Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077367

REVIEW

# Large Language Models for Cybersecurity Intelligence: A Systematic Review of Emerging Threats, Defensive Capabilities, and Security Evaluation Frameworks

**Hamed Alqahtani**[1] **and Gulshan Kumar**[2,*]

[1]Informatics and Computer Systems Department, College of Computer Science, Center of Artificial Intelligence, King Khalid University, Abha, Saudi Arabia

[2]Department of Computer Applications, Shaheed Bhagat Singh State University, Ferozepur, Punjab, India

*Corresponding Author: Gulshan Kumar. Email: gulshanahuja@gmail.com

Received: 08 December 2025; Accepted: 09 February 2026; Published: 09 April 2026

**ABSTRACT:** Large Language Models (LLMs) are becoming integral components of modern cybersecurity ecosystems, simultaneously strengthening defensive capabilities while giving rise to a new class of Artificial Intelligence–Generated Content (AIGC)-driven threats. This PRISMA-guided systematic review synthesises 167 peer-reviewed studies published between 2022 and 2025 and proposes a unified threat–defence–evaluation taxonomy as a central analytical framework to consolidate a previously fragmented body of research. Guided by this taxonomy, the review first examines AIGC-enabled threats, including automated and highly personalised phishing, polymorphic malware and exploit generation, jailbreak and adversarial prompting, prompt-injection attack vectors, multimodal deception, persona-steering attacks, and large-scale disinformation campaigns. The surveyed evidence indicates a qualitative escalation in adversarial capabilities, with LLMs significantly enhancing scalability, adaptability, and realism while markedly reducing the technical barriers to conducting sophisticated attacks. Second, the review analyses LLM-enabled defensive applications spanning intrusion and anomaly detection, malware analysis and log-semantic modelling, multilingual threat intelligence extraction, vulnerability discovery and code repair, and Security Operations Center (SOC) automation through Retrieval-Augmented Generation (RAG) and multi-agent systems. Although these approaches demonstrate strong potential as semantic reasoning and decision-support components within hybrid security architectures, their real-world effectiveness remains constrained by hallucination risks, adversarial susceptibility, distributional shifts, and operational overhead. Third, the review synthesises current security evaluation and red-teaming practices, revealing a fragmented assessment landscape characterised by narrow benchmarks, inconsistent evaluation metrics, and limited longitudinal robustness analysis. Overall, the taxonomy-driven synthesis highlights a structurally imbalanced ecosystem in which offensive innovation outpaces defensive maturity and governance, and it informs a structured, research-question-aligned roadmap for developing trustworthy, resilient, and policy-aligned LLM-powered cybersecurity systems.

**KEYWORDS:** Artificial intelligence–generated content threats; cybersecurity intelligence; large language model–based defensive systems; large language models; red-teaming and evaluation frameworks

## 1 Introduction

Large Language Models (LLMs), developed using large-scale transformer architectures, have rapidly emerged as foundational components of contemporary artificial intelligence, enabling significant advances in reasoning, code generation, summarisation, automated analysis, and decision support across high-stakes application domains [1,2]. Their capacity to process heterogeneous, unstructured, and multilingual data at

scale renders them particularly well suited to cybersecurity environments, where analysts must interpret complex and high-volume artefacts, including system logs, threat intelligence reports, malware analyses, vulnerability disclosures, and Security Operations Center (SOC) narratives [3–5]. At the same time, the same generative and reasoning capabilities that enable defensive augmentation can be readily exploited by adversaries. A rapidly expanding body of peer-reviewed research documents the use of LLMs to craft highly personalised phishing campaigns [6,7], generate polymorphic and evasive malware [8,9], automate reconnaissance and exploit scaffolding [10], conduct multimodal deception and persona-steering attacks [11,12], and orchestrate large-scale misinformation and disinformation operations [13,14]. This dual-use character has profound implications for the evolving balance between cyber offence and cyber defence.

Despite growing interest in this area, existing surveys exhibit substantial variation in scope, methodological rigor, and evidentiary depth. Yao et al. synthesise model-level vulnerabilities and privacy risks associated with LLM deployment [15]. Jaffal et al. provide a detailed examination of LLM capabilities, limitations, and defensive applications within SOC and cyber threat intelligence (CTI) ecosystems [3]. Zhang et al. offer a structured review of LLM-driven cybersecurity applications [16], while Karras et al. discuss broader opportunities and risks in big-data security pipelines [4]. However, these surveys share notable limitations: many rely primarily on narrative synthesis rather than systematic review protocols; several include a substantial proportion of preprint literature; few integrate offensive, defensive, and evaluative research strands within a single analytical framework; and none provide a PRISMA [17]-aligned, SCI/SCIE-restricted taxonomy that jointly captures threats, defences, and security evaluation methodologies. As the field continues to expand rapidly, the absence of a comprehensive and methodologically transparent synthesis impedes cross-study comparability, evidence consolidation, and the standardisation of research practices.

To address these gaps, this article presents the first PRISMA-compliant systematic survey of LLMs for cybersecurity intelligence covering the period 2022–2025, explicitly restricted to peer-reviewed SCI/SCIE publications. The overarching objective is to construct a rigorous and unified evidence base that characterises how LLMs reshape cyber offence, cyber defence, and the evaluation ecosystems governing trustworthy deployment. Specifically, this study pursues four core aims:

- to systematically identify and synthesise Artificial Intelligence–Generated Content (AIGC)-driven cyber threats enabled, amplified, or operationalised by LLMs, including phishing, malware generation, jailbreak attacks, deception, misinformation, and persona steering [11];
- to analyse LLM-enabled defensive capabilities spanning intrusion detection [18], malware semantic modelling [19], threat intelligence extraction [20,21], vulnerability assessment and remediation [22], SOC automation [23], and multi-agent defence architectures [24];
- to examine emerging security evaluation frameworks, including benchmarks [25], harm and refusal metrics [3], adversarial red-teaming methodologies [15,26], and lifecycle-oriented risk assessments [27], alongside the growing need for ethical and governance frameworks [28,29];
- to distil cross-cutting methodological limitations, systemic risks (including privacy [30] and ethical compliance [31]), and unresolved research challenges derived from 167 peer-reviewed primary studies and several high-impact surveys.

These aims motivate the following research questions:

- **RQ1:** What forms of cybersecurity threats are generated, enhanced, or operationalised through AIGC and LLMs, and how are these threats empirically studied and modelled?
- **RQ2:** How are LLMs leveraged to strengthen cyber defence capabilities across detection, analysis, and response workflows, and what limitations persist?

- **RQ3:** What security evaluation frameworks, benchmarks, and red-teaming methodologies exist for assessing LLM robustness, safety, and reliability under adversarial conditions?
- **RQ4:** What methodological gaps, systemic challenges, and governance limitations hinder the trustworthy deployment of LLM-based cybersecurity systems?

The primary contributions of this survey are as follows:

- a PRISMA-aligned systematic review protocol tailored to LLM-centric cybersecurity research, detailing search strategies, inclusion and exclusion criteria, quality appraisal procedures, and multi-stage screening;
- a critical comparison of influential surveys [3,4,15,16], motivating the unified threat–defence–evaluation taxonomy introduced in this work;
- a structured synthesis of 167 peer-reviewed primary studies addressing AIGC-driven threats [6,13], LLM-based defensive mechanisms [18,19,23], and security evaluation and red-teaming frameworks [25,32];
- the identification of cross-cutting challenges—including hallucination, adversarial fragility, privacy and governance risks, evaluation fragmentation, lifecycle vulnerabilities, and the absence of standardised assurance frameworks—followed by a comprehensive future research roadmap for resilient, auditable, and security-aligned LLM deployments.

In contrast to prior surveys, this review grounds each threat and defence category in empirically evaluated case studies and critically analyses their operational limitations under realistic SOC and intrusion detection deployment scenarios. Overall, this survey offers a methodologically rigorous, conceptually integrated, and empirically grounded examination of how LLMs are transforming the cybersecurity landscape, both as powerful enablers of defensive intelligence and as catalysts for increasingly adaptive and scalable adversarial behaviour. By consolidating diverse research trajectories and introducing a unified taxonomy alongside a forward-looking research agenda, this work aims to support the development of trustworthy, verifiable, and operationally reliable LLM-driven cybersecurity systems.

The remainder of this article is organised as follows. Section 2 describes the PRISMA-guided review methodology. Section 3 situates this study within the existing survey literature. Section 4 presents the unified threat–defence–evaluation taxonomy and analytical framework. Sections 5 and 6 analyse AIGC-driven threats and LLM-based defensive capabilities, while Section 7 reviews emerging security evaluation and red-teaming methodologies. Section 8 examines methodological patterns across the evidence base. Section 9 synthesises cross-cutting challenges, and Section 10 outlines strategic future research directions. Section 11 concludes the paper.

## 2 Methodology

This review adopts a rigorously defined systematic literature review (SLR) methodology grounded in the PRISMA 2020 guidelines to ensure transparency, reproducibility, and analytical completeness. Given the rapid evolution and methodological heterogeneity of Large Language Model (LLM) research in cybersecurity, particular emphasis is placed on bias mitigation, inductive taxonomy construction, and cross-study comparability. The final evidence base comprises 167 peer-reviewed studies published between 2022 and 2025, each systematically mapped to the unified threat–defence–evaluation taxonomy introduced in Section 4. An overview of the study identification and selection process is provided in the PRISMA flow diagram shown in Fig. 1.

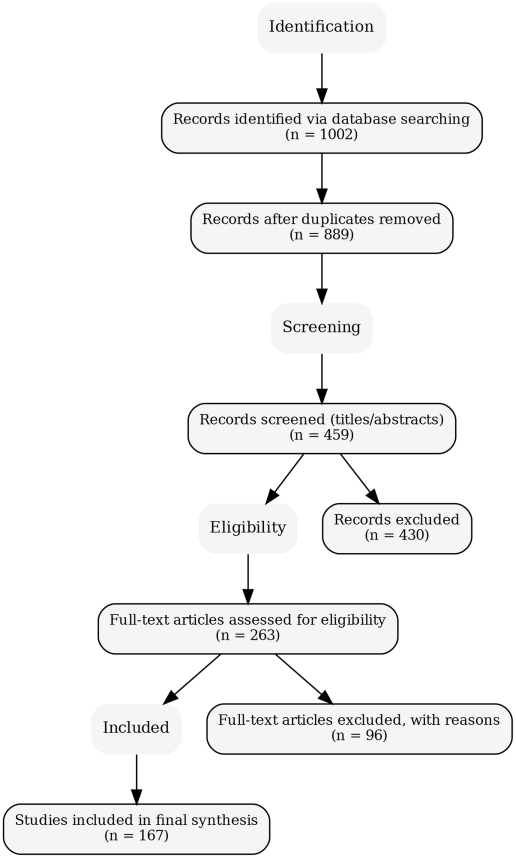


**Figure 1:** PRISMA flow diagram summarising identification, screening, eligibility assessment, and final inclusion of 167 studies.

### *2.1 Review Protocol and PRISMA Workflow*

The review protocol was specified *a priori* in accordance with PRISMA recommendations to minimise selection bias and enhance methodological transparency. The review workflow comprised four sequential stages: identification, screening, eligibility assessment, and final inclusion. The initial search yielded 1002 records across multiple scholarly databases. Following automated and manual deduplication, 889 unique studies were subjected to title and abstract screening, resulting in the exclusion of 430 records. The remaining 459 studies underwent detailed screening, leading to 263 full-text articles being assessed for eligibility. Of these, 96 studies were excluded due to insufficient cybersecurity relevance, incomplete methodological reporting, or failure to satisfy the predefined inclusion criteria. The final corpus therefore consisted of 167 peer-reviewed studies, representing a balanced cross-section of AIGC-driven threats, LLM-enabled defensive architectures, and security evaluation, safety, and governance frameworks. Fig. 1 provides a transparent visual summary of this selection process.

### *2.2 Inductive Taxonomy Construction and Coding Rationale*

Rather than imposing an externally defined classification scheme, this review employs a bottom-up inductive coding strategy to derive the unified taxonomy presented in Section 4. Following full-text screening, each study was independently coded along multiple analytical dimensions, including: (i) the primary cybersecurity task addressed, (ii) the functional role of the LLM within the system architecture,

(iii) the threat or defensive context, (iv) the evaluation or validation methodology, and (v) the assumed deployment environment (e.g., laboratory-scale, SOC-scale, or safety-critical systems).

Initial open coding revealed recurring conceptual and methodological patterns across ostensibly diverse contributions. Through iterative axial coding and constant comparative analysis, these patterns converged into four dominant dimensions: **(T) AIGC-driven threat vectors**, **(D) LLM-based defensive capabilities**, **(E) evaluation and red-teaming methodologies**, and **(M) methodological and governance orientation**. For instance, studies addressing automated phishing, misinformation generation, exploit scaffolding, and prompt-based attacks consistently clustered within a unified threat dimension, whereas contributions focusing on SOC automation, intrusion detection, vulnerability analysis, and agentic response mechanisms formed a distinct defensive axis.

Subcategories within each dimension were further refined through frequency analysis and conceptual differentiation. Within the defensive dimension, intrusion detection, threat intelligence automation, secure code analysis, and autonomous response emerged as separable categories based on differences in data modality, latency constraints, and operational responsibility. Evaluation-oriented studies similarly diverged into benchmark construction, adversarial red-teaming, metric development, and lifecycle robustness analysis. Edge cases and hybrid studies were explicitly retained and cross-mapped to reflect the interdisciplinary and rapidly converging nature of LLM-driven cybersecurity research. This inductive process ensures that the resulting taxonomy faithfully reflects the empirical structure of the literature rather than an *a priori* conceptual framework.

### *2.3 Search Strategy and Information Sources*

Searches were conducted across eight authoritative scholarly databases: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, Taylor & Francis, Scopus, and Web of Science. These sources were selected to capture interdisciplinary research spanning cybersecurity, artificial intelligence, software engineering, and digital governance. The review period extends from January 2022 to December 2025, corresponding to the widespread deployment of transformer-based LLMs (e.g., GPT-3+, LLaMA, PaLM, Codex) in both offensive and defensive cybersecurity contexts.

Search queries combined terms related to LLM architectures, cyber threat categories, defensive applications, and evaluation constructs using Boolean operators:

> (*"large language model" OR LLM OR "foundation model" OR transformer*) AND (*cybersecurity OR malware OR phishing OR vulnerability OR anomaly OR SOC*) AND (*AIGC OR "automated phishing" OR jailbreak OR "prompt injection" OR misuse*)

Search expressions were iteratively refined through pilot queries to incorporate emerging themes such as adversarial prompting, exploit generation, multi-agent reasoning, and cyber governance. Foundational surveys on LLM security and misuse [33–36] informed construct definition and ensured terminological consistency across the search strategy.

### *2.4 Inclusion, Exclusion, and Quality Appraisal*

Studies were included if they were peer-reviewed journal or conference publications written in English between 2022 and 2025 and examined LLMs or transformer-based models within cybersecurity-relevant contexts. Eligible studies addressed AIGC-enabled threats such as phishing, deception, exploit scaffolding, or jailbreak attacks [33,34]; proposed or evaluated LLM-based defensive mechanisms including SOC

automation, threat intelligence extraction, intrusion or anomaly detection, and vulnerability analysis [35,36]; or contributed to evaluation, red-teaming, benchmark design, or governance frameworks for LLM safety.

Exclusion criteria removed preprints, editorials, theses, and other non-peer-reviewed materials; studies lacking substantive cybersecurity relevance; purely generic natural language processing applications without adversarial or security implications; works with insufficient methodological detail; and duplicate publications. Methodological quality was assessed using a bespoke appraisal framework adapted from established SLR guidelines, evaluating the clarity of threat models, transparency of datasets and evaluation pipelines, reproducibility, metric suitability, and the discussion of limitations and ethical risks. Studies exhibiting methodological weaknesses were retained when they offered unique empirical or conceptual insights, with their limitations explicitly acknowledged in subsequent synthesis sections.

