs; the method, however, depends on specific ontologies, and it has not yet been tested on other threat domains. Sai et al. (2024) explored what ChatGPT, DALL-E, and other generative models could offer cybersecurity using exploratory case analysis, without a rigorous quantitative evaluation.

### *2.5 Offensive and Defensive Capabilities*

Other work addresses offensive and defensive capabilities in a more direct way. Iturbe et al. (2024) built an AI-powered mechanism that automatically produces attack-technique code for breach-and-attack simulation; this raises a dual-use concern, since real attackers could misuse the same mechanism. Shenoy and Mbaziira (2024) reviewed prompt-engineering techniques for cyber defense, although their study is descriptive and does not test different prompting strategies against each other empirically. Johnson et al. (2024) developed a modular generative honeypot shell for dynamic deception, which still needs to be tested against more advanced adversaries.

### *2.6 Governance, Agentic Systems, and Emerging GenAI Capabilities*

More recent research has pushed the literature toward governance, agentic systems, and wider GenAI capabilities. Following PRISMA-ScR guidance, Jangirala et al. (2026) carried out a scoping review of LLM threats, defenses, and emerging applications and found an ongoing lack of standardized ethical and regulatory guidance for deployment. Xu, Y. et al. (2026) surveyed the two sides of LLM agent security, namely the security of the agents

themselves and their use to strengthen cybersecurity, and showed that coordination among multiple agents creates new attack surfaces that are not yet well understood. Chowa et al. (2026) reviewed LLMs acting as autonomous agents and tool users and concluded that execution safety and reliable reasoning are still open problems for operational use. Yigit et al. (2025) reviewed GenAI methods, applications, and open challenges in cybersecurity; because the review is so wide-ranging, each application area is covered only in fairly general terms.

Two further systematic reviews support these conclusions. Moia et al. (2026) investigated threats and mitigations that affect real-world LLM-based systems, while Alqahtani and Kumar (2026) brought together emerging threats, defensive capabilities, and security evaluation frameworks for cybersecurity intelligence. Both reviews show that the field is shifting toward combined treatments of threats, defenses, and governance, which matches the thematic structure we follow here. Table 1 summarizes the research problems, methods, and weaknesses of the main secondary and applied studies covered in Sections 2.1–2.6.

**Table 1. Summary of reviewed secondary and applied studies**

| Year | Author(s) | Problem of Study | Method | Main Weakness |
|---|---|---|---|---|
| 2025 | Ferrag et al. | Uses of GenAI in cybersecurity and its vulnerabilities | Broad review covering applications, vulnerabilities, datasets, evaluation, and mitigation | Varying evaluation practices make studies hard to compare |
| 2025 | Zhang, J. et al. | Scope and breadth of LLM research for cybersecurity | Systematic literature review covering 300+ works, 25 LLMs, and 10+ scenarios | Most systems are still prototypes, not validated deployments |
| 2025 | Uddin et al. | Role of GenAI in threat intelligence and security operations | Review covering CTI, automation, anomaly detection, and response | Reliability and hallucination still unsolved for safety-critical use |
| 2025 | Balasubramanian et al. | Use of GenAI for cyber threat intelligence | Review plus case-study analysis of reports, blogs, social media, and dark-web data | Noisy, unstructured open-source data lowers consistency |
| 2025 | Ibrar et al. | GenAI as a double-edged technology | Conceptual comparison of defensive benefits and offensive misuse | Does not propose specific technical countermeasures |
| 2025 | Zhou et al. | Security issues in the use of LLMs | Survey covering bias, disinformation, ethics, misuse, attacks, and privacy | Wide scope; individual attack vectors covered only briefly |
| 2025 | Das et al. | Security and privacy of LLMs | Survey of challenges and of attack and defense mechanisms | Defenses not widely validated in real-world settings |
| 2025 | Karras et al. | LLMs for cybersecurity in Big Data settings | Review of applications and challenges in Big Data security | Large-scale deployment is computationally costly |
| 2025 | Sheng et al. | Vulnerability detection using LLMs | Review of techniques, datasets, benchmarks, and metrics | No standardized benchmarks, which limits reproducibility |
| 2025 | Wang et al. | LLM-supported collaborative vulnerability remediation | Field study of a team-based remediation process supported by LLMs | Generic advice less helpful for simple issues; risk of over-reliance |
| 2025 | Yigit et al. | Methods of GenAI in cybersecurity | Review of methods, applications, and open challenges | Wide scope; application areas discussed at a general level |
| 2024 | Schmitt & Flechais | GenAI-enabled digital deception | Analysis of the role of GenAI in social engineering and phishing | Mainly conceptual; little measurement of attack success |

| Year | Author(s) | Problem of Study | Method | Main Weakness |
|---|---|---|---|---|
| 2024 | Chen et al. | LLMs in cyber threat detection | Survey of the roles LLMs play in threat detection | False positives; weak grounding in production data |
| 2024 | McIntosh et al. | Governance and compliance of LLMs | Assessment of frameworks from COBIT to ISO/IEC 42001 | Frameworks are qualitative and not linked to measurable controls |
| 2024 | Hu et al. | Building CTI knowledge graphs | Knowledge-graph construction assisted by LLMs (LLM-TIKG) | Restricted to specific ontologies; generalization not tested |
| 2024 | Sai et al. | Generative AI tools for cybersecurity | Exploratory analysis of ChatGPT, DALL-E, and other models | Lacks rigorous quantitative evaluation |
| 2024 | Iturbe et al. | Generating attack techniques with offensive AI | Automatic generation of attack code for breach-and-attack simulation | Dual-use risk if real attackers misuse it |
| 2024 | Shenoy & Mbaziira | Prompt engineering for cyber defense | Extended review of prompt-engineering techniques | Descriptive; strategies not compared empirically |
| 2024 | Johnson et al. | Honeypots and deception driven by LLMs | Design of a modular generative honeypot shell | Not yet tested against advanced adversaries |
| 2026 | Jangirala et al. | Threats, defenses, and regulation of LLMs | Scoping review following PRISMA-ScR | Lack of standardized ethical and regulatory guidance |
| 2026 | Xu, Y. et al. | Dual nature of LLM agent security | Survey of agent self-security and agent-empowered cybersecurity | Multi-agent coordination introduces poorly understood attack surfaces |
| 2026 | Chowa et al. | Autonomy and tool use in LLMs | Review of LLMs acting as autonomous agents and tool users | Execution safety and reasoning reliability still unresolved |