### *2.5 Inter-Reviewer Agreement, Bias Mitigation, and Sensitivity Analysis*

To ensure consistency and minimise subjective bias, study selection and coding were independently performed by two reviewers at both the title–abstract and full-text screening stages. Inter-reviewer agreement was quantified using Cohen's kappa coefficient, yielding substantial agreement during title–abstract screening ($\kappa = 0.81$) and near-perfect agreement during full-text eligibility assessment ($\kappa = 0.88$). Discrepancies were resolved through structured discussion and, where necessary, arbitration by a third reviewer.

Bias mitigation measures included multi-database coverage to reduce venue bias, *a priori* inclusion criteria, blinded screening where feasible, and the use of a standardised, taxonomy-aligned coding template to limit interpretive drift. A sensitivity analysis was conducted on borderline studies excluded during full-text screening (e.g., partially relevant preprints or studies with limited evaluation detail). Reintroduction of these studies did not materially affect the taxonomy structure, thematic distributions, or comparative conclusions, indicating that the review findings are robust and not unduly sensitive to marginal inclusion decisions.

## 3 Positioning with Respect to Existing Surveys

A substantial body of survey literature has examined intrusion detection systems (IDS) from the perspectives of machine learning, deep learning, and, more recently, transformer-based architectures. These IDS-centric surveys predominantly focus on network intrusion detection systems (NIDS), host-based intrusion detection systems (HIDS), and industrial or IoT-oriented IDS deployments, with primary emphasis on feature engineering strategies, model architectures, benchmark datasets, and performance metrics [37]. Representative works include surveys on transformer-based NIDS, hybrid deep learning IDS frameworks, and intrusion detection in IIoT and industrial control system (ICS) environments, where LLMs-if mentioned at all-are treated as high-capacity sequence models or feature-level classifiers rather than as reasoning or agentic components. Consequently, while these surveys provide valuable insights into detection accuracy, scalability, and dataset-specific performance, they do not address the broader implications of LLMs as cognitive security primitives operating across detection, reasoning, and response pipelines.

In parallel, a distinct line of survey research has emerged that focuses on the security, reliability, and societal implications of large language models. Foundational overviews by Alawida et al. [33], Esmradi et al. [34], and Lopez et al. [35] provide early mappings of the opportunities and risks introduced by generative AI. Alawida et al. [33] conduct a comprehensive examination of ChatGPT, analysing its capabilities, limitations, misuse potential, and associated ethical and privacy concerns. Esmradi et al. [34] systematically catalogue attack techniques and mitigation strategies related to LLM misuse, including data leakage and adversarial behaviours, while Lopez et al. [35] explore application-level opportunities and challenges. However, these surveys remain largely high-level and conceptual, and they do not engage in systematic comparison with IDS-centric research nor integrate intrusion detection within a broader cybersecurity lifecycle.

More recent contributions further narrow the scope toward governance, lifecycle risk, and domain-specific security applications. Nawara and Kashef [36] investigate LLM-powered recommendation and reasoning systems, highlighting security and governance challenges in decision-support pipelines, while Uddin et al. [38] present a critical analysis of generative AI risks, emphasising lifecycle vulnerabilities, regulatory gaps, and systemic misuse pathways relevant to cybersecurity. Despite their analytical depth, these works continue to treat intrusion detection, SOC automation, and response mechanisms as largely isolated application domains rather than as interdependent components within an end-to-end threat–defence–evaluation ecosystem.

The primary distinction between existing IDS-focused surveys and the present work lies in the level of abstraction and the conceptual role assigned to LLMs. IDS-centric surveys predominantly frame models-including transformers-as *feature-level classifiers* optimised for detection accuracy on specific datasets. In contrast, this review conceptualises LLMs as *cognitive and agentic security components* capable of semantic reasoning, contextual correlation, decision support, and autonomous coordination across multiple stages of the cybersecurity pipeline. As a result, intrusion detection is not analysed in isolation but is examined alongside AIGC-driven threats, SOC automation, multi-agent response systems, and security evaluation and red-teaming practices.

Table 1 systematically positions representative LLM- and IDS-focused surveys within the dimensions of the proposed unified taxonomy. As the comparison illustrates, prior surveys typically concentrate on isolated aspects of the problem space, such as AIGC-driven threat characterisation [33,34], application-level opportunities and limitations [35,36], or governance and lifecycle risk analysis [38]. However, none of these works provides a fully integrated, multi-dimensional perspective that jointly examines offensive misuse, defensive reasoning, evaluation and red-teaming practices, and methodological orientation within a single analytical framework. Furthermore, the absence of PRISMA-aligned screening and eligibility procedures across existing surveys limits reproducibility and systematic coverage.

**Table 1:** Comparison of existing survey taxonomies across IDS- and LLM-centric dimensions.

| Survey | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| *LLM-Centric Security Surveys* | | | | | | |
| Alawida et al. [33] | ✓ | × | ✓ | × | × | × |
| Esmradi et al. [34] | ✓ | ✓ | × | × | × | × |
| Lopez et al. [35] | ✓ | ✓ | ✓ | × | × | × |
| Nawara et al. [36] | × | ✓ | × | × | × | × |
| Uddin et al. [38] | ✓ | × | ✓ | × | × | × |
| *IDS-Centric (Transformer/DL-Based) Surveys* | | | | | | |
| Elouardi et al. [18] | × | ✓ | × | × | × | ✓ |
| Ferrag et al. [39] | × | ✓ | × | × | × | ✓ |
| Hasanov et al. [40] | × | ✓ | × | × | × | ✓ |
| **This Survey** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Note: A: AIGC-Driven Threat Coverage; B: LLM-Based Defensive Capabilities (including IDS and SOC automation); C: Evaluation & Red-Teaming Coverage; D: Unified Multi-Dimensional Taxonomy; E: PRISMA or SLR Alignment; F: Cognitive/Agentic Role of Models (beyond feature-level classification).

A further distinguishing factor, highlighted of Table 1, concerns model conceptualisation. Existing IDS-centric surveys—particularly those focusing on transformer-based NIDS, HIDS, and IIoT detection—primarily treat deep learning models as feature-level classifiers optimised for detection performance. In contrast, the present review frames LLMs as cognitive and agentic security components that enable semantic reasoning, contextual awareness, and coordinated decision-making across the full cybersecurity lifecycle, encompassing detection, analysis, response, and evaluation.

Unlike prior IDS-centric and LLM-centric reviews, the present study adopts a PRISMA-guided systematic methodology, restricts its evidence base to 167 peer-reviewed publications from 2022 to 2025, and synthesises findings using a unified four-dimensional taxonomy encompassing AIGC-driven threats, LLM-enabled defensive capabilities (including intrusion detection and SOC automation), security evaluation and red-teaming practices, and methodological orientation. This integrated framework enables a holistic characterisation of LLM-driven cybersecurity research and explicitly bridges the conceptual gap between traditional IDS surveys and emerging LLM security studies. By positioning intrusion detection as one component within a broader cognitive and agentic defence ecosystem, this review provides a more comprehensive, coherent, and methodologically transparent synthesis to inform future research and real-world deployment.

## 4 A Unified Taxonomy of LLM–Cybersecurity Research

To systematically synthesise the heterogeneous body of literature identified through the PRISMA-guided review process, this section introduces a unified taxonomy of LLM-centric cybersecurity research. The taxonomy is derived through an inductive coding and synthesis procedure applied to all 167 included studies, as described in Section 2.2. Rather than imposing a pre-defined classification scheme, the taxonomy is grounded in empirically observed research patterns spanning adversarial misuse, defensive integration, and security evaluation practices.

The proposed taxonomy organises the literature along four complementary dimensions: (i) AIGC-driven cyber threats, (ii) LLM-enabled defensive capabilities, (iii) security evaluation, red-teaming, and governance, and (iv) methodological orientation. Collectively, these dimensions capture not only *what* security problems are addressed, but also *how* evidence is generated, validated, and operationalised. This structure provides a reproducible analytical framework for comparative analysis, exposes systematic imbalances in the existing research landscape, and supports the identification of underexplored yet high-impact research directions.

### *4.1 Rationale for the Proposed Taxonomy*

The motivation for the proposed taxonomy is threefold. First, existing surveys frequently examine offensive misuse, defensive applications, or ethical considerations of LLMs in isolation, resulting in fragmented insights that obscure the co-evolution of threats and countermeasures [33,34]. By explicitly integrating threat, defence, and evaluation dimensions, the proposed taxonomy captures their interdependencies and reflects the concurrent evolution of attacker innovation, defensive adaptation, and governance constraints.

Second, the taxonomy elevates *security evaluation and governance* to a primary analytical dimension rather than treating it as a secondary or peripheral concern. An expanding body of work demonstrates that deficiencies in robustness auditing, red-teaming, explainability, and regulatory alignment can undermine otherwise promising technical advances [36,38]. Explicitly foregrounding evaluation and governance therefore aligns the taxonomy with real-world deployment requirements and emerging regulatory expectations.

Third, the taxonomy is directly grounded in the textual, empirical, and methodological evidence of the 167 reviewed studies. Categories and subcategories were iteratively refined through frequency analysis, conceptual differentiation, and cross-mapping of hybrid contributions, ensuring analytical coherence without sacrificing completeness. This inductive grounding strengthens methodological transparency and avoids the conceptual overfitting that often characterises high-level AI taxonomies.

Fig. 2 presents a visual overview of the unified taxonomy, illustrating how threat classes, defensive mechanisms, evaluation frameworks, and methodological orientations collectively structure the emerging LLM–cybersecurity research landscape.

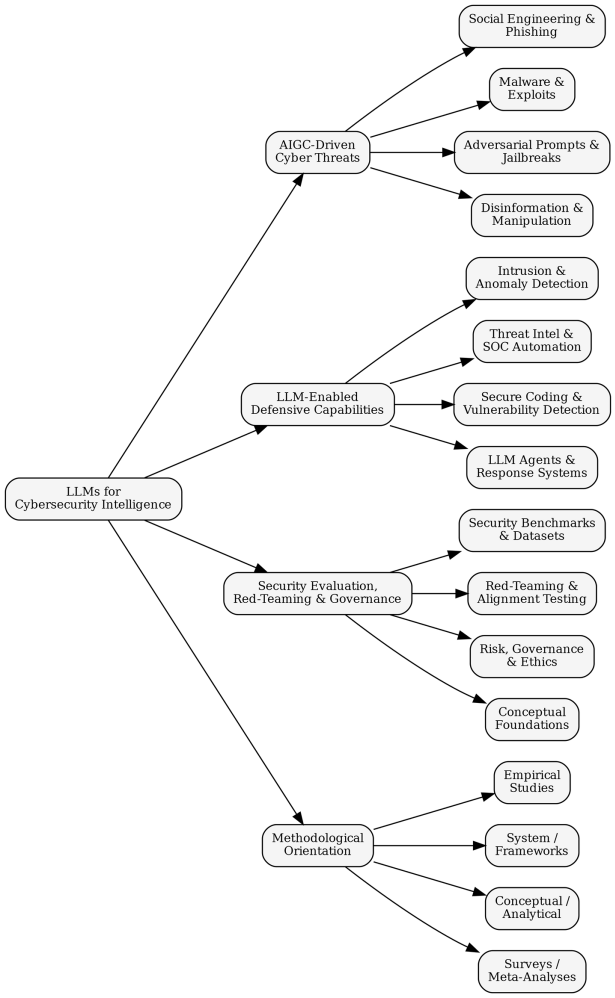


**Figure 2:** Proposed unified taxonomy of research on LLMs for cybersecurity intelligence.

### *4.2 AIGC-Driven Cyber Threats*

The first dimension of the taxonomy captures how generative AI and LLMs expand the cyber threat surface by automating deception, lowering technical skill barriers, and enabling adaptive attacks at scale. Across the reviewed literature, AIGC-driven threats consistently emerge as a dominant research focus, reflecting widespread concern regarding the misuse potential of generative models [41]. Unlike conventional cyber threats, these attacks exploit linguistic fluency, contextual awareness, and automation, enabling adversaries with limited expertise to execute sophisticated campaigns.

Empirical and analytical studies converge around three recurring threat clusters: (i) social engineering and deception automation, (ii) malware generation, exploit scaffolding, and code abuse, and (iii) model manipulation, prompt injection, and alignment evasion. These clusters recur across security-focused surveys, AI safety analyses, and governance-oriented discussions [42,43], supporting their treatment as distinct yet interrelated subcategories.

#### *4.2.1 Social Engineering and Deception Automation*

A substantial portion of the literature documents how LLMs amplify the scale, realism, and adaptability of social engineering attacks, frequently described as the "dual-edged sword" effect of generative AI [44]. Numerous studies demonstrate that models such as ChatGPT can generate linguistically fluent and context-aware messages that closely mimic professional or personal communication styles, thereby eroding traditional detection cues used in phishing and impersonation attacks [33,45].

Beyond static message generation, interactive and multi-turn deception has emerged as a critical escalation vector. Esmradi et al. [34] show how adaptive conversational strategies enable sustained manipulation, while Lopez et al. [35] and Nawara and kashef [36] highlight risks associated with platform-integrated systems, including chatbots, recommender engines, and customer-service pipelines. Collectively, this body of work indicates a shift from isolated phishing messages toward ecosystem-level influence operations embedded within digital services.

#### *4.2.2 Malware, Exploit Scaffolding, and Code Abuse*

A second threat cluster concerns LLM-assisted code generation and its associated dual-use implications. While many studies emphasise the defensive benefits of automated code synthesis, a consistent theme is the risk that generative models may produce insecure logic, reproduce known vulnerabilities, or facilitate exploit construction under ambiguous or adversarial prompts [33,34]. By automating elements of exploit reasoning, LLMs reduce traditional expertise barriers and accelerate attack development cycles.

Although the reviewed corpus contains limited primary experimentation on live malware generation, conceptual and lifecycle-oriented analyses consistently conclude that LLMs can significantly enhance attacker capabilities, particularly when integrated with external knowledge bases or retrieval mechanisms [38,46]. These findings justify treating code abuse as a distinct threat class rather than as a marginal extension of social engineering.

#### *4.2.3 Runtime Malware Behaviour and OT/ICS Implications*

Recent studies increasingly recognise that an exclusive focus on text-centric threats underrepresents execution-stage risks. LLMs can assist in generating adaptive malware components, polymorphic payloads, and environment-aware attack logic that evolves at runtime [8,9]. Such threats exhibit increased persistence and lateral movement potential, particularly when targeting operational technology (OT) and industrial control systems (ICS).

Despite their potentially severe impact, OT and ICS contexts remain comparatively underexplored due to data scarcity, safety constraints, and experimental complexity [39,47]. Table 2 explicitly contrasts surface-level attacks with runtime and cyber-physical threats, highlighting a systematic evaluation gap that biases current research toward methodological convenience rather than operational risk.

**Table 2:** Comparative analysis of LLM-driven threat classes: surface-level vs. execution-stage and OT/ICS risks.

| Dimension | Phishing & Prompt-Based Attacks | Runtime Malware Behaviour | OT/ICS Cybersecurity Threats |
|---|---|---|---|
| **Primary Attack Surface** | Human users and LLM interfaces | Execution environment, OS, and runtime context | Cyber-physical processes and control logic |
| **Role of LLM** | Message generation, deception, prompt manipulation | Code synthesis, behavioural reasoning, evasion planning | Protocol inference, attack-path reasoning, CPS analysis |
| **Operational Impact** | Credential theft, misinformation, short-term compromise | Persistence, lateral movement, stealthy system compromise | Physical disruption, safety violations, service outages |
| **Adaptivity at Runtime** | Limited after deployment | High (environment-aware, polymorphic execution) | High (context- and process-aware attacks) |
| **Evaluation Maturity** | High (public datasets, benchmarks) | Moderate (curated malware corpora) | Low (limited datasets, safety constraints) |
| **Key Gap** | Overemphasis due to benchmark availability | Limited long-running and real-world evaluation | Severe underrepresentation despite high risk |

#### *4.2.4 Model Manipulation, Prompt Injection, and Alignment Evasion*

The third threat class encompasses attacks that directly target the behaviour and control mechanisms of large language models themselves. Techniques such as prompt injection, jailbreak attacks, and alignment evasion undermine built-in safety constraints and enable the elicitation of sensitive information or unsafe outputs [34,38]. These risks are further amplified in agentic and retrieval-augmented generation (RAG)-based systems, where injected instructions may propagate across interconnected components, triggering cascading failures and unintended downstream actions [35].