### *2.7 Primary Empirical Studies and Benchmarks*

Since most of the studies discussed so far are secondary reviews, this subsection focuses on influential primary studies and benchmarks that offer direct empirical evidence (Table 2). We located them through backward citation searching (Section 3.3) and included them because the reviews above repeatedly cite them as foundational evidence.

Regarding software security, Pearce et al. (2022) tested code produced by GitHub Copilot in 89 scenarios linked to high-risk software weaknesses and reported that about 40% of the 1,689 generated programs were vulnerable, which shows that AI-assisted coding can spread insecure patterns. Turning to attacks on LLM-integrated applications, Greshake et al. (2023) introduced the concept of indirect prompt injection and showed that malicious instructions placed in retrieved data can steer model behavior, leading to data theft and other compromises in deployed systems. Carlini et al. (2024) demonstrated that poisoning web-scale training datasets is realistic, because attackers can take control of a small portion of widely used datasets at low cost.

From the offensive perspective, Heiding et al. (2024) ran a phishing experiment with 112 participants. Emails generated by GPT-4 reached click-through rates of 30–44%, against 19–28% for generic control emails, and when LLM generation was combined with human-designed persuasion principles, rates rose to as much as 81%. The same study also showed that LLMs were able to recognize malicious intent in phishing emails, at times doing better than the human participants. Fang et al. (2024) found that an agent based on GPT-4 was able to exploit 87% of 15 real-world one-day vulnerabilities when it received the CVE description, but only 7% without it; the small sample, however, limits how far these results can be generalized. For defensive assessment, Deng et al. (2024) built PentestGPT, a modular penetration-testing tool driven by LLMs, which completed considerably more tasks than a baseline LLM, although the authors reported problems with keeping track of context over long, multi-step tasks.

A related body of work has focused on building benchmarks. CyberSecEval measures how likely LLMs are to produce insecure code and to go along with requests to

help with cyberattacks (Bhatt et al., 2023). CTIBench tests LLMs on threat-intelligence tasks, for example mapping vulnerabilities to weaknesses, predicting severity scores, and attributing threat actors (Alam et al., 2024). Cybench assesses LLM agents on 40 professional-level capture-the-flag tasks taken from four competitions (Zhang, A. K. et al., 2025). These benchmarks move the field closer to the standardized evaluation that several reviews have asked for, even though they do not yet share common tasks and metrics.

**Table 2. Primary empirical studies and benchmarks**

| Year | Author(s) | Study Type | Focus | Key Finding | Limitation |
|---|---|---|---|---|---|
| 2022 | Pearce et al. | Empirical evaluation | Security of code written by AI | About 40% of 1,689 generated programs had vulnerabilities | One tool and model version; models have changed since |
| 2023 | Greshake et al. | Attack demonstration | Indirect prompt injection | Retrieved content can take over LLM-integrated applications | No systematic evaluation of defenses |
| 2023 | Bhatt et al. | Benchmark | Insecure code and compliance with attack requests | LLMs often propose insecure code | Preprint; test cases defined by developers |
| 2024 | Carlini et al. | Empirical attack | Poisoning of training data | Poisoning web-scale datasets at low cost is feasible | Downstream effect differs by model and task |
| 2024 | Heiding et al. | Human-subject experiment (n = 112) | Phishing generated by LLMs | GPT-4 emails reached 30–44% click-through vs. 19–28% for controls | Small sample and limited population |
| 2024 | Fang et al. | Agent evaluation | Exploiting one-day vulnerabilities | 87% of 15 CVEs exploited when the description was provided | Small set of vulnerabilities; preprint |
| 2024 | Deng et al. | Tool design and evaluation | Automated penetration testing | Modular design raises task completion | Loses context in long, multi-step tasks |
| 2024 | Alam et al. | Benchmark | Cyber threat intelligence tasks | Benchmark covering multiple CTI tasks for LLMs | Restricted to predefined task formats |
| 2025 | Zhang, A. K. et al. | Benchmark | Offensive capability of agents | 40 capture-the-flag tasks at professional level | CTF tasks may not represent enterprise environments |

## 3. Research Methodology

### *3.1 Research Approach*

We used a structured thematic literature review to synthesize research on the applications, opportunities, risks, and governance of GenAI in cybersecurity. The main focus was on academic and institutional literature from 2024 to 2026, with older sources included when they offered foundational empirical evidence or methodological guidance. The review does not apply a complete PRISMA systematic-review protocol; instead, it follows the PRISMA 2020 reporting principles to make the identification, screening, and selection of literature transparent (Page et al., 2021). The thematic design allows the paper to move from the existing evidence toward a joined-up discussion of trustworthy, human-centered GenAI adoption, ending with the Human–AI Collaboration Model in Section 4.7.

### *3.2 Search Strategy and Data Sources*

We searched Scopus, Web of Science, IEEE Xplore, the ACM Digital Library, and Google Scholar in a structured way. Two sets of terms were joined with Boolean operators: (“generative artificial intelligence” OR “generative AI” OR “large language model*” OR LLM OR ChatGPT) AND (cybersecurity OR “cyber security” OR “threat intelligence” OR “vulnerability detection” OR phishing OR “prompt injection” OR “LLM security” OR “AI governance”). Institutional guidance was obtained directly from the websites of NIST, OWASP, ENISA, CISA, the UK National Cyber Security Centre, the European Union, ISO, and MITRE. Table 3 lists the eligibility criteria.