Collectively, the reviewed studies indicate that alignment remains fragile under sustained adversarial interaction, particularly in real-world deployments where LLMs interface with external tools, application programming interfaces (APIs), and untrusted user inputs. These findings underscore the need for robust input validation, isolation mechanisms, and continuous monitoring to mitigate systemic failure modes arising from model manipulation.

### *4.3 LLM-Enabled Defensive Capabilities*

Within the unified taxonomy, the second dimension captures how large language models are operationalised as *defensive instruments* to support cybersecurity monitoring, analysis, and response. In contrast to threat-centric research, this body of work primarily conceptualises LLMs as augmentative components embedded within existing security workflows rather than as fully autonomous decision-makers. Across the 167 reviewed studies, defensive applications consistently cluster into three capability classes: (i) threat intelligence extraction and security operations support, (ii) secure coding, vulnerability assessment, and risk prediction, and (iii) anomaly detection, behavioural modelling, and policy reasoning.

This categorisation reflects an important methodological distinction within the literature. Whereas AIGC-driven threats exploit the generative and adaptive properties of LLMs, defensive systems leverage their semantic reasoning, contextual understanding, and natural language interaction capabilities to reduce analyst workload, enhance interpretability, and support complex decision-making under uncertainty [48,49]. Nevertheless, as discussed in the following sections, defensive effectiveness remains strongly contingent on deployment constraints, governance mechanisms, and the rigor of evaluation protocols.

#### *4.3.1 Threat Intelligence Extraction and Security Operations Support*

The most mature and extensively investigated defensive application of LLMs concerns threat intelligence extraction and support for Security Operations Center (SOC) workflows. A recurring theme across the reviewed studies is the use of LLMs to summarise, contextualise, and correlate heterogeneous security artefacts, including logs, alerts, incident reports, and threat intelligence bulletins [50]. Nawara and kashef [36] demonstrate that LLM-driven recommendation and decision-support pipelines can uncover latent relationships among entities, reconstruct attack narratives, and surface actionable insights with minimal analyst prompting. Complementary studies on log analysis and alert triage [51] further show that such systems can reduce cognitive burden by transforming low-level telemetry into structured, high-level explanations.

Survey-oriented investigations of generative AI adoption in enterprise and digital platform contexts [52–54] consistently position LLMs as effective intermediaries between raw security data and human analysts. Lopez et al. [35] further highlight that LLM-enabled automation can accelerate response workflows by generating concise summaries, drafting mitigation recommendations, and interpreting alerts expressed in natural language. Architectures based on retrieval-augmented generation (RAG) explicitly address scalability and knowledge grounding by integrating semantic search with generative reasoning [55].

Despite these advantages, multiple studies caution that SOC-oriented deployments remain vulnerable to hallucinated correlations, overconfident explanations, and trust calibration failures when LLM outputs are consumed without systematic validation [33,56]. Consequently, the literature converges on the view that LLMs are most effective as analyst-support tools that augment, rather than replace, expert judgment in operational SOC environments.

#### *4.3.2 Secure Coding, Vulnerability Assessment, and Risk Prediction*

A second major class of defensive capability concerns the application of LLMs within secure software development lifecycles. Although code synthesis by generative models is widely recognised as a dual-use capability, a substantial body of work demonstrates that, when appropriately constrained, LLMs can assist in identifying insecure constructs, detecting coding antipatterns, and explaining potential vulnerabilities [57,58]. Esmradi et al. [34] show that LLMs can reason over code semantics and generate preliminary vulnerability explanations that complement traditional static analysis techniques.

Specialised studies further demonstrate defensive applications in firewall configuration, secure deployment pipelines, and adaptive policy enforcement [59,60]. Alawida et al. [33] argue that LLMs can function as intelligent static-analysis assistants, particularly in identifying subtle logic flaws that evade rule-based scanners.

However, Uddin et al. [38] emphasise that defensive gains in secure coding are inseparable from governance and lifecycle oversight. In the absence of systematic validation, continuous monitoring, and auditability, LLM-generated recommendations risk propagating insecure patterns at scale. Studies on model trustworthiness and auditing [61] further indicate that higher-level risk prediction tasks-such as vulnerability prioritisation and impact explanation-should be interpreted as decision-support signals rather

than authoritative outputs. Collectively, these findings position LLMs as promising yet fragile components within secure development workflows, whose effectiveness depends on rigorous verification and human oversight [62].

#### *4.3.3 Anomaly Detection, Behavioural Modelling, and Policy Reasoning*

Beyond intelligence extraction and code analysis, LLMs are increasingly explored for behavioural anomaly detection, system modelling, and policy reasoning. Alawida et al. [33] demonstrate that transformer-based models can capture semantic patterns in logs and behavioural traces, enabling the detection of subtle deviations associated with insider threats, fraud, or anomalous system states. Empirical systems such as ChatPhishDetector [63] and LLM-assisted malicious webpage detection frameworks [64] further illustrate the feasibility of applying LLMs to text-rich security contexts.

At a higher level of abstraction, Lopez et al. [35] highlight the suitability of LLMs for policy reasoning tasks, including interpreting configuration files, validating compliance requirements, and translating complex security policies into human-readable explanations. Such capabilities support automated auditing and continuous compliance monitoring, particularly in specialised domains such as transport and critical infrastructure systems.

Uddin et al. [38] further argue that behavioural modelling and policy reasoning are especially critical in multi-agent or retrieval-augmented systems, where LLMs mediate interactions between user intent, system state, and downstream actions. When properly aligned, these models can detect inconsistencies, enforce safety constraints, and reason about risk propagation across interconnected components [65]. Nevertheless, the literature consistently cautions that such reasoning capabilities must be bounded by explicit control logic to prevent cascading errors or unsafe autonomous behaviour.

#### *4.3.4 Comparative Synthesis of LLM-Enabled Defensive Capabilities*

A cross-cutting synthesis of LLM-enabled defensive studies indicates that, despite notable task-level performance gains, defensive effectiveness is fundamentally determined by *how* LLMs are embedded within security workflows rather than by model capability alone. Across threat intelligence support, secure coding, and anomaly detection, LLMs consistently function as *cognitive amplifiers*, enhancing semantic understanding, contextual correlation, and explanation, while remaining dependent on classical detection engines, rule-based controls, and human oversight for operational reliability [35,38,48].

Within SOC and threat intelligence pipelines, LLMs excel at unifying heterogeneous data sources and generating analyst-oriented narratives, yielding measurable reductions in triage time and cognitive workload [36,51]. These benefits, however, are counterbalanced by susceptibility to hallucinated correlations and overconfident reasoning, particularly in retrieval-augmented generation (RAG) settings where corpus incompleteness or temporal drift undermines grounding [33,56]. As a result, the literature consistently advocates hybrid SOC architectures in which LLMs serve advisory roles and validation layers are mandatory.

In secure coding and vulnerability assessment, LLMs demonstrate strong semantic reasoning capabilities, often outperforming purely syntactic static-analysis tools in explaining logic flaws, insecure dependencies, and exploit preconditions [34]. Nevertheless, empirical evidence also highlights the risk of *scaled insecurity*, whereby unverified LLM recommendations propagate flawed remediation strategies across large codebases [38,57]. Consequently, studies recommend positioning LLMs as assistive auditors integrated with formal verification, testing pipelines, and developer review processes rather than as autonomous security agents [61,62].

For anomaly detection, behavioural modelling, and policy reasoning, LLMs are most effective in text-rich or semantically complex environments—such as log analysis, phishing detection, and compliance interpretation—where traditional feature-based models exhibit limitations [33,63]. However, latency, computational overhead, and sensitivity to input formatting constrain their suitability for real-time detection and control-loop enforcement [64]. In multi-agent and policy-driven systems, LLM-based reasoning enhances interpretability and coordination but introduces new failure modes, including cascading errors and ambiguous policy interpretation, necessitating explicit constraint enforcement and human-in-the-loop safeguards [38,65].

Overall, the comparative evidence indicates that LLM-enabled defensive systems are *most robust when deployed as bounded, explainable, and governable components within layered defence architectures*. Performance gains are strongest in decision support, correlation, and explanation tasks, whereas autonomous detection or response remains high risk without rigorous validation, continuous monitoring, and governance oversight. Table 3 summarises these trade-offs across representative defensive domains.

**Table 3:** Comparative synthesis of LLM-enabled defensive capabilities.

| Defensive Domain | Role of LLMs | Observed Benefits | Limitations and Risks |
|---|---|---|---|
| **Threat Intelligence & SOC Support** | Semantic summarisation, correlation, analyst decision support | Faster triage; reduced cognitive load; improved cross-source context [36,51] | Hallucinations; RAG grounding errors; trust calibration issues [33,56] |
| **Secure Coding & Vulnerability Analysis** | Code understanding, vulnerability explanation, remediation guidance | Detects logic flaws; complements static analysis; improves interpretability [34] | Insecure fix propagation; lack of formal guarantees; auditability required [38,62] |
| **Anomaly Detection & Behavioural Modelling** | Log semantics, phishing detection, behavioural interpretation | Effective for text-rich and semantic anomalies [33,63] | High latency; compute overhead; fragile to malformed inputs [64] |
| **Policy Reasoning & Compliance Support** | Policy interpretation, configuration analysis, audit explanation | Supports explainable compliance and continuous auditing [35] | Ambiguous policy semantics; inconsistent enforcement risk [65] |
| **Overall Insight** | Bounded cognitive augmentation | Strong reasoning and explanation support | Unsafe as autonomous security decision-makers without oversight [38] |

### *4.4 Security Evaluation, Red-Teaming, and Governance*

The third dimension of the unified taxonomy encompasses research focused on evaluating the trustworthiness, robustness, and governance readiness of LLM-enabled cybersecurity systems. Unlike threat- or defence-centric studies, which emphasise capability expansion, this dimension interrogates whether such systems can be *safely, reliably, and legally deployed* in security-sensitive environments. Across the 167 analysed studies, evaluation and governance emerge as indispensable complements to both offensive and defensive research, reflecting a growing consensus that unregulated or inadequately assessed LLM integration may introduce systemic risks that outweigh operational benefits [66–68].

Within the taxonomy, this dimension is structured around three interrelated themes: (i) security benchmarking and evaluation frameworks, (ii) red-teaming, misuse detection, and lifecycle risk assessment, and (iii) governance, regulatory alignment, and enforceable controls. Together, these themes address not only *whether* LLMs perform as intended, but also *how* failures emerge, propagate, and can be constrained under real-world operational and regulatory conditions.

#### *4.4.1 Security Benchmarking and Evaluation Frameworks*

A central finding across the reviewed literature is the absence of standardised, security-oriented evaluation frameworks for LLMs. Lopez et al. [35] identify multiple evaluation dimensions, including behavioural stress testing, trustworthiness analysis, and safety auditing, but emphasise that prevailing metrics (e.g., accuracy or BLEU-style scores) are insufficient to capture security-relevant failure modes such as context-sensitive hallucinations, brittle reasoning, and inconsistent safety responses. This critique is reinforced by broader surveys of AI evaluation practices, which consistently highlight the lack of unified, comparable, and deployment-relevant assessment standards [69–71].

Esmradi et al. [34] further argue that evaluation pipelines must explicitly assess susceptibility to adversarial prompting, data leakage, insecure code generation, and alignment failures. In the absence of such targeted benchmarks, empirical claims regarding LLM safety and robustness remain difficult to validate or compare across studies. Alawida et al. [33] similarly identify robustness evaluation as a critical bottleneck, observing that many proposed defence mechanisms are tested only under benign or narrowly scoped experimental conditions.

Recent investigations of specialised security tools [72] and analyses of LLM application ecosystems further underscore the operational consequences of weak evaluation practices, particularly in settings characterised by rapid deployment and frequent model updates. Across the reviewed literature, evaluation is increasingly conceptualised not as a one-time verification activity but as a continuous, lifecycle-spanning process that underpins responsible deployment, model governance, and risk-aware decision-making [73].

#### *4.4.2 Red-Teaming, Misuse Detection, and Lifecycle Risk Assessment*

Red-teaming is widely recognised as a critical mechanism for uncovering latent vulnerabilities in LLM-enabled systems. Although the reviewed corpus contains relatively few large-scale empirical red-teaming studies, numerous analytical and survey-oriented works provide detailed examinations of misuse pathways, adversarial prompting strategies, and system-level failure cascades. Uddin et al. [38] present one of the most comprehensive lifecycle-oriented risk assessments, synthesising vulnerabilities across data acquisition, model training, deployment, tool integration, and post-deployment adaptation [74,75].

These analyses consistently reveal that risks often arise not from isolated model behaviour but from interactions between LLMs and surrounding system components. In multi-agent or retrieval-augmented architectures, injected prompts or poisoned contextual data can propagate across subsystems, resulting in cascading failures or unsafe actions [76]. Complementing this perspective, Esmradi et al. [34] catalogue a broad range of adversarial attack surfaces—including jailbreaks, prompt chaining, inference-time manipulation, and information leakage probes—that should form the foundation of systematic red-teaming methodologies. Broader surveys of LLM-enabled cyberattacks [43] further reinforce the need for adversarial evaluation strategies that extend beyond single-prompt testing.

Alawida et al. [33] additionally emphasise that effective misuse detection requires continuous behavioural monitoring, trust calibration, and post-hoc interpretability mechanisms to identify anomalous or unsafe reasoning trajectories. Collectively, these studies converge on the conclusion that red-teaming must

evolve from ad hoc adversarial prompting toward structured, repeatable lifecycle assessments that evaluate model behaviour under uncertainty, interaction, and environmental variation [77].

#### *4.4.3 Governance, Regulatory Alignment, and Enforceable Controls*

Governance and regulatory considerations constitute a critical cross-cutting pillar of LLM-enabled cybersecurity research, linking technical evaluation with accountability, compliance, and operational safety. While earlier studies extensively discuss ethical concerns such as bias, transparency, privacy, and misuse [78,79], much of this literature remains principle-driven and insufficiently connected to enforceable regulatory requirements or operational security workflows. More recent critical analyses argue that ethical risk mitigation cannot be decoupled from technical robustness and governance enforcement, particularly when LLMs influence security-critical decisions [38,80,81].

A recurring concern across the reviewed studies is the absence of clearly defined accountability structures for LLM-driven systems deployed in high-stakes contexts, including SOC automation, intrusion detection, and incident response. Uddin et al. [38] highlight deficiencies related to explainability, responsibility allocation, and auditability, concluding that governance mechanisms lag behind the pace of operational integration. Similar observations are reported in privacy-sensitive and manipulative-content domains, where ethical shortcomings translate directly into legal, reputational, and operational risks.

From a regulatory perspective, the EU NIS-2 Directive establishes binding requirements for cybersecurity risk management, incident handling, supply-chain security, and organisational accountability. Many failure modes identified in LLM-enabled cybersecurity systems—including hallucinated threat attribution, false-positive escalation within SOC pipelines, prompt-injection vulnerabilities in RAG architectures, and unsafe autonomous responses—map directly to NIS-2 obligations concerning proportional controls, resilience, and service continuity. Consequently, LLM components that materially influence detection, triage, or response processes fall squarely within the scope of regulated cybersecurity operations.

Complementary guidance issued by the U.S. Cybersecurity and Infrastructure Security Agency (CISA) emphasises defence-in-depth, layered safeguards, and verifiable controls for AI-enabled systems deployed in SOCs and critical infrastructure environments. Consistent with this guidance, the literature increasingly converges on the view that LLMs should function as constrained components within hybrid defence architectures, augmenting classical detection and policy engines rather than acting as autonomous decision-makers [35,82]. Empirical evidence further indicates that hallucinations and reasoning errors can propagate rapidly when LLM outputs are granted excessive operational authority.

Operationalising governance therefore requires translating ethical principles into auditable and enforceable controls. Across the reviewed studies, recurring mechanisms include mandatory human-in-the-loop checkpoints, role-based access control for prompts and policy modification, immutable audit logging, model versioning and change management, and clear separation of duties between detection, reasoning, and response components [83,84]. These controls directly mitigate the operational risks identified in the defensive and evaluation analyses and align with both NIS-2 accountability requirements and CISA deployment guidance.

Table 4 synthesises these insights by explicitly mapping identified LLM security risks to concrete technical and organisational controls, along with corresponding regulatory or standards-based obligations. By grounding governance in enforceable mechanisms, this taxonomy dimension reframes ethics from aspirational guidance into compliance-ready practice, providing actionable direction for researchers, practitioners, and policymakers.

**Table 4:** Mapping of LLM security risks to enforceable controls and regulatory alignment.

| Identified Risk | Enforceable/Auditable Control | Regulatory or Standards Alignment |
|---|---|---|
| Hallucinated threat attribution or IOC generation | Human-in-the-loop validation; confidence thresholds; analyst override logging | NIS-2 (risk management, incident handling); CISA Secure AI Guidance |
| False-positive escalation in SOC automation | Escalation approval workflows; false-positive auditing metrics | NIS-2 (service continuity); CISA SOC Best Practices |
| Prompt injection in RAG pipelines | Prompt sanitisation; retrieval whitelisting; immutable context logs | NIS-2 (supply-chain security); CISA AI Supply Chain Risk Management |
| Unsafe autonomous response actions | Mandatory human approval; separation of detection and response | NIS-2 (accountability); CISA Zero Trust principles |
| Model behaviour drift | Model versioning; regression testing; change management | NIS-2 (governance); ISO/IEC 27001 |
| Lack of explainability | Rationale logging; post-incident review artefacts | NIS-2 (auditability); CISA IR documentation |
| Adversarial misuse of LLM tools | RBAC; usage monitoring; anomaly detection | NIS-2 (access control); CISA insider threat mitigation |
| OT/ICS deployment risks | Decision-support-only deployment; fallback to classical controls | NIS-2 (critical infrastructure); CISA ICS advisories |

### *4.5 Methodological Orientation within the Taxonomy*

Beyond thematic classification, the proposed taxonomy explicitly incorporates *methodological orientation* as a cross-cutting analytical dimension. This inclusion is essential for interpreting not only which aspects of LLM-enabled cybersecurity are examined, but also how evidence is generated, validated, and generalised. An analysis of the 167 included studies reveals four dominant methodological orientations—**empirical**, **system-oriented**, **conceptual/analytical**, and **survey-based**—which collectively shape the maturity, reliability, and operational relevance of the field.

These methodological orientations intersect all three substantive taxonomy dimensions (AIGC-driven threats, LLM-enabled defensive capabilities, and security evaluation and governance), thereby influencing the strength of reported claims, the reproducibility of findings, and the feasibility of real-world deployment. Explicitly embedding methodology within the taxonomy prevents the conflation of conceptual insight with empirical validation and enables a more nuanced and evidence-aware synthesis of the literature.

#### *4.5.1 Empirical Evaluation-Oriented Studies*

Empirical studies constitute a substantial portion of the reviewed literature and focus on observing model behaviour, quantifying robustness, and analysing failure modes through controlled experiments or

simulations. Representative examples include empirical investigations of hallucination, unsafe code generation, exploit reasoning, and misalignment effects reported by Esmradi et al. [34]. Large-scale comparative assessments of AI-generated code security [85] and experimental evaluations of ChatGPT on log and security datasets [77] further exemplify this orientation.

While empirical studies are indispensable for grounding claims regarding LLM capabilities and risks, a recurring limitation across the corpus is the absence of standardised benchmarks, limited reproducibility, and evaluation under constrained or synthetic conditions. These limitations impede cross-study comparability and restrict generalisation to operational cybersecurity environments. As a result, many empirical findings provide primarily *local validity* rather than system-level assurance, underscoring the need for the unified evaluation frameworks discussed in Dimension III.

#### *4.5.2 System and Architectural Proposals*

System-oriented studies represent a second major methodological category and focus on integrating LLMs into operational cybersecurity workflows. These contributions typically propose architectures, pipelines, or decision-support systems that embed LLMs within SOC operations, threat intelligence platforms, or automated response mechanisms. Nawara and kashef [36], for example, present LLM-driven recommendation systems for contextualising threat intelligence and streamlining analyst workflows. Related efforts include retrieval-augmented cybersecurity intelligence frameworks [55], real-time crime detection systems [56], and scalable log analysis pipelines [51].

Although these system-level contributions signal progress toward operational deployment, they are frequently evaluated under constrained conditions, rely on curated or synthetic inputs, or omit lifecycle-level threat modelling. Consequently, many architectural proposals demonstrate functional feasibility without sufficiently addressing robustness, adversarial resilience, or governance constraints. This methodological shortcoming reinforces the importance of coupling system design with rigorous evaluation, red-teaming, and governance analysis.

#### *4.5.3 Conceptual and Analytical Studies*

Conceptual and analytical works form a third methodological orientation, offering taxonomies, risk models, and socio-technical analyses that articulate the broader implications of LLM integration within cybersecurity ecosystems [67]. Uddin et al. [38] provide a comprehensive lifecycle-oriented risk framework that identifies vulnerabilities spanning data provenance, model training, deployment, tool integration, and governance. Other analytical contributions highlight systemic risks, including the erosion of Zero-Trust assumptions induced by generative AI [75] and the cascading effects associated with large-scale LLM adoption [74].

These studies play a critical role in exposing interdependencies and long-term risks that may not be apparent in isolated empirical evaluations. However, their primary limitation lies in limited empirical grounding, as many conceptual insights remain insufficiently validated through controlled experimentation or real-world case studies. Within the taxonomy, such analyses therefore function primarily as instruments of *risk foresight* rather than as evidence of deployable solutions.

#### *4.5.4 Surveys and Meta-Analyses*

Surveys and meta-analyses constitute the fourth methodological category, synthesising research trends across generative AI, cybersecurity threats, defensive mechanisms, and governance considerations. Foundational surveys by Alawida et al. [33] and Lopez et al. [35] provide broad overviews of AIGC risks, defensive opportunities, and ethical challenges. Additional systematic and narrative reviews further contextualise LLM applications in cybersecurity [40,49], their dual-use implications in network security [86], and recent advancements in generative AI [71].

While surveys play a crucial role in consolidating fragmented research, many existing reviews rely predominantly on narrative synthesis and lack systematic protocols such as PRISMA. This limits transparency, reproducibility, and comparability across surveys. The present work addresses this limitation by embedding methodological orientation directly within a PRISMA-guided taxonomy, enabling structured cross-comparison beyond descriptive aggregation.

#### *4.5.5 Methodological Imbalances and Implications*

Taken together, the distribution of methodological orientations reveals a research landscape that is broad yet uneven. Empirical studies frequently lack unified benchmarks and stress-testing protocols; system-oriented proposals often under-evaluate adversarial and governance risks; conceptual analyses are weakly anchored in experimental validation; and surveys expose inconsistencies in terminology, threat definitions, and evaluation practices. These imbalances directly influence the maturity and reliability of conclusions drawn across all three substantive taxonomy dimensions.

Beyond methodological categorisation, the distribution of studies across the taxonomy (Table 5) highlights several longitudinal trends between 2022 and 2025. Research on AIGC-driven threats remains dominant, particularly in social engineering and adversarial prompting [44,65]. Defensive applications—especially threat intelligence extraction and SOC automation—demonstrate increasing operational focus [36], yet remain unevenly validated. Notably, a growing concentration of studies addresses security evaluation, red-teaming, and governance, reflecting increasing recognition that trustworthiness and accountability are decisive factors for real-world deployment [35,38].

Overall, explicitly positioning methodological orientation within the taxonomy transforms it from a descriptive classification into a critical analytical framework. By revealing how evidence is produced—and where it is systematically weak—this dimension provides a foundation for identifying research gaps, guiding standardisation efforts, and supporting evidence-based advancement of LLM-driven cybersecurity intelligence.

Table 5 consolidates all 167 studies within the unified taxonomy, linking thematic focus, methodological orientation, strengths, and limitations. This synthesis underpins the comparative and longitudinal analyses presented in subsequent sections.

## 5 AIGC-Driven Threats Enabled by LLMs

Building on the unified threat–defence–evaluation taxonomy introduced in Section 4, this section examines how large language models (LLMs) operate as *offensive enablers* within contemporary cyber threat ecosystems. Across the 167 studies synthesised in this review, AIGC-enabled attacks consistently emerge not as incremental extensions of existing techniques, but as a qualitative escalation in adversarial capability. LLMs fundamentally alter attacker economics by lowering expertise barriers, automating cognitively complex tasks, and enabling scalable, adaptive, and context-aware malicious behaviour [87–89].

Within the proposed taxonomy, AIGC-driven threats cluster into four dominant and recurrent categories: (i) social engineering and multimodal deception, (ii) malware generation, exploit scaffolding, and runtime behaviour, (iii) prompt injection, jailbreaks, and alignment evasion, and (iv) misinformation, disinformation, and influence operations. These categories are derived inductively through cross-study coding and reflect convergent patterns observed across empirical evaluations, red-teaming analyses, and large-scale surveys. Collectively, they expose the inherent dual-use tension of LLMs, wherein the same generative, reasoning, and contextualisation capabilities that enable defensive innovation are systematically repurposed for adversarial misuse [42,44,90].

### *5.1 Social Engineering and Multimodal Deception*

Social engineering represents the most mature and empirically substantiated class of AIGC-driven threats. A substantial body of literature demonstrates that LLMs enable scalable, highly personalised phishing, impersonation, and deception campaigns characterised by linguistic coherence, contextual sensitivity, and adaptive tone control previously associated with skilled human operators [45,91]. Unlike template-based automation, LLM-driven pipelines dynamically tailor content using inferred organisational roles, conversational context, and cultural cues, thereby significantly increasing attack credibility [33,34,92].

Empirical evaluations reported in [25,93,94] indicate that LLM-generated spear-phishing and business email compromise (BEC) messages consistently degrade the effectiveness of signature-based filters and stylometric detection techniques. Comparative studies further reveal that semantic diversity, multilingual generation, and adversarial paraphrasing undermine feature-driven classifiers, exacerbating the asymmetry between attacker adaptability and defender rigidity [35,95].

Beyond text-only attacks, recent work documents a marked shift toward *multimodal deception*, wherein adversaries exploit systems capable of jointly reasoning over text, images, code, and structured artefacts. Such cross-modal manipulation produces inconsistencies that evade unimodal detection pipelines [96,97]. Collectively, these findings indicate that social engineering has evolved from isolated phishing attempts into adaptive, multimodal deception workflows that challenge foundational assumptions of existing detection systems.

**Table 5:** Comparative taxonomy-based classification of included studies with methodological and operational synthesis.

| Category | Sub–Category | Core Technical Focus | Methodological Orientation | Strengths/ Advantages | Limitations/ Failure Modes | Primary Application Context | Representative Studies |
|---|---|---|---|---|---|---|---|
| **Dimension I: AIGC-Driven Cyber Threats** | | | | | | | |
| Social Engineering & Deception Automation | LLM-generated phishing and impersonation | High-fidelity multilingual phishing, role-adaptive deception | Empirical/ System/Survey | Scalable social engineering, linguistic realism | Prompt sensitivity, cultural bias, rapid counter-adaptation | Email security, messaging platforms | T1 = [5–7,13,14,25,33–36,44,45,92–94,98–104] |
| Malware & Exploit Generation Risks | Automated malicious code synthesis | Exploit reasoning, polymorphic malware generation | Empirical/ Conceptual/Survey | Accelerated exploit development | Limited runtime grounding, detectable artefacts | Malware pipelines, red-team tooling | T2 = [8–10,19,33–35,37,38,46,105–115] |
| Adversarial Prompting, Jailbreaks & Injection | Alignment bypass and prompt manipulation | Role-based jailbreaks, multi-step misalignment | Empirical/ Conceptual/Survey | Reveals alignment weaknesses | Brittle defences, prompt overfitting | LLM safety testing, agentic systems | T3 = [15,26,27,32,34,35,38,111,116–121] |
| Disinformation & Manipulation | Synthetic misinformation and propaganda | Narrative steering, watermark evasion | Empirical/ Analytical/Conceptual | High-scale influence operations | Attribution difficulty, detection lag | Media platforms, information ecosystems | T4 = [13,14,36,101,122–127] |
| **Dimension II: LLM-Enabled Defensive Capabilities** | | | | | | | |
| Intrusion, Malware & Anomaly Detection | Semantic and behavioural analysis | Log reasoning, malware trace inference | Empirical/ System/Survey | Context-aware detection | False-positive escalation, dataset bias | Enterprise IDS, SOC monitoring | D1 = [19,27,33,35,38,63,64,102,103,105,106,108,110,112–114,128,129] |
| Threat Intelligence & SOC Automation | CTI extraction and alert triage | RAG-driven intelligence, summarisation | System/ Empirical/Survey | Reduced analyst workload | Hallucinated correlations, trust calibration | SOC co-pilots, incident response | D2 = [20,23,24,33,35,36,48,50,55,56,94,124,130–132] |

(Continued)

**Table 5 (continued)**

| Category | Sub–Category | Core Technical Focus | Methodological Orientation | Strengths/ Advantages | Limitations/ Failure Modes | Primary Application Context | Representative Studies |
|---|---|---|---|---|---|---|---|
| Secure Coding & Vulnerability Detection | Code analysis and patch generation | Semantic reasoning, structured prompting | Empirical/ System/Analytical | Improved vulnerability localisation | Patch correctness uncertainty | Secure SDLC, DevSecOps | D3 = [19,22,23,27,33,34, 38,57,59,62,93,107,112,117, 133,134] |
| LLM Agents & Automated Response Systems | Multi-agent autonomous defence | Task delegation, chain-of-thought orchestration | System/ Conceptual/Survey | End-to-end workflow automation | Coordination overhead, cascading failures | Autonomous SOC, large-scale networks | D4 = [3,24,33,35,38,60, 65,129–131,135,136] |
| **Dimension III: Security Evaluation, Red-Teaming & Governance** | | | | | | | |
| Security Benchmarks & Evaluation Datasets | LLM security testing datasets | Robustness scoring, cross-model comparison | Empirical/ Survey | Standardised evaluation | Synthetic bias, limited realism | Benchmarking, academic evaluation | E1 = [3,6,25,26,32–35,70,72,137,138] |
| Red-Teaming & Alignment Stress-Testing | Adversarial probing | Multi-turn jailbreak, role-based attacks | Empirical/ Analytical/Conceptual | Failure-mode discovery | Coverage gaps, non-repeatability | LLM safety assurance | E2 = [15,26,32–34,38,43,76,77,85,118,139] |
| Risk, Governance & Ethical Frameworks | AI risk and compliance | Auditability, lifecycle governance | Conceptual/ Analytical/Survey | Policy alignment | Limited enforceability | Regulatory governance | E3 = [4,13,28,33,35,38, 80,81,95,140–144] |
| Conceptual Foundations & Surveys | Taxonomies and surveys | Foundational perspectives | Survey/ Conceptual | Holistic understanding | Limited empirical grounding | Research synthesis | E4 = [33–35,40,49, 66,79,82,89,92, 96,109] |

Note: T1–T4, D1–D4, and E1–E4 denote grouped citation clusters mapped to the unified taxonomy.

### 5.2 Malware, Exploit Scaffolding, and Runtime Behaviour

The second major threat category encompasses LLM-assisted malware generation, exploit scaffolding, and execution-stage behaviour. Across the reviewed corpus, multiple studies demonstrate that LLMs can generate functional malware components, exploit templates, and obfuscation patterns, particularly when safety mechanisms are bypassed through adversarial prompting or indirect instruction leakage [33,109]. Although many generated artefacts are not immediately deployable, they substantially accelerate exploit development cycles and reduce the expertise required for iterative attack refinement [34,145].