**Table 3. Eligibility criteria**

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Topic | Applications, risks, or governance of GenAI or LLMs in cybersecurity | Only non-generative machine learning; no clear link to cybersecurity |
| Publication type | Peer-reviewed journal and conference papers, plus institutional reports or standards | Editorials, opinion articles, and non-scholarly web material |
| Period | 2024–2026 as the main focus; older foundational empirical studies found via citation searching | Earlier versions of a work that have been superseded |
| Preprints | Highly cited preprints with primary evidence not available elsewhere | Preprints that duplicate peer-reviewed versions |
| Language | English | Sources not in English |
| Evidence | Studies offering substantial evidence on applications, dual-use risks, or governance | Studies lacking enough methodological detail |

### *3.3 Study Selection*

In the first step, records were screened on their titles and abstracts, and those that passed were then read in full and checked against the criteria in Table 3. We then used backward citation searching (snowballing) on the included reviews to find primary studies and benchmarks that they cited often; this added several foundational empirical works from 2022–2024 (Section 2.7). The final set contains 49 sources: 22 secondary reviews or surveys, 15 primary empirical, design, or analytical studies (benchmarks included), 11 institutional frameworks or standards, and one methodological reference.

### *3.4 Data Analysis*

We analyzed the selected literature thematically across five linked areas: defensive applications, operational opportunities, security and reliability risks, governance requirements, and emerging research gaps. LLMs received particular attention because they are well suited to text-heavy security tasks, workflows that involve analysts directly, and tool-connected and agentic systems. We also classified each source as (a) a secondary review or survey, (b) a primary empirical, design, or analytical study, or (c) institutional guidance. This allowed us to separate claims made in surveys from findings shown empirically, and it forms the basis of the evidence-maturity assessment in Table 4.

### *3.5 Validity and Reliability*

To keep the analysis consistent, we grouped the evidence by recurring themes and compared studies according to their research problem, method, and reported weaknesses (Tables 1 and 2). The bibliographic details of every source were checked against publisher records and DOI metadata. Separating secondary from primary evidence also lowers the risk of overstating claims that appear in many reviews but rest on little original data.

### *3.6 Limitations of the Review*

The review is subject to several limitations. To begin with, it is based on published academic and institutional literature, so unpublished industry practice, proprietary security research, and classified threat intelligence may be under-represented. In addition, much of the corpus is made up of reviews or surveys and not primary empirical studies; Section 2.7 brings in key primary evidence, but the synthesis still partly resembles a review of reviews. Furthermore, even though we gave priority to literature from 2024–2026, GenAI changes so quickly that newer models and techniques may overtake some findings. The review also does not use a full systematic-review protocol, such as screening by two independent reviewers or a formal risk-of-bias assessment, which reduces reproducibility. Lastly, the database coverage is broad but not complete, and limiting the review to English-language sources may have left out relevant regional studies.

## 4. Results and Discussion

Drawing on the literature presented in Section 2, this section brings the evidence together under six connected themes. Section 4.7 then introduces an integrated framework that links these themes into a practical workflow for cybersecurity organizations.

### *4.1 Defensive Applications (RQ1)*

According to the literature, GenAI and LLMs can assist with many security functions, such as cyber threat intelligence, threat detection, vulnerability assessment and remediation, secure software development, phishing defense, penetration testing, deception, security operations, and incident response. The strength of the evidence, though, differs a

great deal from one function to another (Table 4). Case studies, field studies, tools, and benchmarks support work on threat intelligence, vulnerability management, and penetration testing, while security operations, incident response, and malware analysis are backed mostly by discussion in surveys. In every area, evidence from long-term operational deployment is still limited.

**Table 4. Application areas, representative evidence, and evidence maturity**

| Application Area | Representative Evidence | Evidence Maturity |
|---|---|---|
| Cyber threat intelligence | Balasubramanian et al. (2025); Hu et al. (2024); Alam et al. (2024) | Moderate: case studies, prototype, benchmark |
| Threat and anomaly detection | Chen et al. (2024); Uddin et al. (2025); Karras et al. (2025) | Low–moderate: mainly surveys |
| Vulnerability detection and remediation | Sheng et al. (2025); Mirtaheri et al. (2026); Wang et al. (2025) | Moderate: benchmarks and one field study |
| Secure software development | Pearce et al. (2022); Bhatt et al. (2023) | Moderate: empirical, but focused on risks |
| Phishing and social-engineering defense | Heiding et al. (2024); Schmitt & Flechais (2024) | Low–moderate: one controlled experiment |
| Penetration testing and red teaming | Deng et al. (2024); Iturbe et al. (2024); Zhang, A. K. et al. (2025) | Moderate: tools and benchmarks |
| Deception and honeypots | Johnson et al. (2024) | Low: prototype |
| Malware analysis | Zhang, J. et al. (2025); Xu, H. et al. (2025) | Low: survey-level only |
| Security operations and incident response | Uddin et al. (2025); Shenoy & Mbaziira (2024) | Low: mainly conceptual |

### *4.2 Operational Opportunities*

The reviewed studies mostly present GenAI as a way to augment human work. It can summarize reports, extract indicators of compromise, connect pieces of security information, explain vulnerabilities, produce defensive code, and support workflows with several steps (Uddin et al., 2025; Balasubramanian et al., 2025). The biggest opportunities are lighter analyst workloads, faster triage, and passing specialist knowledge on to less experienced staff. Even so, real operational value depends on outputs being reliable, on proper integration with security infrastructure, and on validation by humans. The field study by Wang et al. (2025) shows this clearly: LLM support increased engagement in vulnerability remediation, yet its generic advice was of little help for simple issues and created a risk of over-reliance. In the same way, PentestGPT completed more tasks but had trouble keeping context over long tasks (Deng et al., 2024).