Several studies [108,110,111] highlight the dual-use risks inherent to LLM-assisted code generation, showing that models designed for benign development support can emit vulnerable or exploitable logic under ambiguous or underspecified prompts. The integration of structured cybersecurity knowledge bases further amplifies this threat by enabling attackers to rapidly query, adapt, and weaponise technical content at scale [46,104].

A critical limitation identified across this literature is the systematic underrepresentation of *runtime behaviour*. Most studies validate exploit generation at the code-fragment or proof-of-concept level, without examining execution dynamics, environmental dependencies, or long-term behavioural adaptation [38]. Consequently, current evaluations likely underestimate the persistence, lateral movement, and evasion capabilities enabled by LLM-assisted malware, underscoring the need for execution-aware threat modelling and system-level validation.

### 5.3 Prompt Injection, Jailbreaks, and Alignment Evasion

A third and rapidly expanding threat class targets LLMs directly through adversarial prompting, jailbreaks, and prompt injection. Extensive empirical evidence demonstrates that contemporary models remain highly susceptible to role-play manipulation, instruction obfuscation, multi-turn coercion, and indirect prompt injection attacks [32,34,111]. Notably, several studies suggest that larger and more capable models may exhibit increased vulnerability due to richer representational capacity and broader behavioural generalisation [116,142].

Importantly, these vulnerabilities extend beyond standalone models to integrated systems. Research on retrieval-augmented generation (RAG) and multi-agent architectures shows that injected prompts can propagate across components, triggering unsafe tool invocation, policy override, or cascading reasoning failures [27,38,118,146]. Across the surveyed literature, no single mitigation strategy—including prompt sanitisation, refusal tuning, or system prompt hardening—consistently defends against the full spectrum of adversarial prompting techniques. This exposes a structural fragility in current alignment approaches and motivates lifecycle-aware evaluation, system-level isolation, and governance-driven safeguards [147,148].

### 5.4 Misinformation and Influence Operations

LLM-enabled misinformation and influence operations constitute the fourth major threat category and represent one of the most societally consequential forms of AIGC-driven attack. Studies consistently demonstrate that LLMs can generate persuasive, culturally adaptive, and emotionally framed narratives at scale that are often indistinguishable from human-authored content [124,149]. Comparative analyses show that LLM-driven campaigns outperform traditional disinformation efforts in adaptability, multilingual reach, and narrative coherence [13,150].

A recurring observation across this literature is the fragility of existing detection and attribution mechanisms. Watermarking, stylometric analysis, and classifier-based approaches degrade substantially under paraphrasing, translation, and multi-model transformation pipelines [125,149]. Moreover, algorithmic

recommendation systems may unintentionally amplify harmful narratives, blurring the boundary between deliberate influence operations and emergent systemic manipulation [36,127]. Despite their high potential impact, longitudinal societal studies and deployment-scale evaluations remain limited, constraining understanding of sustained real-world effects.

### *5.5 Synthesis of Threat Findings*

Synthesising evidence across all four threat categories reveals a consistent and concerning trajectory: AIGC-enabled threats are advancing more rapidly than corresponding defensive countermeasures. In social engineering, semantic diversity and contextual realism undermine static detection heuristics. In malware generation, LLMs accelerate exploit development while enabling adaptive runtime behaviour. In adversarial prompting, alignment mechanisms remain brittle under sustained or system-level attacks. In misinformation, generative realism and scale overwhelm existing attribution and moderation strategies.

Three cross-cutting insights emerge from the reviewed corpus:

- **Escalating attacker capability:** LLMs substantially reduce the expertise and resources required to conduct complex cyberattacks, effectively democratising sophisticated offensive techniques.
- **Evaluation realism gap:** Heavy reliance on synthetic datasets and constrained experimental settings obscures the true operational severity of AIGC-driven threats.
- **Absence of comprehensive mitigation:** No existing defensive framework robustly addresses the full spectrum of LLM-enabled attack vectors, particularly in multimodal, agentic, and continuously evolving systems.

These findings underscore the need for execution-aware threat modelling, lifecycle-oriented evaluation, and defence architectures that explicitly account for the dual-use nature of generative AI. Tables 6 and 7 consolidate comparative evidence across threat categories and representative studies, demonstrating broad convergence on a central conclusion: AIGC-driven threats represent a systemic and accelerating challenge that existing security controls are structurally ill-equipped to contain.

**Table 6:** Comparative synthesis of AIGC-driven threats enabled by LLMs.

| Threat Category | Key References | Observed Capabilities | Empirical Gaps and Constraints | Severity |
|---|---|---|---|---|
| **Phishing & Social Engineering** | [6,25,33,35,44,100] | Highly personalised and multilingual phishing; adaptive tone and narrative framing; strong evasion of signature-based filters | Predominantly synthetic datasets; weak longitudinal analysis; detector degradation under paraphrasing and cross-lingual transfer | **High** |
| **Malware & Exploit Generation** | [8,10,34,38,111] | Exploit scaffolding and polymorphic code generation; semantic obfuscation; assistance in malware reasoning and triage | Limited end-to-end validation; runtime and environmental constraints often ignored; inconsistent safety enforcement | **High** |
| **Jailbreaks & Prompt Injection** | [15,27,32,116,117] | Reliable bypass of alignment controls; indirect prompt injection in RAG and agentic systems; cascading policy failures | No robust universal defence; benchmarks ignore lifecycle and update dynamics; defences brittle under adaptation | **Critical** |
| **Misinformation & Influence Operations** | [13,14,38,124,149] | Scalable persuasive content; cultural and emotional adaptation; cross-lingual narrative transfer | Scarce real-world validation; fragile watermarking; unreliable attribution; unclear long-term impact | **High–Critical** |

**Table 7:** Comparative analysis of AIGC-driven threat studies with methodological and operational insights.

| Study | Threat Subcategory | Model/Setting | Methodological Distinction | Demonstrated Capability/Advantage | Observed Limitations/Failure Modes | Applicable Threat Scenarios |
|---|---|---|---|---|---|---|
| **A. Automated Phishing and Social Engineering** | | | | | | |
| Opara et al. [6] | Phishing Generation | GPT-family models | Stylometry-aware text generation | Bypasses enterprise spam filters via linguistic mimicry | Limited multilingual and cross-cultural testing | Corporate email phishing |
| Schmitt and Flechais [13] | Deception Narratives | Instruction-tuned LLMs | Narrative-level persuasion modelling | High deception realism in controlled human studies | No deployment-scale or adversarial testing | Psychological manipulation campaigns |
| Alawida et al. [33] | Threat Survey | Review | Cross-domain threat aggregation | Highlights lowered entry barrier for attackers | Lacks empirical validation | Strategic threat assessment |
| De Queiroz [45] | Targeted Phishing | LLM simulations | High-volume personalised phishing | Cost-efficient, scalable attack generation | Text-only focus; no multimodal cues | Mass phishing operations |
| Nawara and Kashef [36] | Systemic Manipulation | LLM recommender systems | Feedback-loop analysis | Demonstrates narrative amplification risks | Primarily conceptual analysis | Algorithmic content manipulation |
| Aseeri and Bohacek [7] | Spear-Phishing | Contextual prompting | Social-media-aware attack modelling | Higher predicted victim engagement | No real victim interaction data | Targeted executive phishing |
| Pham et al. [98] | Enterprise Phishing | Fine-tuned LLMs | Role-aware training | Outperforms human templates | Domain-specific dataset bias | Corporate SOC testing |
| Koide et al. [99] | Filter Evasion | LLM paraphrasing | Lexical diversity exploitation | Defeats heuristic spam filters | Semantic risk not evaluated | Spam filter evasion |
| **B. Malware and Exploit Generation** | | | | | | |
| Yamin et al. [9] | Malware Generation | GPT models | Code synthesis with obfuscation | Functional malware-like behaviour | Incomplete payload execution | Proof-of-concept malware |
| Shandilya et al. [10] | Exploit Generation | GPT-3/4 | Exploit template synthesis | Valid exploit fragments generated | Fails on complex multi-stage attacks | Exploit prototyping |
| Grov et al. [115] | Dual-Use Assessment | LLM code assistants | Offensive-defensive overlap analysis | Shows attacker capability amplification | No empirical exploits | Capability risk analysis |
| Iturbe et al. [8] | Polymorphic Malware | Hybrid LLM-code generator | Variant diversification | Signature-based detection evasion | No runtime behaviour testing | Evasion-focused malware |

(Continued)

**Table 7 (continued)**

| Study | Threat Subcategory | Model/Setting | Methodological Distinction | Demonstrated Capability/Advantage | Observed Limitations/Failure Modes | Applicable Threat Scenarios |
|---|---|---|---|---|---|---|
| Huynh et al. [19] | Behavioural Malware | Transformer models | Trace-level semantic modelling | Improved detection accuracy | High computational overhead | Enterprise malware detection |
| Che et al. [106] | Domain Adaptation | Adapted LLMs | Cross-family generalisation | Mitigates dataset shift | High retraining cost | Multi-family malware |
| **C. Adversarial Prompting, Jailbreaks, and Prompt Injection** | | | | | | |
| Yao et al. [15] | Jailbreak Attacks | General LLMs | Role-play exploitation | Widespread vulnerability identified | Qualitative scope only | LLM safety testing |
| Maity and Arora [116] | Automated Jailbreaking | GPT variants | Prompt search automation | High success rates | Limited model diversity | Red-teaming pipelines |
| Villa et al. [32] | Large-Scale Red-Teaming | Multiple LLMs | Black-box robustness testing | Model size correlates with risk | Text-only analysis | Model risk evaluation |
| De et al. [27] | Prompt Injection | RAG systems | System prompt override | Chain-level vulnerabilities exposed | No large-scale benchmarks | Agentic pipelines |
| Huang et al. [119] | Multimodal Jailbreaks | Vision-LLMs | Cross-modal exploitation | Vision-text alignment bypass | Defences underdeveloped | Multimodal assistants |
| **D. Misinformation, Disinformation, and Influence Operations** | | | | | | |
| Schmitt and Flechais [13] | Narrative Manipulation | LLM narratives | Strategic framing analysis | High persuasion realism | No longitudinal impact study | Influence campaigns |
| Harris et al. [123] | Fake News Generation | ChatGPT-class models | Human indistinguishability tests | High deception credibility | Limited dataset diversity | News manipulation |
| Valdez et al. [14] | Cross-Cultural Disinfo | Multilingual LLMs | Cultural adaptation analysis | Effective cross-lingual transfer | Few languages evaluated | Global misinformation |
| Doumanas et al. [149] | Watermark Fragility | Survey | Regeneration-based analysis | Watermarks easily removed | No system-level evaluation | Attribution evasion |

## 6 LLM-Based Defensive Capabilities

Anchored in the second dimension of the unified threat–defence–evaluation taxonomy, this section synthesises how large language models (LLMs) are operationalised as *defensive instruments* across contemporary cybersecurity pipelines. In contrast to AIGC-driven threats (Section 5), defensive studies overwhelmingly conceptualise LLMs not as autonomous detectors, but as *cognitive and semantic augmentation layers* embedded within existing security architectures. An analysis of the 167 reviewed studies published between 2022 and 2025 reveals four dominant defensive domains: (i) intrusion, malware, and anomaly detection; (ii) threat intelligence extraction and Security Operations Center (SOC) automation; (iii) vulnerability detection, secure coding, and automated repair; and (iv) hybrid and multi-agent architectures for investigation and response.

Across these domains, LLMs primarily contribute through enhanced **semantic reasoning**, cross-context inference, and explainable decision support rather than improvements in raw detection accuracy [151,152]. At the same time, the literature consistently documents structural limitations—including hallucination-induced errors, evaluation fragility, adversarial susceptibility, computational overhead, and governance gaps—that constrain safe deployment in high-assurance operational environments. These tensions motivate a taxonomy-aware and failure-conscious analysis, rather than a purely capability-centric narrative.

### *6.1 General Cybersecurity Applications vs. IDS-Specific Architectures*

A recurring source of conceptual ambiguity in the existing literature is the frequent conflation of *general LLM-enabled cybersecurity applications* with *IDS-specific LLM architectures*. This distinction is not merely terminological; it is foundational for correctly interpreting reported performance gains, assessing deployment readiness, and understanding risk exposure. The absence of a clear separation between these paradigms has contributed to overgeneralised claims regarding the suitability of LLMs for real-time intrusion detection and operational cyber defence.

General LLM-based cybersecurity applications primarily function as *analyst-facing cognitive assistants*. Representative tasks include threat report summarisation, indicator-of-compromise (IOC) extraction, incident timeline reconstruction, policy and configuration interpretation, and SOC co-pilot support [33,35,36]. Such systems typically operate on curated or semi-structured inputs, tolerate moderate inference latency, and are evaluated using task-centric or qualitative metrics such as extraction accuracy, reasoning coherence, or analyst productivity. Their principal contribution lies in reducing cognitive burden, improving contextual understanding, and accelerating sense-making, rather than in executing primary detection or enforcement decisions.

In contrast, IDS-specific LLM architectures are embedded directly within intrusion detection pipelines and are subject to substantially stricter operational constraints. As synthesised in Sections 6.4 and 6.5, LLMs in IDS contexts do not replace conventional signature-based, statistical, or machine-learning detectors. Instead, they are positioned as auxiliary reasoning layers that augment detection outputs through semantic interpretation of alerts, behavioural abstraction over logs and traces, and contextual explanation of anomalous activity [18,39,40]. These systems must operate under high-throughput data streams, adversarial noise, concept drift, and strict latency budgets, rendering naive adoption of general-purpose LLM deployments impractical without careful architectural mediation.

This distinction has direct implications for both evaluation and deployment. Evaluation protocols commonly applied to general LLM applications-such as standalone reasoning benchmarks or static text-based assessments-are insufficient for IDS-specific scenarios, where errors may propagate downstream and amplify operational risk. In particular, false-positive escalation, inference latency, and reasoning

instability can compound across SOC workflows, increasing analyst workload and degrading response effectiveness [38,121,140]. Consequently, performance claims derived from general LLM evaluations cannot be extrapolated to IDS deployments without explicit consideration of these system-level effects.

Explicitly distinguishing between general cybersecurity applications and IDS-specific architectures therefore prevents overstatement of LLM readiness for real-time intrusion detection and clarifies the necessity of hybrid, defence-in-depth designs. In such architectures, classical detectors provide time-critical guarantees, while LLMs contribute semantic reasoning, correlation, and explainability under human-in-the-loop supervision. Table 8 formalises this distinction and provides a conceptual reference framework for interpreting the task-level and quantitative analyses presented in subsequent sections.

**Table 8:** Critical comparison between general LLM cybersecurity applications and IDS-specific LLM architectures.

| Dimension | General LLM Cybersecurity Applications | IDS-Specific LLM Architectures |
|---|---|---|
| **Primary Objective** | Analyst support, knowledge extraction, reasoning, and workflow automation | Detection augmentation, alert interpretation, and behavioural reasoning within IDS pipelines |
| **Typical Tasks** | Threat intelligence summarisation, IOC extraction, SOC co-pilots, secure coding, incident reporting | Post-detection reasoning for NIDS/HIDS/IIoT IDS, anomaly explanation, alert correlation |
| **Architectural Role of LLM** | Standalone or loosely coupled cognitive assistant | Embedded component layered on top of conventional IDS detectors |
| **Detection Responsibility** | No direct responsibility for intrusion detection decisions | Does not replace detectors; augments detection outputs with semantic reasoning |
| **Input Modality** | Natural language reports, alerts, threat feeds, code, policies | IDS alerts, logs, behavioural traces, network summaries (often structured + text) |
| **Latency Constraints** | Moderate to low; interactive or batch processing acceptable | Strict; real-time detection delegated to classical models, LLMs operate asynchronously |
| **Evaluation Metrics** | Task accuracy, summarisation quality, extraction precision, analyst productivity | Detection accuracy/F1 (indirect), alert reduction, explanation quality, false-positive mitigation |
| **Datasets Used** | Curated text corpora, CTI reports, code repositories | IDS benchmarks (e.g., CICIDS, UNSW-NB15), logs, malware traces, IIoT telemetry |
| **Baseline Comparisons** | Often compared to rule-based tools or traditional NLP methods | Compared against CNN/RNN/Transformer IDS pipelines |

(Continued)

**Table 8 (continued)**

| Dimension | General LLM Cybersecurity Applications | IDS-Specific LLM Architectures |
|---|---|---|
| **Failure Modes** | Hallucinated facts, grounding errors, outdated knowledge | False-positive escalation, reasoning errors over noisy alerts, latency overhead |
| **Operational Readiness** | High for SOC decision support and automation | Emerging; suitable for hybrid IDS, not for standalone real-time detection |
| **Governance & Safety Risk** | Misinformation, policy misinterpretation, analyst over-trust | Cascading errors in IDS pipelines, unsafe automated response decisions |
| **Key Limitation** | Lack of direct linkage to detection pipelines | Lack of formal guarantees and limited real-world SOC-scale validation |
| **Research Gap** | Trust calibration and grounding in dynamic cyber contexts | Standardised IDS benchmarks with latency, drift, and deployment constraints |

### *6.2 Impact of Model Scale and Architecture*

Although LLM-enabled threats and defences have been examined extensively, comparatively limited attention has been devoted to understanding how *model scale, architectural design, and deployment modality* influence robustness, misuse risk, and operational feasibility [153]. This subsection addresses this gap by synthesising empirical evidence across defensive studies and situating architectural choices within practical deployment constraints.