### *4.3 Security and Reliability Risks (RQ2)*

Across the literature, the same group of risks keeps appearing as a barrier to safe adoption. Prompt injection stands out as the main risk at the application level, since direct and indirect injections can override system instructions and lead to data exfiltration or unauthorized actions (Greshake et al., 2023; OWASP GenAI Security Project, 2025). Data and model poisoning endangers model integrity starting from the training stage, and Carlini et al. (2024) demonstrated that poisoning web-scale datasets is affordable for attackers. Insecure output and code generation is another issue, because code produced by AI often includes vulnerabilities (Pearce et al., 2022; Bhatt et al., 2023). Hallucination and reliability are still unsolved problems in safety-critical contexts (Uddin et al., 2025; Chen et al., 2024), and sensitive information disclosure and privacy leakage come up again and again in surveys on security and privacy (Das et al., 2025; Zhang, R. et al., 2025). Lastly, the OWASP lists give weight to supply-chain weaknesses and excessive agency; in the 2026 edition, excessive agency is placed third, which reflects the spread of tool-connected systems (OWASP GenAI Security Project, 2026). In security-critical environments, these risks are especially serious, because wrong or manipulated outputs can carry over into later decisions or automated actions.

### *4.4 Offensive and Dual-Use Capabilities (RQ2)*

GenAI is dual-use by nature. Earlier conceptual analyses (Schmitt & Flechais, 2024; Ibrar et al., 2025) are now supported by primary evidence: phishing emails generated by LLMs receive more clicks than generic phishing messages (Heiding et al., 2024), LLM agents can exploit known vulnerabilities when they have access to public descriptions (Fang et al., 2024), and code for attack techniques can be produced automatically (Iturbe et al., 2024). Taken together, these results suggest that GenAI reduces the cost and the level of skill needed for several stages of an attack. These same studies, however, also reveal defensive potential, as LLMs can spot malicious intent in phishing emails (Heiding et al., 2024) and can support breach-and-attack simulation for defensive testing (Iturbe et al., 2024). This highlights why defensive usefulness and the potential for misuse should be assessed together.

### *4.5 Governance and Ethical Considerations (RQ3)*

Institutional guidance has evolved quickly and now spans several complementary layers: risk management, secure development, threats at the application level, monitoring of the threat landscape, adversarial knowledge bases, management systems, and binding regulation (Table 5). These instruments agree strongly at the conceptual level on lifecycle risk management, secure-by-design development, least privilege, transparency, and human oversight. They do, however, vary in legal status, level of detail, and intended audience, and, as McIntosh et al. (2024) point out, they are still mainly qualitative and seldom linked to measurable technical controls. Likewise, Jangirala et al. (2026) report a continuing lack of standardized ethical and regulatory guidance for deploying LLMs in cybersecurity. As a result, ethical issues such as accountability for automated decisions, the privacy of security data handled by GenAI systems, and the responsible management of dual-use capabilities are only partly dealt with in practice.

**Table 5. Comparison of governance frameworks and standards relevant to GenAI security**

| Instrument | Type and Issuer | Primary Contribution to GenAI Security |
|---|---|---|
| NIST AI 600-1: Generative AI Profile (NIST, 2024a) | Voluntary framework; U.S. NIST | Sets out risks that are specific to or worsened by GenAI and links recommended actions to the AI RMF functions (Govern, Map, Measure, Manage) |
| NIST SP 800-218A (NIST, 2024b) | Voluntary guidance; U.S. NIST | Expands the Secure Software Development Framework to cover GenAI and dual-use foundation models |
| OWASP Top 10 for LLM Applications (2023, 2025, 2026) | Community standard; OWASP | Ranked list of application-level risks such as prompt injection, disclosure, poisoning, and excessive agency |
| ENISA Threat Landscape 2024 (ENISA, 2024) | Threat report; EU agency | Tracks how the threat landscape is changing, AI-enabled threats included |
| Roadmap for AI (CISA, 2023) | Strategy; U.S. CISA | Encourages beneficial use of AI, secure-by-design AI, and the protection of critical infrastructure |
| Guidelines for Secure AI System Development (NCSC et al., 2023) | Multinational guidance; UK NCSC, U.S. CISA, and partners | Covers the secure design, development, deployment, and operation of AI systems |
| EU AI Act, Regulation (EU) 2024/1689 (European Parliament and Council of the European Union, 2024) | Binding regulation; European Union | Sets risk-based obligations and requires accuracy, robustness, and cybersecurity from high-risk AI systems |
| ISO/IEC 42001:2023 (ISO/IEC, 2023) | Certifiable management-system standard; ISO/IEC | AI management system at the organizational level, addressing risk, accountability, and continual improvement |
| MITRE ATLAS (MITRE, 2024) | Knowledge base; MITRE | Catalogue of adversary tactics and techniques targeting AI systems, used for threat modeling and red teaming |

### *4.6 Emerging Agentic Systems (RQ2)*

Recent research on autonomous, tool-using LLM agents takes GenAI past content generation and into planning and acting. The studies we reviewed show that memory, tool access, planning, and coordination among multiple agents bring extra attack surfaces as well as open questions about execution safety and the reliability of reasoning (Xu, Y. et al., 2026; Chowa et al., 2026). Benchmarks for agents point to a steady rise in offensive capability (Zhang, A. K. et al., 2025; Fang et al., 2024), and the higher ranking of excessive agency in the 2026 OWASP list shows that practitioners share this concern (OWASP GenAI Security Project, 2026). These results strengthen the case for least-privilege tool access, ongoing monitoring, and human sign-off for actions with high impact.

### *4.7 Toward an Integrated Framework: The Human–AI Collaboration Model*

The six themes above are closely connected. Defensive applications and operational opportunities (Sections 4.1–4.2) only deliver value if the related risks (Sections 4.3–4.4) are actively controlled through governance mechanisms (Section 4.5) and through safeguards suited to systems that are becoming more autonomous (Section 4.6). To bring these themes together in a form that security organizations can put into practice, we propose a four-stage Human–AI Collaboration Model, shown in Figure 2 and summarized in Table 6.

During **Stage 1, GenAI Assistance**, the system carries out a limited, clearly defined task, for example summarizing a threat report, drafting defensive code, extracting indicators of compromise, or explaining a vulnerability. The model does not act on its own at this point; it produces an artifact or recommendation that will be reviewed, in line with the augmentation role discussed throughout Section 2.

During **Stage 2, Human Validation**, a qualified analyst checks the output for factual accuracy, relevance, and completeness before it is used. This responds directly to the risks of hallucination, insecure output, and low reliability discussed in Sections 4.3 and 5.3. The depth of validation depends on the level of risk: outputs with higher impact, such as suggested remediations or generated content resembling exploits, are examined more carefully than lower-impact outputs, such as a draft summary meant only for internal use.

During **Stage 3, Controlled Action**, only outputs that have been validated move on to execution, and only through auditable channels with least privilege. In agentic or tool-using systems, this stage matches the human-approval gates and controlled permissions described in Sections 4.6 and 5.4. It makes sure that autonomous execution cannot sidestep human accountability and reduces the damage a successful prompt injection could cause.

During **Stage 4, Continuous Feedback**, the results of every action are recorded, including whether the recommendation was right, whether the action worked, and any errors that occurred. These results are fed back into monitoring and evaluation and, where relevant, into refining the model or the prompts. This stage puts into practice the continuous monitoring and evaluation principles that the governance literature stresses (Section 4.5), and it supports the benchmarking and continuous red-teaming directions set out in Section 6.

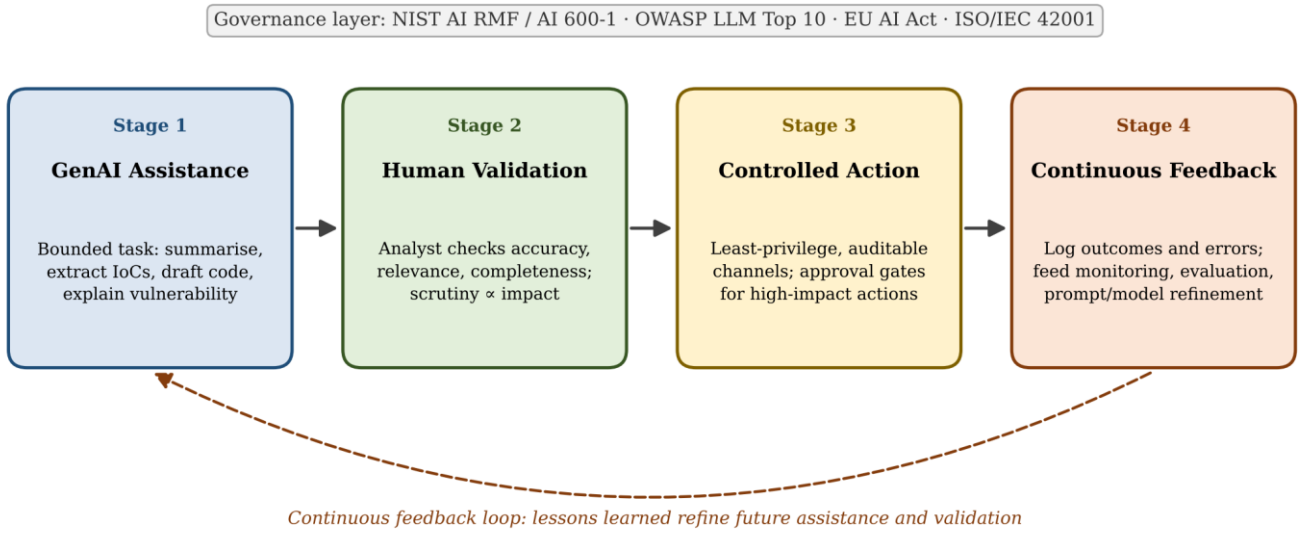



**Figure 2. The proposed four-stage Human–AI Collaboration Model for GenAI-enabled cybersecurity.**

**Table 6. Stages of the Human–AI Collaboration Model**

| Stage | Description | Primary Risk or Gap Addressed |
|---|---|---|
| 1. GenAI Assistance | The system carries out a limited, clearly defined task, such as summarizing a report, drafting code, extracting indicators of compromise, or explaining a vulnerability, without acting autonomously. | Establishes the augmentation role (Sections 2 and 4.1–4.2) |
| 2. Human Validation | A qualified analyst checks the output for accuracy, relevance, and completeness, and the depth of review matches the possible impact of the output. | Hallucination, insecure output, and reliability (Sections 4.3 and 5.3) |
| 3. Controlled Action | Only validated outputs are executed, through least-privilege, auditable, and pre-approved channels; actions with high impact need explicit human approval. | Excessive agency, prompt injection, and agentic security (Sections 4.6 and 5.4) |
| 4. Continuous Feedback | Results, including correctness, success, and errors, are recorded and fed back into monitoring, evaluation, and refinement of the model or prompts. | Governance, monitoring, and benchmarking gaps (Sections 4.5, 5.1, and 5.8) |

We deliberately kept the model generic so that it can be applied in different ways for different use cases. A workflow for summarizing threat intelligence, a workflow for code remediation, and an autonomous incident-response agent would each fill the four stages with their own specific controls, while still following the same sequence of assistance, validation, controlled action, and feedback. The model is also consistent with existing governance instruments. Stage 3, for instance, applies the least-privilege principles of the secure AI development guidelines (NCSC et al., 2023), and Stage 4 supports the Measure and Manage functions of the NIST Generative AI Profile (NIST, 2024a). In this way, the model reflects the main conclusion of this review: the value of GenAI in cybersecurity depends less on how capable the model is and more on how carefully organizations build human oversight around it. Testing the model empirically in operational settings is left for future work.