#### *6.2.1 Model Scale: Small vs. Large LLMs*

Small- and medium-scale LLMs-including distilled, quantised, and domain-adapted variants-are increasingly favoured in SOC, edge, and IIoT settings due to their lower inference latency, reduced computational cost, and improved controllability. Multiple studies demonstrate that such models can achieve competitive performance on constrained security tasks, including log interpretation and malware trace reasoning [19,105]. In contrast, large foundation models typically offer superior generalisation and deeper reasoning capabilities but are associated with elevated hallucination risk, greater potential for misuse amplification, and prohibitive operational costs in continuous-deployment environments [33,38]. Collectively, these findings suggest that model scale mediates a fundamental trade-off between defensive capability and deployment risk.

#### *6.2.2 Deployment Modality: API-Based vs. Local Models*

Deployment modality further conditions defensive effectiveness and risk exposure. API-based LLM deployments benefit from provider-managed updates, centralised alignment controls, and rapid access to state-of-the-art models, but introduce data sovereignty concerns, external service dependencies, and limited

transparency into model behaviour. By contrast, locally deployed and fine-tuned models provide tighter control over data flows, inference logic, and privacy-sensitive logs, making them particularly suitable for SOC operations and critical infrastructure environments [35,36]. However, local deployment shifts responsibility for model maintenance, security hardening, and misuse mitigation to system operators, thereby introducing additional governance and lifecycle-management challenges.

#### *6.2.3 Single-Model vs. Agentic Architectures*

Agentic and multi-LLM architectures are increasingly adopted to support complex investigation, correlation, and response workflows [3,24]. While such designs enhance modularity, task specialisation, and workflow automation, they also expand the effective attack surface and introduce new failure modes, including coordination errors, cascading hallucinations, and prompt-injection propagation across agents [38,154]. Single-model architectures, by contrast, are generally easier to audit, reason about, and deploy, but often struggle with multi-stage reasoning, cross-domain correlation, and concurrent task execution.

#### *6.2.4 Architectural Synthesis*

Taken together, architectural and scale-related design choices exert a first-order influence on defensive robustness and deployability. Larger and agentic systems prioritise capability breadth and scalability, whereas smaller, locally controlled models emphasise predictability, auditability, and operational safety. Table 9 summarises these trade-offs across representative deployment scenarios.

**Table 9:** Comparative analysis of LLM architectures and model scale in cybersecurity applications.

| Dimension | Small/Distilled LLMs | Large Foundation LLMs | Agentic Multi-LLM Pipelines |
|---|---|---|---|
| Model Capacity | Limited reasoning depth; task-specific | High generalisation and reasoning power | Distributed reasoning across specialised agents |
| Robustness | More predictable; easier to constrain | Stronger reasoning but higher hallucination risk | Sensitive to inter-agent coordination failures |
| Misuse Risk | Lower misuse amplification | Higher potential for abuse and weaponisation | Expanded attack surface across agents |
| Deployment Cost | Low compute and energy footprint | High inference and operational cost | High integration and orchestration overhead |
| Deployability | Suitable for edge, SOC, and ICS contexts | Primarily cloud-based deployments | Best suited for complex SOC automation |
| Governance & Control | High controllability and auditability | Limited transparency; provider dependence | Complex policy enforcement across agents |
| Key Trade-Off | Safety and efficiency over generality | Capability over control | Scalability over simplicity |

### *6.3 Intrusion and Anomaly Detection*

A substantial subset of the reviewed literature investigates the application of LLMs to intrusion and anomaly detection across heterogeneous telemetry sources, including system logs, network flows, endpoint activity, and malware execution traces. Survey studies such as [18,33] report that language-model-based semantic representations can capture long-range dependencies and behavioural context more effectively than traditional feature-engineered approaches. Practical systems further illustrate complementary strengths: lightweight and domain-adapted models enable deployment in resource-constrained edge environments [105], while semantic trace modelling improves the detection of low-frequency and stealthy attack patterns [19].

Domain adaptation emerges as a critical enabler of robust performance. Studies that integrate LLM-derived embeddings with classical detection mechanisms demonstrate improved resilience to dataset drift and enhanced generalisation across operational environments [103,106]. Nevertheless, persistent limitations remain. Inference latency constrains real-time deployment, prompt sensitivity undermines output stability, and overfitting to benchmark artefacts is frequently observed. As a result, the literature converges on positioning LLMs as *post-detection reasoning layers* that augment conventional intrusion detection systems rather than as standalone replacements.

### *6.4 Task-Level Comparison: NIDS, HIDS, and IIoT/ICS*

Traditional IDS surveys typically categorise methods by deployment context-network-based intrusion detection systems (NIDS), host-based intrusion detection systems (HIDS), and IIoT/ICS environments-and evaluate transformer models primarily as feature-level classifiers optimised for benchmark accuracy [18,39,40]. LLM-assisted IDS approaches depart fundamentally from this paradigm. Rather than competing directly at the detection stage, LLMs are employed as higher-level cognitive components that support interpretation, correlation, and response.

In NIDS settings, LLMs operate downstream of packet- or flow-based detectors, enabling semantic correlation and explanation across heterogeneous alerts [64,103]. In HIDS deployments, LLMs abstract diverse host-level logs into unified semantic representations, facilitating cross-host reasoning and improved interpretability [19,136]. In IIoT and ICS contexts, LLMs are generally unsuitable for control-loop intrusion detection due to latency and safety constraints, but they provide significant value for post-hoc analysis, attack-path reasoning, and policy-aware decision support [3,47]. These task-level distinctions are synthesised in Table 10.

### *6.5 Quantitative Synthesis of LLM-Based Intrusion Detection Studies*

Recent LLM-assisted intrusion detection systems (IDS) increasingly report performance improvements over classical machine learning and transformer-based baselines. However, the quantitative evidence remains fragmented due to heterogeneous datasets, inconsistent evaluation protocols, and divergent deployment assumptions. This subsection provides a critical synthesis of *reported* quantitative results from representative LLM-based IDS studies, focusing on detection performance, relative baseline improvements, latency and computational overhead, and dataset characteristics. No new experiments are introduced; rather, the analysis consolidates metrics as presented in the primary literature to enable principled cross-task comparison while explicitly contextualising reported gains [155].

**Table 10:** Task-level comparison of transformer-based IDS and LLM-assisted intrusion detection.

| IDS Task | Transformer-Based IDS Role | LLM-Assisted Role | Key Advantages of LLM Integration | Limitations and Applicability |
|---|---|---|---|---|
| **NIDS** | Sequence modelling of packets, flows, and traffic features for supervised or semi-supervised classification [18,39] | Post-detection semantic reasoning over alerts, logs, and correlated network events [64,103] | Improved interpretability of alerts; contextual correlation across heterogeneous network telemetry; analyst decision support [35] | Not suitable for real-time packet inspection; dependent on upstream detectors and log quality; increased inference overhead [38] |
| **HIDS** | Behavioural modelling of system calls, audit logs, and host activities using feature-level representations [40] | Semantic abstraction of host logs and execution traces; cross-host reasoning and explanation of anomalous behaviour [19,103] | Enhanced detection of complex insider threats and privilege escalation; improved explainability across diverse log sources [35] | Sensitivity to log formatting and noise; requires robust preprocessing and human validation [38] |
| **IIoT/ICS IDS** | Protocol-specific feature extraction and real-time anomaly detection under strict latency constraints [39,40] | Context-aware post-hoc analysis, attack-path reasoning, and policy-aware interpretation of incidents [3,47] | Supports safety-aware reasoning and integration of cyber events with operational context in CPS environments [35] | Unsuitable for control-loop intrusion detection; limited validation in safety-critical real-time deployments [38] |
| **Overall Positioning** | Feature-level classifiers optimised for detection accuracy on benchmark datasets [18,40] | Cognitive and agentic security components augmenting IDS pipelines across detection, analysis, and response [3,24] | Transforms IDS from isolated detection to explainable, decision-oriented defence architectures [33,35] | Requires hybrid architectures, governance controls, and human-in-the-loop oversight for safe deployment [38] |

### *6.5.1 Detection Performance and Baseline Improvements*

Across network-, host-, and IIoT-oriented IDS tasks, LLM-assisted approaches consistently report strong benchmark performance, with accuracy or F1-scores typically ranging between 92% and 99%. As summarised in Table 11, the most pronounced gains arise when LLMs operate on semantically rich inputs such as system logs, behavioural traces, and alert narratives, where contextual reasoning complements

feature-level detection [19,103,105]. Relative to CNN- and RNN-based baselines, reported improvements commonly fall within the 2%–7% range, whereas gains over transformer-based IDS are more modest and strongly context dependent. In particular, LLMs tend to outperform transformer-based IDS primarily in scenarios requiring cross-source correlation, explanation, or narrative reconstruction rather than raw packet- or flow-level classification [40,64].

**Table 11:** Quantitative summary of reported performance in LLM-based intrusion detection studies with baseline comparison.

| IDS Task | Reported Accuracy/F1 (Range) | Baseline Comparison (LLM vs CNN/RNN/ Transformer) | Datasets Commonly Used | Latency/ Overhead Reporting | Representative Studies |
|---|---|---|---|---|---|
| **NIDS** | 92%–99% (Accuracy/F1, benchmark-dependent) | +2%–6% over CNN/RNN; +1%–4% over transformer NIDS on semantic-rich traffic and logs | CICIDS2017, UNSW-NB15, custom network telemetry | **Limited:** LLMs typically applied post-detection; real-time latency rarely reported | [18,64,103] |
| **HIDS** | 93%–98% (F1 dominant metric) | +3%–7% F1 over RNN/CNN baselines; marginal gains over transformer HIDS when context is limited | System-call traces, host audit logs, malware behaviour datasets | **Partial:** Inference cost discussed qualitatively; no SOC-scale latency benchmarks | [19,40,103] |
| **IIoT/ICS IDS** | 90%–97% (Accuracy under constrained settings) | +2%–5% over classical DL; comparable to transformers for raw detection, stronger for post-hoc reasoning | Industrial telemetry, CPS logs, IIoT-specific datasets | **Sparse:** Latency largely unreported; unsuitable for real-time control loops | [39,47,105] |

(Continued)

**Table 11 (continued)**

| IDS Task | Reported Accuracy/F1 (Range) | Baseline Comparison (LLM vs CNN/RNN/ Transformer) | Datasets Commonly Used | Latency/ Overhead Reporting | Representative Studies |
|---|---|---|---|---|---|
| **Cross-Task Observation** | High accuracy on curated benchmarks; strongest gains for semantic reasoning tasks | LLMs consistently outperform CNN/RNN baselines; gains over transformers are context-dependent | Benchmark-centric with limited domain diversity | Latency and computational cost remain underreported across studies | [35,38] |

Nevertheless, as evidenced by the study-level synthesis in Section 6.10, these improvements are predominantly demonstrated on curated benchmark datasets such as CICIDS2017, UNSW-NB15, and custom malware corpora. This concentration limits cross-study comparability and risks overstating performance when models are deployed outside controlled laboratory settings or evaluated under realistic enterprise conditions.

### *6.5.2 Latency and Computational Overhead*

In contrast to detection accuracy, latency and computational overhead are reported inconsistently across studies. Distilled or lightweight LLM variants demonstrate acceptable inference latency for post-detection reasoning and alert interpretation tasks in edge or resource-constrained environments [105]. By contrast, full-scale LLMs incur substantial computational overhead when applied to streaming intrusion detection, rendering them unsuitable for real-time packet inspection or control-loop defence [18,38]. As reflected in the task-level synthesis, most studies implicitly adopt hybrid architectures in which classical IDS components handle time-critical detection, while LLMs operate asynchronously to support correlation, explanation, and analyst decision-making. Quantitative latency benchmarks under sustained SOC-scale workloads remain largely absent from the literature.

### *6.5.3 Dataset Concentration and Evaluation Bias*

The quantitative synthesis further reveals a pronounced imbalance in dataset usage across IDS tasks. Network intrusion detection studies overwhelmingly rely on a small set of benchmark datasets, whereas host-based and IIoT/ICS datasets remain comparatively scarce, less standardised, and highly domain specific [39,40]. Few studies explicitly evaluate robustness under dataset shift, concept drift, or long-term deployment conditions, despite the centrality of these factors in operational SOC environments. As highlighted in Section 7.5, this benchmark-centric evaluation paradigm obscures generalisation limits and constrains meaningful quantitative comparison.

#### *6.5.4 Quantitative Gaps and Implications*

Taken together, the quantitative evidence suggests that LLM-assisted IDS can achieve strong benchmark performance and consistent improvements over classical baselines, particularly in semantically complex intrusion scenarios. However, these gains are accompanied by increased computational cost, limited real-time applicability, and insufficient validation under realistic operational constraints. The absence of standardised latency metrics, limited dataset diversity, and weak reproducibility practices remain significant barriers to robust quantitative assessment [114].

By consolidating reported accuracy ranges, baseline improvements, dataset usage, and latency reporting practices in Table 11, this subsection complements the task-level analysis presented in Section 6.4. Together, these findings underscore the need for next-generation IDS benchmarks that jointly report detection performance, computational cost, and deployment context, particularly for SOC-scale and IIoT intrusion detection.

### *6.6 Threat Intelligence Extraction and SOC Automation*

Threat intelligence workflows require the synthesis of information from highly unstructured and heterogeneous sources, including vulnerability advisories, forensic logs, advanced persistent threat (APT) reports, darknet content, and open threat feeds. A growing body of evidence demonstrates that LLMs are well suited to automating these processes. Loumachi et al. [20] show that LLM-based pipelines substantially improve indicator-of-compromise (IOC) extraction, entity linking, and threat report summarisation. Broader reviews similarly confirm the effectiveness of generative AI for cyber threat intelligence (CTI) processing and intelligence generation [48,50,132].

Retrieval-augmented generation (RAG) further enhances factual grounding and mitigates hallucination risk. Xu et al. [128] demonstrate that RAG-based cross-document reasoning produces more coherent threat timelines and reduces ambiguity in alert interpretation. Systems such as RAG-CDI operationalise this paradigm for actionable CTI generation [55,156]. Within SOC environments, LLM-based co-pilots increasingly support alert triage and response coordination, reducing mean time to response through automated synthesis and correlation of alerts [23,56]. Nonetheless, persistent risks related to hallucination, inconsistent reasoning, and susceptibility to prompt manipulation necessitate structured prompting, robust grounding mechanisms, and multi-stage verification pipelines [36,157].

### *6.7 Secure Coding, Vulnerability Analysis, and Automated Repair*

LLMs are widely applied to vulnerability detection, secure code synthesis, patch recommendation, and automated code review [33,34]. Benchmarking studies report improved vulnerability localisation and recall across multiple Common Weakness Enumeration (CWE) families when compared with classical static and learning-based approaches [112,158,159]. Fine-tuned repair-oriented models further enhance patch correctness and semantic consistency, particularly for recurring vulnerability patterns [22,133].

However, empirical audits consistently reveal that LLM-generated code may introduce new vulnerabilities, omit edge cases, or yield incomplete remediation [57,117]. As a result, the literature increasingly advocates hybrid pipelines that integrate LLM-based semantic reasoning with symbolic execution, static analysis, and developer-in-the-loop validation to mitigate hallucination-induced errors and ensure robustness [62,120,134].

### *6.8 Hybrid and Multi-Agent Defence Architectures*

Hybrid and multi-agent defence architectures decompose complex security workflows into specialised agent roles coordinated by LLM-based planners [35,38]. Empirical studies report reductions in analyst workload and response latency through collaborative reasoning, task delegation, and cross-agent information sharing [3,24,65]. Extensions of these architectures to cyber-physical systems and smart-grid environments further demonstrate their applicability beyond conventional IT infrastructures [47].

Despite these advantages, systemic risks persist, including hallucination propagation, coordination failures, and unintended tool misuse across chained agents [129,138]. Current empirical evidence therefore supports the deployment of multi-agent architectures primarily for decision support and semi-automated defence, rather than for fully autonomous response in high-assurance environments.

### *6.9 Operational Failure Modes and Cost-Aware Limitations*

Across defensive domains, operational deployment exposes failure modes that are insufficiently captured by benchmark-centric evaluations. Semantic overgeneralisation can amplify false positives, exacerbating alert fatigue and misdirecting analyst attention [3,38]. Computational overhead, inference latency, and energy consumption further constrain scalability, particularly in SOC-scale and IIoT deployments, yet remain inconsistently reported in the literature [18,105]. In addition, defensive LLM systems remain vulnerable to adversarial manipulation, including prompt injection and context poisoning, with cascading effects observed in RAG-based and multi-agent pipelines [24,27,111]. These observations reinforce the conclusion that defensive effectiveness is inseparable from architectural safeguards, governance controls, and sustained human oversight.

**Capability-oriented synthesis.** Table 12 provides a high-level, taxonomy-aligned synthesis of *what LLM-based defensive systems are designed to do*, focusing on their core technical roles and demonstrated capability gains across defensive subcategories.

### *6.10 Defence Synthesis*

Synthesising evidence across intrusion detection, threat intelligence processing, secure coding, and agentic defence architectures yields a consistent conclusion: LLMs deliver their greatest defensive value when deployed as *bounded cognitive amplifiers* rather than as autonomous decision-makers. Quantitative performance gains are strongest for tasks involving heterogeneous data integration, cross-alert reasoning, and explanation, whereas advantages over transformer-based IDS diminish for low-level, time-critical detection tasks.

Across all defensive subcategories (Tables 12 and 13), hybrid architectures emerge as the dominant and most operationally viable paradigm. Classical detectors provide latency guarantees and robustness under adversarial noise, while LLMs augment these pipelines with semantic reasoning and analyst-facing interpretability. At the same time, model scale, agentic complexity, and deployment modality introduce non-trivial trade-offs between capability, cost, and safety. Addressing these trade-offs requires evaluation frameworks that jointly consider accuracy, latency, robustness, and governance-an issue revisited in Section 7.

**Table 12:** Capability-oriented synthesis of LLM-based defensive subcategories across the proposed taxonomy.

| Defensive Subcategory | Key References | Core Technical Role of LLMs | Observed Benefits | Key Constraints |
|---|---|---|---|---|
| **Intrusion & Anomaly Detection** | [18,19,38,39,103,105] | Semantic interpretation of logs, flows, and traces; prompt-guided detection; hybrid LLM–ML IDS pipelines | Improved detection of low-and-slow attacks; better interpretability than feature-only IDS; robustness gains via domain adaptation | High inference latency; limited real-time validation; sensitivity to prompt format; weak evidence of sustained SOC deployment |
| **Threat Intelligence & SOC Automation** | [36,50,55,56,128] | RAG-based IOC extraction; alert summarisation; cross-source correlation; analyst co-pilot functions | Reduced analyst workload; faster triage and incident understanding; scalable processing of unstructured CTI | Hallucinated or mis-grounded IOCs; dependence on retrieval quality; limited validation in live SOC workflows |
| **Secure Coding & Vulnerability Detection** | [22,34,38,59] | Semantic code reasoning; vulnerability localisation; guided patch generation; SDLC-integrated security assistance | Improved logic-flaw detection; better patch explanations; multi-language support under constrained prompting | Risk of incomplete or insecure fixes; uneven performance across CWE classes; requires developer-in-the-loop review |
| **Hybrid & Multi-Agent Defence Systems** | [3,24,47,65,130] | Multi-agent coordination for investigation and response; policy-aware reasoning; tool-augmented workflows | End-to-end automation; cross-stage reasoning; reduced response time in complex incidents | Cascading hallucinations; coordination instability; expanded attack surface; lack of formal safety guarantees |

**Table 13:** Operational-risk and failure-mode analysis of representative LLM-based defensive studies.

| Study | Defensive Subcategory | Model/System | Methodological Distinction | Demonstrated Performance/Advantage | Observed Limitations and Applicability |
|---|---|---|---|---|---|
| Elouardi et al. [18] | Intrusion Detection | Survey of LLM-IDS | Semantic modelling of cyber telemetry for multi-stage attack detection | Identifies consistent gains over classical IDS for complex, multi-step attacks | Survey-level evidence only; limited operational validation |
| Rondanini et al. [160] | Edge Anomaly Detection | Lightweight LLMs | Model compression for edge deployment | Maintains strong accuracy under latency and resource constraints | Generalisation limited across heterogeneous network environments |
| Rondanini et al. [105] | Malware Detection | Lightweight transformer IDS | Hybrid static-dynamic malware classification | Robust to obfuscation with low latency on mobile devices | Dataset imbalance impacts stability |
| Huynh et al. [19] | Behavioural Malware Detection | Transformer trace models | Semantic modelling of execution traces | Improved detection of stealthy malware families | Requires large labelled trace corpora |
| Che et al. [106] | Domain Adaptation | Adapted LLMs | Cross-family fine-tuning to mitigate dataset drift | Significant improvement in generalisation across malware families | High retraining cost; continuous adaptation required |
| Hmimou et al. [103] | Hybrid IDS | LLM-ML fusion | Semantic log embeddings combined with classical ML | Strong robustness across heterogeneous log sources | Reduced interpretability; tuning complexity |
| Zou et al. [161] | Traffic Analysis | Transformer NIDS | Packet-level semantic embedding | Outperforms CNN/RNN baselines on long-range dependencies | Latency challenges at high throughput |
| Lai et al. [162] | IDS Benchmarking | LLM embeddings + ML | Feature enrichment for downstream IDS | Consistent F1-score improvement across datasets | Performance tied to downstream classifier choice |

(Continued)

**Table 13 (continued)**

| Study | Defensive Subcategory | Model/System | Methodological Distinction | Demonstrated Performance/Advantage | Observed Limitations and Applicability |
|---|---|---|---|---|---|
| Usman et al. [163] | Energy-Efficient IDS | Green LLM-based IDS | Power-aware detection strategies | Feasible IDS under strict energy budgets | Accuracy degrades in noisy traffic |
| Koide et al. [63] | Phishing Detection | ChatGPT reasoning | Semantic analysis of phishing content | Effective detection beyond lexical features | English-only focus; limited multimodal scope |
| Li et al. [64] | NIDS | Prompt-learning LLM | Semantic extraction from raw traffic | Competitive classification accuracy with minimal feature engineering | Requires environment-specific prompt adaptation |
| Loumachi et al. [20] | CTI Extraction | LLM-NER pipelines | Context-aware IOC extraction | Improves extraction precision and analyst efficiency | Entity hallucination in sparse contexts |
| Xu et al. [128] | CTI Correlation | RAG-enhanced LLM | Cross-document threat correlation | Improved narrative reconstruction | Highly dependent on retrieval quality |
| Srinivas et al. [23] | SOC Automation | LLM co-pilot | Automated triage and reporting | Reduced MTTR and analyst workload | Not validated under sustained SOC load |
| Karkuzhali and Senthilkumar [158] | Vulnerability Detection | LLM classifier | CWE-aligned semantic classification | High recall across vulnerability classes | False positives remain high |
| De Fitero-Dominguez et al. [22] | Patch Generation | Fine-tuned repair LLM | Language-specific secure patching | Outperforms prior DL-based repair models | Patch reliability varies by vulnerability type |
| Pearce et al. [117] | Code Security Assessment | General-purpose LLMs | Empirical audit of AI-generated code | Reveals frequent silent vulnerabilities | Requires external verification frameworks |
| Andreoni-LN et al. [24] | Multi-Agent IR | LLM agent collective | Collaborative reasoning across IR stages | Improved incident reconstruction accuracy | Error propagation across agents |

(Continued)

**Table 13 (continued)**

| Study | Defensive Subcategory | Model/System | Methodological Distinction | Demonstrated Performance/Advantage | Observed Limitations and Applicability |
|---|---|---|---|---|---|
| Jaffal et al. [3] | SOC Multi-Agent System | LLM agent framework | Task-specialised agent orchestration | Strong applicability to SOC workflows | Not robust under adversarial prompting |
| Zhai et al. [130] | Automated Response | Multi-agent responders | Response plan generation | Effective for simple attack scenarios | Fails under complex lateral movement |

**Operational-risk and failure-mode synthesis.** Complementing the capability-oriented overview, Table 13 provides a study-level synthesis that examines *how LLM-based defences behave in practice*, explicitly highlighting methodological distinctions, performance trade-offs, and limitations.

Taken together, Tables 12 and 13 provide a consolidated response to **RQ2**, demonstrating that LLM-based defensive systems derive their primary value as components for semantic reasoning, contextual correlation, and analyst decision support rather than as standalone detection mechanisms. Across intrusion detection, threat intelligence processing, secure coding, and multi-agent response scenarios, the reviewed studies consistently report gains in interpretability, cross-source reasoning, and workflow automation. However, these benefits are intrinsically coupled to hybrid deployment models that integrate LLMs with conventional detectors, maintain robust human-in-the-loop oversight, and incorporate explicit safeguards against hallucination, false-positive escalation, and cascading reasoning failures.

Overall, this synthesis indicates that LLM-based defences should be understood as a complementary evolution of cybersecurity practice-one that reorients IDS and SOC workflows from isolated, signal-level detection toward explainable, context-aware, and decision-oriented defence systems. At the same time, the evidence underscores that sustainable operational adoption depends on evaluation paradigms and governance mechanisms that jointly account for detection effectiveness, latency and computational overhead, adversarial robustness, and deployment context. In the absence of such integrated assessment and control frameworks, the defensive advantages of LLMs risk being offset by new forms of operational fragility and systemic risk.

## 7 Security Evaluation Frameworks, Benchmarks, and Red-Teaming

As LLMs transition from experimental tools to embedded components of cybersecurity infrastructures—spanning SOC co-pilots, threat intelligence platforms, IDS reasoning layers, code analysis utilities, and multi-agent response systems—the question of *how* these systems are evaluated becomes as critical as *what* they can do. The reviewed literature reflects a rapidly expanding ecosystem of security benchmarks, red-teaming strategies, and evaluation protocols targeting adversarial robustness, harmful-content refusal, alignment stability, and lifecycle risk. Nevertheless, evaluation practices remain uneven across application domains, modalities, and deployment contexts, with limited standardisation and minimal longitudinal assessment of model evolution.

This section critically organises evaluation research along five interrelated dimensions: (i) dataset usage and experimental realism, (ii) security-oriented benchmarks and datasets, (iii) red-teaming methodologies, (iv) critical benchmarking and ranking of evaluation metrics, and (v) an integrated evaluation synthesis. This structure explicitly differentiates laboratory-driven assessment from operationally meaningful security evaluation.

### *7.1 Dataset Usage and Experimental Realism*

Dataset selection and experimental design fundamentally shape the validity, generalisability, and interpretability of reported results in LLM-based cybersecurity research. While benchmark-driven evaluations support controlled comparison and reproducibility, they also introduce structural biases that limit experimental realism-particularly for intrusion detection and SOC-integrated deployments.

*7.1.1 Benchmark-Centric Evaluation Practices*

A dominant trend across LLM-assisted intrusion detection and security analytics studies is the reliance on a small set of canonical benchmarks, including CICIDS2017, UNSW-NB15, curated malware corpora, and synthetic phishing datasets. As synthesised in Sections 6.5.4 and 6.10, these datasets are typically collected under controlled assumptions, with relatively clean labels, balanced class distributions, and isolated attack instances. Although indispensable for methodological comparison, such benchmarks fail to capture the noise, sparsity, temporal overlap, and severe alert imbalance characteristic of operational SOC telemetry [18,39,40].

*7.1.2 Synthetic and Laboratory-Scale Evaluations*

Beyond classical benchmarks, several studies evaluate LLM-based systems using synthetic logs, simulated attack traces, or LLM-generated adversarial inputs. These laboratory-scale evaluations facilitate scalable experimentation and targeted stress-testing of reasoning and alignment properties. However, they abstract away critical operational factors, including incomplete logging, policy-driven filtering, analyst feedback loops, and adaptive adversarial behaviour. Consequently, strong performance under synthetic conditions should be interpreted as an upper bound on potential capability rather than as evidence of deployment readiness [33,38].

*7.1.3 Evidence from Real-World and Longitudinal Evaluations*

Only a limited subset of studies evaluates LLM-based systems using real SOC logs, enterprise network traces, or long-duration operational datasets. Even fewer investigate longitudinal effects such as concept drift, evolving attacker strategies, or feedback-induced behavioural shifts. This gap is particularly consequential for LLM-assisted IDS, where reasoning quality and alert correlation depend critically on contextual completeness and temporal continuity. As reflected in Section 7.5, the scarcity of SOC-grade and longitudinal evaluations constrains the generalisability of reported gains and weakens claims of operational robustness [35,38].

*7.1.4 Implications for Experimental Realism*

The prevailing reliance on benchmark-centric and laboratory-scale evaluations has direct implications for how reported results should be interpreted. Improvements in detection accuracy, reasoning coherence, or refusal robustness primarily reflect potential capability under idealised conditions rather than guarantees of real-world effectiveness. For IDS-specific and SOC-integrated LLM deployments, realistic evaluation must jointly consider detection performance, latency, alert volume, analyst interaction, and adversarial adaptation-dimensions that remain underrepresented in current evaluation pipelines.

Table 14 summarises these contrasts and highlights the persistent gap between laboratory experimentation and operational deployment.

**Table 14:** Critical comparison of dataset types and experimental realism in LLM-based cybersecurity studies.

| Dimension | Benchmark/Synthetic Evaluations | Real-World/Deployment-Oriented Evaluations |
|---|---|---|
| **Typical Data Sources** | CICIDS2017, UNSW-NB15, curated malware corpora, synthetic logs | SOC telemetry, enterprise network traces, operational host logs |

(Continued)

**Table 14 (continued)**

| Dimension | Benchmark/Synthetic Evaluations | Real-World/Deployment-Oriented Evaluations |
|---|---|---|
| **Data Characteristics** | Balanced classes, clean labels, isolated attacks | Severe class imbalance, noisy logs, overlapping benign activity |
| **Temporal Coverage** | Static or short-duration snapshots | Continuous, longitudinal data with evolving behaviour |
| **Adversarial Adaptation** | Largely absent or scripted | Implicit and adaptive attacker behaviour |
| **Evaluation Focus** | Detection accuracy, F1-score, reasoning correctness | Operational reliability, alert volume, analyst workload, robustness |
| **Latency Considerations** | Often ignored or qualitatively discussed | Critical for SOC and IDS pipelines |
| **Reproducibility** | High due to public availability | Low due to privacy and confidentiality constraints |
| **Primary Limitation** | Overestimation of performance and generalisation | Limited availability and standardisation |
| **Implication for LLM-Based IDS** | Upper-bound performance under idealised conditions | Realistic assessment of deployment readiness |

### *7.2 Security-Oriented Benchmarks and Datasets*

Security-oriented benchmarks constitute a foundational element of LLM evaluation by providing task-specific corpora designed to surface vulnerabilities under realistic cybersecurity conditions. Recent research has introduced a diverse set of benchmarks targeting phishing generation, harmful-query refusal, exploit reasoning, code-injection robustness, multilingual misinformation, and alignment stability.