## 5. Research Gaps (RQ4)

Our review points to several connected research gaps that stand in the way of reliable and responsible GenAI adoption in cybersecurity. They relate not only to the technical performance of GenAI systems but also to operational validation, security, privacy, explainability, governance, and evaluation.

### *5.1 Evaluation Inconsistency*

Because studies rely on different datasets, models, tasks, and evaluation metrics, it is hard to compare their results in a meaningful way (Ferrag et al., 2025; Sheng et al., 2025).

### *5.2 Limited Real-World Validation*

A large number of proposed systems have not moved beyond the prototype or controlled-experiment stage, and evidence from long-term operational use is missing (Zhang, J. et al., 2025; Das et al., 2025). Field studies like the one by Wang et al. (2025) are still uncommon.

### *5.3 Reliability and Hallucination*

In security-critical applications, better factual accuracy, uncertainty estimation, and evidence-grounded generation are needed to limit the impact of wrong outputs (Uddin et al., 2025; Chowa et al., 2026).

### *5.4 Security of Agentic Systems*

Tool use, memory, planning, and coordination among multiple agents open up further attack surfaces, among them indirect prompt injection via retrieved content, and these need to be evaluated systematically (Greshake et al., 2023; Xu, Y. et al., 2026).

### *5.5 Privacy-Preserving Architectures*

More work is required on deployment methods that keep sensitive security data safe through local inference, data minimization, and controlled retrieval (Das et al., 2025; Zhang, R. et al., 2025).

### *5.6 Explainability and Evidence Traceability*

Security analysts need AI-generated recommendations to be backed by evidence and reasoning that they can check for themselves, but only a small number of studies assess whether GenAI outputs can be traced back to their sources.

### *5.7 Operational Governance*

Organizations require practical governance mechanisms that tie technical controls to accountability, risk ownership, monitoring, and compliance, whereas current frameworks are still mostly qualitative (McIntosh et al., 2024).

### *5.8 Standardized Benchmarking*

Benchmarks such as CyberSecEval, CTIBench, and Cybench mark real progress (Bhatt et al., 2023; Alam et al.,

2024; Zhang, A. K. et al., 2025), but there are still no agreed benchmarks that measure both defensive usefulness and security robustness using common metrics.

### *5.9 Regulatory Harmonization*

The guidance issued by NIST, OWASP, ENISA, CISA, and the NCSC, along with the EU AI Act and ISO/IEC 42001, agrees at the conceptual level, yet these instruments have not been combined into one actionable framework that works across jurisdictions (Jangirala et al., 2026).

## 6. Future Research Directions

The gaps described above offer a clear starting point for future work. Instead of handling technical capability, operational deployment, and governance as separate research areas, future studies should look for solutions that tackle these limitations together.

### *6.1 Trustworthy and Evidence-Grounded GenAI*

Build mechanisms for checking facts, estimating uncertainty, and grounding generation in evidence, in order to close the reliability gap (Section 5.3).

### *6.2 Privacy-Preserving LLM Deployment*

Study secure local deployment, data minimization, and controlled retrieval as ways to protect sensitive security information (Section 5.5).

### *6.3 Standardized Evaluation and Benchmarking*

Create shared datasets, tasks, and metrics that allow GenAI systems to be compared in a reproducible way, extending the benchmarks that already exist (Sections 5.1 and 5.8).

### *6.4 Secure Agentic AI*

Design least-privilege tool access, strong protection against direct and indirect prompt injection, monitoring mechanisms, and human approval gates for autonomous and multi-agent systems (Section 5.4).

### *6.5 Multimodal Cybersecurity Models*

Explore models that combine text, source code, logs, network data, and visual evidence so that security analysis can be more complete.

### *6.6 Continuous Red-Teaming and Adversarial Evaluation*

Put in place ongoing security testing across development and deployment, using adversarial knowledge bases such as MITRE ATLAS to organize test coverage (MITRE, 2024).

### *6.7 Integration with Security Infrastructure*

Examine controlled, auditable integration with SIEM, SOAR, EDR, vulnerability-management, and threat-intelligence platforms to support stronger real-world validation (Section 5.2).

### *6.8 Human-Centered GenAI Design*

Assess analyst workload, trust, decision quality, over-reliance, and error rates to learn how GenAI shapes human decision-making, including empirical testing of the Human–AI Collaboration Model from Section 4.7.

### *6.9 Secure AI Supply Chains*

Create security mechanisms for models, datasets, APIs, plugins, packages, and the infrastructure behind them, including protection against the poisoning of training data (Carlini et al., 2024).

### *6.10 AI-versus-AI Cyber Competition*

Study how offensive and defensive AI systems interact and what this means for security strategies in the future.

### *6.11 Cross-Jurisdictional Governance Harmonization*

Develop practical ways to map and align governance frameworks and standards across jurisdictions and to connect them with measurable technical controls (Sections 5.7 and 5.9).

## 7. Conclusion

GenAI is turning into an important technology for cybersecurity, as it can produce, summarize, interpret, and transform security information at a scale that strengthens the work of human analysts. The literature we reviewed shows uses in threat intelligence, threat detection, vulnerability assessment and remediation, secure software development, phishing defense, penetration testing, deception, security operations, and incident response. These capabilities open the door to automation, quicker analysis, better scalability, and wider access to security expertise, even though the strength of the evidence differs considerably between application areas.

At the same time, GenAI is dual-use by nature. Primary studies indicate that it can make phishing more effective and less costly, help in exploiting known vulnerabilities, and create new attack surfaces in AI-enabled applications. As a result, prompt injection, sensitive information disclosure, data and model poisoning, supply-chain vulnerabilities, insecure output, hallucination, misinformation, and excessive agency are key concerns. Current governance instruments agree in principle, but they are still mostly qualitative and have not yet been harmonized across jurisdictions.

The answers to the research questions can be summarized as follows. For RQ1, GenAI has been used in cyber threat intelligence, threat detection, vulnerability assessment and remediation, secure software development, phishing defense, penetration testing, deception, malware analysis, and security operations, but the evidence behind these uses is of low to moderate maturity, and evidence from long-term operational deployment is still scarce (Section 4.1, Table 4). For RQ2, the main risks are prompt injection, data and model poisoning, insecure output, hallucination, sensitive information disclosure, supply-chain weaknesses, and excessive agency, in addition to attackers misusing GenAI

to scale up phishing and vulnerability exploitation (Sections 4.3, 4.4, and 4.6). For RQ3, current frameworks and standards share common ground on lifecycle risk management, secure-by-design development, and human oversight, but they are still mainly qualitative, are seldom linked to measurable technical controls, and lack harmonization across jurisdictions (Section 4.5, Table 5). For RQ4, the key gaps relate to consistent evaluation, real-world validation, reliability, the security of agentic systems, privacy, explainability, operational governance, benchmarking, and regulatory harmonization, and each one is matched to a future research direction (Sections 5 and 6).

We conclude that the future of GenAI in cybersecurity relies on adoption that is trustworthy and centered on people. Organizations should pair GenAI capabilities with secure-by-design development, least-privilege access, privacy protection, ongoing evaluation, evidence-based validation, and clear human accountability. A key contribution of this paper is the Human–AI Collaboration Model in Section 4.7, which turns these principles into a practical four-stage process linking GenAI assistance, human validation, controlled action, and continuous feedback. Future work should give priority to standardized benchmarks, privacy-preserving architectures, secure autonomous agents, multimodal security, and operational validation in realistic settings.

## References


1. Alam, M. T., Bhusal, D., Nguyen, L., & Rastogi, N. (2024). CTIBench: A benchmark for evaluating LLMs in cyber threat intelligence. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 50805–50825). Curran Associates.
2. Alqahtani, H., & Kumar, G. (2026). Large language models for cybersecurity intelligence: A systematic review of emerging threats, defensive capabilities, and security evaluation frameworks. *Computers, Materials & Continua, 87*(3). https://doi.org/10.32604/cmc.2026.077367
3. Balasubramanian, P., Liyana, S., Sankaran, H., Sivaramakrishnan, S., Pusuluri, S., Pirttikangas, S., et al. (2025). Generative AI for cyber threat intelligence: Applications, challenges, and analysis of real-world case studies. *Artificial Intelligence Review, 58*, Article 336. https://doi.org/10.1007/s10462-025-11338-z
4. Bhatt, M., Chennabasappa, S., Nikolaidis, C., Wan, S., Evtimov, I., et al. (2023). *Purple Llama CyberSecEval: A secure coding benchmark for language models* (arXiv:2312.04724). arXiv. https://doi.org/10.48550/arXiv.2312.04724
5. Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2024). Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 407–425). IEEE.
6. Chen, Y., Cui, M., Wang, D., Cao, Y., et al. (2024). A survey of large language models for cyber threat detection. *Computers & Security, 145*, Article 104016. https://doi.org/10.1016/j.cose.2024.104016
7. Chowa, S. S., Alvi, R., Rahman, S. S., Rahman, M. A., Raiaan, M. A. K., Islam, M. R., Hussain, M., & Azam, S. (2026). From language to action: A review of large language models as autonomous agents and tool users. *Artificial Intelligence Review, 59*, Article 71. https://doi.org/10.1007/s10462-025-11471-9
8. Cybersecurity and Infrastructure Security Agency [CISA]. (2023). *Roadmap for artificial intelligence*. U.S. Department of Homeland Security.
9. Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys, 57*(6), Article 152. https://doi.org/10.1145/3712001
10. Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., & Rass, S. (2024). PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *Proceedings of the 33rd USENIX Security Symposium* (pp. 847–864). USENIX Association.
11. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.
12. European Union Agency for Cybersecurity [ENISA]. (2024). *ENISA threat landscape 2024*. ENISA.
13. Fang, R., Bindu, R., Gupta, A., & Kang, D. (2024). *LLM agents can autonomously exploit one-day vulnerabilities* (arXiv:2404.08144). arXiv. https://doi.org/10.48550/arXiv.2404.08144
14. 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*, 1–46. https://doi.org/10.1016/j.iotcps.2025.01.001
15. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79–90). ACM. https://doi.org/10.1145/3605764.3623985

16. Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access, 12*, 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882

17. Hu, Y., Zou, F., Han, J., Sun, X., & Wang, Y. (2024). LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model. *Computers & Security, 145*, Article 103999. https://doi.org/10.1016/j.cose.2024.103999

18. Ibrar, W., Mahmood, D., Al-Shamayleh, A. S., Ahmed, G., Alharthi, S. Z., & Akhunzada, A. (2025). Generative AI: A double-edged sword in the cyber threat landscape. *Artificial Intelligence Review, 58*, Article 285. https://doi.org/10.1007/s10462-025-11285-9

19. International Organization for Standardization & International Electrotechnical Commission [ISO/IEC]. (2023). *Information technology — Artificial intelligence — Management system* (ISO/IEC Standard No. 42001:2023). ISO.

20. Iturbe, E., Llorente-Vázquez, O., Rego, A., Rios, E., & Toledo, N. (2024). Unleashing offensive artificial intelligence: Automated attack technique code generation. *Computers & Security, 147*, Article 104077. https://doi.org/10.1016/j.cose.2024.104077

21. Jangirala, S., Gupta, V., & Mandal, A. (2026). Large language models in cybersecurity: A PRISMA-ScR-guided scoping review of threats, defenses, and emerging applications. *IEEE Access, 14*, 117961–117990. https://doi.org/10.1109/ACCESS.2026.3719163

22. Johnson, S., Hassing, R., Pijpker, J., & Loves, R. (2024). A modular generative honeypot shell. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE. https://doi.org/10.1109/CSR61664.2024.10679411

23. Karras, A., Theodorakopoulos, L., Karras, C., Theodoropoulou, A., Kalliampakou, I., & Kalogeratos, G. (2025). LLMs for cybersecurity in the Big Data era: A comprehensive review of applications, challenges, and future directions. *Information, 16*(11), Article 957. https://doi.org/10.3390/info16110957

24. McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., Xu, D., Liu, D., Nowrozy, R., & Halgamuge, M. N. (2024). From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *Computers & Security, 144*, Article 103964. https://doi.org/10.1016/j.cose.2024.103964

25. Mirtaheri, S. L., Movahed, N., Shahbazian, R., Pascucci, V., & Pugliese, A. (2026). Cybersecurity in the age of generative AI: A systematic taxonomy of AI-powered vulnerability assessment and risk management. *Future Generation Computer Systems, 175*, Article 108107. https://doi.org/10.1016/j.future.2025.108107

26. MITRE. (2024). *ATLAS: Adversarial threat landscape for artificial-intelligence systems*. The MITRE Corporation. https://atlas.mitre.org

27. Moia, V. H. G., Sanz, I. J., Rebello, G. A. F., de Meneses, R. D., Hitaj, B., & Lindqvist, U. (2026). LLM in the middle: A systematic review of threats and mitigations to real-world LLM-based systems. *Computer Science Review, 61*, Article 100916. https://doi.org/10.1016/j.cosrev.2026.100916

28. National Cyber Security Centre [NCSC], Cybersecurity and Infrastructure Security Agency, et al. (2023). *Guidelines for secure AI system development*. NCSC.

29. National Institute of Standards and Technology [NIST]. (2024a). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

30. National Institute of Standards and Technology [NIST]. (2024b). *Secure software development practices for generative AI and dual-use foundation models: An SSDF community profile* (NIST SP 800-218A). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-218A

31. OWASP Foundation. (2023). *OWASP Top 10 for large language model applications* (Version 1.1). Open Worldwide Application Security Project.

32. OWASP GenAI Security Project. (2025). *OWASP Top 10 for LLM and generative AI applications: 2025*. Open Worldwide Application Security Project.

33. OWASP GenAI Security Project. (2026). *OWASP GenAI LLM Top 10 2026*. Open Worldwide Application Security Project. https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

34. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

35. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 754–768). IEEE. https://doi.org/10.1109/SP46214.2022.9833571

36. 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*, *12*, 53497–53516. https://doi.org/10.1109/ACCESS.2024.3385107

37. Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, *57*, Article 324. https://doi.org/10.1007/s10462-024-10973-2

38. Sheng, Z., Chen, Z., Gu, S., Huang, H., Gu, G., & Huang, J. (2025). LLMs in software security: A survey of vulnerability detection techniques and insights. *ACM Computing Surveys*, *58*(5), Article 134. https://doi.org/10.1145/3769082

39. Shenoy, N., & Mbaziira, A. V. (2024). An extended review: LLM prompt engineering in cyber defense. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE. https://doi.org/10.1109/ICECET61485.2024.10698605

40. Uddin, M., Irshad, M. S., Kandhro, I. A., Alanazi, F., Ahmed, F., Maaz, M., Hussain, S., et al. (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. *Artificial Intelligence Review*, *58*, Article 236. https://doi.org/10.1007/s10462-025-11219-5

41. Vadisetty, R., & Polamarasetti, A. (2024). Generative AI-driven distributed cybersecurity frameworks for AI-integrated global Big Data systems. In *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)* (pp. 595–600). IEEE. https://doi.org/10.1109/EmergIN63207.2024.10961616

42. Wang, X., Tian, Y., Huang, K., & Liang, B. (2025). Practically implementing an LLM-supported collaborative vulnerability remediation process: A team-based approach. *Computers & Security*, *148*, Article 104113. https://doi.org/10.1016/j.cose.2024.104113

43. Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2025). Large language models for cyber security: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3769676

44. Xu, Y., Zhuang, Y., Liu, X., Zhang, T., Xiao, B., Xu, X., Jiang, D., Wang, J., et al. (2026). LLM agents security duality: A comprehensive survey of self-security and empowered cybersecurity. *Artificial Intelligence Review*, *59*, Article 174. https://doi.org/10.1007/s10462-026-11563-0

45. 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

46. Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., et al. (2025). Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.

47. Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., Li, L., Yang, Y., Zhu, H., & Meng, D. (2025). When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, *8*, Article 55. https://doi.org/10.1186/s42400-025-00361-w

48. Zhang, R., Li, H.-W., Qian, X.-Y., Jiang, W.-B., & Chen, H.-X. (2025). On large language models safety, security, and privacy: A survey. *Journal of Electronic Science and Technology*, *23*(1), Article 100301. https://doi.org/10.1016/j.jnlest.2025.100301

49. Zhou, W., Zhu, X., Han, Q.-L., Li, L., Chen, X., Wen, S., & Xiang, Y. (2025). The security of using large language models: A survey with emphasis on ChatGPT. *IEEE/CAA Journal of Automatica Sinica*, *12*(1), 1–26. https://doi.org/10.1109/JAS.2024.124983