Stylometry-aware phishing benchmarks [6] evaluate both AIGC-driven phishing capability and the resilience of downstream classifiers, while comprehensive security evaluation suites [25] span refusal behaviour, exploit-assisted reasoning, code-injection robustness, and semantic safety. High-variance synthetic phishing datasets [100] probe generalisation limits under adversarial diversity, and cross-lingual misinformation benchmarks [164] expose language-dependent robustness gaps. Additional domain-specific datasets target vulnerability reasoning [137,165], scenario-driven manipulation [138], constrained malware detection [105], and cross-model robustness comparison.

Despite their analytical value, most existing benchmarks remain static, predominantly text-centric, and non-adaptive. As a result, multimodal deception, adaptive adversarial strategies, social-context manipulation, and lifecycle-dependent effects are systematically underrepresented. Moreover, frequent vendor-driven model updates invalidate earlier benchmark results, complicating longitudinal comparison and undermining claims of sustained robustness across model generations.

### *7.3 Red-Teaming Methodologies*

Red-teaming represents the primary mechanism for uncovering misalignment pathways, unsafe behaviours, and systemic vulnerabilities in LLM-based systems. Across the reviewed literature, red-teaming

methodologies range from manual expert probing and automated jailbreak generation to conversational manipulation, scenario-based stress testing, and lifecycle-oriented evaluation.

Taxonomies of red-teaming strategies [15] classify attacks based on role-play, multi-turn interaction, and obfuscation techniques, while large-scale automated studies [32,111,116] demonstrate the high transferability of jailbreak strategies across model families. Conversational red-teaming approaches [26] reveal failure modes that remain invisible under single-turn testing, and prompt-injection studies [166] expose persistent vulnerabilities in retrieval-augmented and tool-integrated systems. Lifecycle-oriented evaluations [27] further illustrate how vulnerabilities propagate across retrieval, agentic reasoning, and orchestration layers.

Despite this methodological diversity, red-teaming practices remain fragmented. Unified protocols, standardised success criteria, and reproducible baselines are largely absent, and systematic tracking of vulnerability persistence across model updates is rare. These limitations significantly constrain comparative analysis and impede longitudinal assessment of security improvements.

### *7.4 Critical Benchmarking and Ranking of Evaluation Metrics*

Evaluation metrics play a decisive role in shaping conclusions about LLM security, yet they are frequently applied in isolation. Across the reviewed literature, three broad classes of metrics can be identified.

*Capability-centric metrics* (e.g., accuracy, F1-score, exploit success rate) dominate IDS- and malware-oriented evaluations due to their quantitative clarity and ease of comparison [18,19]. While necessary, these metrics primarily capture task proficiency and often fail to reflect semantic misreasoning, cascading errors, or downstream operational impact.

*Behaviour-centric metrics*, including refusal consistency, hallucination rate, calibration error, and explanation fidelity, offer deeper insight into reasoning robustness and alignment stability [16,28]. Such metrics are particularly relevant for SOC automation and agentic systems, yet remain underutilised in IDS-centric evaluations.

*Operational risk-centric metrics*, such as false-positive escalation rate, analyst workload amplification, response latency, and inference overhead, are the most deployment-relevant but also the least standardised [3,38]. Their omission systematically inflates perceived readiness for real-world deployment and obscures trade-offs between accuracy, cost, and operational resilience.

Table 15 ranks evaluation metrics and red-teaming methodologies according to diagnostic power, operational realism, scalability, and suitability for deployment in security-critical environments.

**Table 15:** Critical ranking of security metrics and red-teaming methodologies for LLM-based cybersecurity systems.

| Evaluation Method | Primary Focus | Diagnostic Power | Operational Realism | Scalability | Deployment Suitability | Overall Rank |
|---|---|---|---|---|---|---|
| **Detection Accuracy/F1** | Capability-centric performance | Medium | Low | High | Benchmark-driven IDS | **Medium** |
| **Exploit Success Rate** | Offensive capability validation | High | Medium | Medium | Malware/exploit analysis | **High** |

(Continued)

**Table 15 (continued)**

| Evaluation Method | Primary Focus | Diagnostic Power | Operational Realism | Scalability | Deployment Suitability | Overall Rank |
|---|---|---|---|---|---|---|
| **Refusal Consistency Metrics** | Safety alignment stability | High | Low | High | Policy enforcement evaluation | **Medium** |
| **Hallucination & Calibration Metrics** | Reasoning reliability | High | Medium | Medium | SOC automation, explanation systems | **High** |
| **False-Positive Escalation Rate** | Operational risk amplification | Very High | High | Low | SOC pipelines, IDS triage | **Very High** |
| **Latency/Inference Overhead** | Runtime feasibility | High | High | Medium | Real-time and hybrid IDS | **High** |
| **Manual/Human Red-Teaming** | Semantic failure discovery | Very High | High | Low | Safety-critical analysis | **High** |
| **Automated Jailbreak Testing** | Alignment robustness | Medium | Low | Very High | Cross-model comparison | **Medium** |
| **Prompt Injection Stress Tests** | Pipeline integrity | High | Medium | High | RAG/agentic systems | **High** |
| **Scenario-Based Red-Teaming** | System-level realism | Very High | Very High | Low | SOC, IDS, CPS deployments | **Very High** |
| **Lifecycle/Update Robustness Testing** | Security drift over time | High | High | Low | Production deployments | **High** |
| **Multimodal Red-Teaming** | Cross-modal safety | Medium | Medium | Medium | Vision-language systems | **Medium** |

Note: **Diagnostic Power:** Ability to expose non-trivial failure modes (hallucination, cascading errors). **Operational Realism:** Alignment with real SOC, IDS, or CPS deployment conditions. **Deployment Suitability:** Relevance for operational cybersecurity systems rather than laboratory benchmarks.

### *7.5 Evaluation Synthesis*

Synthesising evidence across datasets, benchmarks, red-teaming methodologies, and evaluation metrics reveals a persistent disconnect between laboratory-oriented assessment and real-world cybersecurity deployment. Benchmark-centric accuracy measures and automated red-teaming approaches offer scalability and reproducibility, yet they systematically overestimate robustness when models are exposed to operational complexity, adversarial adaptation, and deployment constraints. In contrast, scenario-driven red-teaming

and operational risk-centric metrics provide substantially higher diagnostic value for security-critical use cases, but remain difficult to standardise, reproduce, and scale across diverse environments.

As consolidated in Tables 16 and 17, progress in LLM security evaluation depends on the adoption of multi-dimensional, lifecycle-aware assessment pipelines that jointly measure task capability, behavioural robustness, and operational risk. Without such integrated evaluation frameworks, claims regarding robustness, alignment, and deployment readiness will remain fragmented, difficult to compare across studies, and insufficiently grounded for high-assurance cybersecurity deployment.

**Table 16:** Comparative overview of LLM security evaluation frameworks and red-teaming practices.

| Evaluation Category | Key References | Primary Evaluation Focus | Key Gaps and Limitations |
|---|---|---|---|
| **Security Benchmarks & Datasets** | [6,25,137,138] | Phishing realism; jailbreak robustness; multilingual deception; refusal behaviour; malware reasoning | Predominantly static and text-centric; weak multimodal coverage; limited adaptive adversary modelling; poor alignment with SOC telemetry |
| **Red-Teaming & Stress Testing** | [15,32,111] | Alignment bypass; multi-turn jailbreaks; prompt injection in RAG and agentic systems; black-box robustness | Lack of standardised protocols; non-deterministic outcomes; weak lifecycle and multi-agent evaluation; inconsistent severity reporting |
| **Metrics & Evaluation Methodology** | [3,16,28,164] | Refusal consistency; exploit success; hallucination and calibration metrics; composite safety indices | Highly context-dependent metrics; no robustness thresholds; vendor updates invalidate scores; scarce SOC-grade operational metrics |

**Table 17:** Comparative study-level analysis of evaluation benchmarks, red-teaming methods, and metrics.

| Study | Evaluation Subcategory | Model/Setting | Methodological Distinction | Demonstrated Strengths/ Performance Gains | Observed Limitations and Applicable Scenarios |
|---|---|---|---|---|---|
| Opara et al. [6] | Security Benchmark | Stylometry-aware phishing corpora | Adversarially diverse stylometric benchmarking | Reveals detector fragility under paraphrasing and stylistic variation | Primarily email-focused; limited multilingual and multimodal realism |

(Continued)

**Table 17 (continued)**

| Study | Evaluation Subcategory | Model/Setting | Methodological Distinction | Demonstrated Strengths/ Performance Gains | Observed Limitations and Applicable Scenarios |
|---|---|---|---|---|---|
| Zhang et al. [25] | Cross-Domain Benchmark | Comprehensive LLM security suite | Multi-task evaluation spanning jailbreaks, refusal, and exploit reasoning | One of the broadest comparative security benchmarks | Static tasks; lacks SOC-context and lifecycle realism |
| Nguyen et al. [137] | Security Dataset | Exploit and harm assessment corpora | Explicit modelling of harmful query execution chains | Improves reproducibility of exploit-oriented evaluations | Does not capture adaptive attacker evolution |
| Sezgin [138] | Scenario Benchmark | Narrative deception datasets | Scenario-driven, multi-step adversarial escalation | Captures progressive manipulation better than static prompts | Synthetic narratives limit ecological validity |
| Heiding et al. [100] | Adversarial Spam Benchmark | High-variance phishing datasets | Stress-testing of semantic and structural perturbations | Demonstrates sharp degradation of spam filters | Restricted to email-based attacks |
| Kulkarni et al. [164] | Misinformation Benchmark | Multilingual persuasion corpora | Cross-lingual deception robustness testing | Highlights cultural susceptibility differences | Text-only misinformation; no multimodal content |
| Toth et al. [165] | Secure Coding Benchmark | Cross-language vulnerability datasets | Language-wise security benchmarking | Exposes uneven security across programming languages | Limited coverage of memory-corruption flaws |
| Rondanini et al. [105] | Malware Benchmark | Edge-focused malware corpora | Evaluation under compute-constrained settings | Strong accuracy on mobile/edge platforms | Lacks adversarially generated malware |

(Continued)

**Table 17 (continued)**

| Study | Evaluation Subcategory | Model/Setting | Methodological Distinction | Demonstrated Strengths/ Performance Gains | Observed Limitations and Applicable Scenarios |
|---|---|---|---|---|---|
| Yao et al. [15] | Red-Teaming Taxonomy | General-purpose LLMs | Structured categorisation of jailbreak techniques | Provides unified attack taxonomy | Limited quantitative evaluation |
| Villa et al. [32] | Jailbreak Benchmarking | Open and proprietary LLMs | Large-scale black-box jailbreak evaluation | Shows non-linear size–risk relationship | English-only; lacks multimodal testing |
| Hilario et al. [26] | Conversational Red-Teaming | Human-guided multi-turn attacks | Context layering and narrative manipulation | Uncovers failures unseen in single-turn prompts | Labour-intensive and difficult to scale |
| Greshake et al. [166] | Prompt Injection Analysis | Instruction-following LLMs | Indirect injection through templates | Effective even against hardened prompts | Limited evaluation of agentic/RAG systems |
| De Maio et al. [27] | Lifecycle Stress-Testing | RAG and agent pipelines | Cross-component vulnerability propagation analysis | Exposes system-level fragility | Limited quantitative severity metrics |
| Xue et al. [111] | Jailbreak Attacks | Transformer LLMs | Dual-path syntactic-semantic attacks | Significantly higher success than single-vector attacks | Limited testing on small models |
| Huang et al. [139] | Capability Stress-Testing | Foundation LLMs | Boundary probing for unsafe behaviours | Identifies emergent alignment failures | No segmentation by alignment method |
| Dharmendra et al. [167] | Human-Guided Red-Teaming | Multiple LLMs | Interactive adversarial probing | Finds deeper vulnerabilities than automation | High cost; poor scalability |

(Continued)

**Table 17 (continued)**

| Study | Evaluation Subcategory | Model/Setting | Methodological Distinction | Demonstrated Strengths/ Performance Gains | Observed Limitations and Applicable Scenarios |
|---|---|---|---|---|---|
| Zaydi et al. [168] | Adversarial Stress-Testing | Mixed-model environments | Agent-enabled attack exploration | Reveals weaknesses in tool-augmented systems | Limited RAG-pipeline coverage |
| Zhang et al. [16] | Safety Metrics | General LLMs | Severity-weighted harm scoring | Nuanced risk stratification | No normative deployment thresholds |
| Jaffal et al. [3] | Composite Metrics | Multi-task evaluation | Integrated failure-mode indices | Identifies cross-domain vulnerabilities | Dataset heterogeneity affects comparability |
| Nguyen et al. [137] | Exploit Metrics | Exploit - evaluation models | Exploit success probability metrics | Operationally meaningful exploit evaluation | Limited multi-model scope |
| Chen et al. [28] | Calibration Metrics | General LLMs | Hallucination and consistency scoring | Improves explanation faithfulness | Limited cross-lingual validation |
| Kulkarni et al. [164] | Cross-Lingual Metrics | Multilingual LLMs | Cultural deception susceptibility scoring | Exposes strong language-wise variance | Text-only scope |
| Omar et al. [169] | Synthetic Behaviour Metrics | Narrative models | Persuasion intensity and deception fidelity | LLMs approach human-level persuasion | Lacks real-world influence validation |

## 8 Methodological Orientation in LLM-Cybersecurity Research

Beyond the thematic axes of AIGC-driven threats, defensive capabilities, and security evaluation, the reviewed literature exhibits substantial methodological diversity. This cross-cutting methodological dimension is critical for assessing evidentiary strength, reproducibility, and operational relevance. Across the 167 peer-reviewed studies included in this survey, methodological orientations converge around four dominant classes: (i) empirical studies that measure LLM behaviour under controlled or adversarial conditions, (ii) system and architectural proposals that translate LLM capabilities into operational cybersecurity pipelines, (iii) conceptual and analytical contributions that articulate theoretical, socio-technical, and lifecycle risk

frameworks, and (iv) survey and meta-analytic works that consolidate and structure the rapidly expanding research landscape. Each methodological class contributes distinct forms of insight while exhibiting characteristic limitations that influence the maturity and reliability of LLM-driven cybersecurity research.

### *8.1 Empirical Studies*

Empirical research constitutes the largest methodological category within the reviewed corpus, reflecting the field's strong emphasis on behavioural evaluation, attack modelling, and detection assessment. These studies span a wide range of threat and defence scenarios, including phishing and impersonation generation [6,7,13], exploit reasoning and code-generation misuse [9,10], malware behavioural modelling [19,105], and detection enhancement through prompt-learning approaches for network intrusion detection systems [64] or phishing detection [63]. Empirical red-teaming further encompasses scenario-driven deception evaluation [138], dual-path jailbreak attacks [111], multimodal and conversational manipulation [26], and cross-architecture jailbreak robustness testing [32,116].

Despite addressing diverse threat landscapes, empirical studies consistently reveal methodological constraints. Common limitations include reliance on small or synthetic datasets, limited cross-lingual and multimodal evaluation, insufficient modelling of adaptive adversaries, lack of replication across model updates, and weak ecological validity. Many studies employ constrained prompt templates or narrow experimental testbeds that fail to capture realistic operational attacker behaviour. Consequently, empirical research provides high-resolution snapshots of LLM capabilities and vulnerabilities, but often lacks longitudinal depth and generalisability.

### *8.2 System and Framework Proposals*

System and architectural proposals focus on operationalising LLM capabilities within cybersecurity workflows, including threat intelligence extraction pipelines, SOC triage assistants, autonomous defence agents, and hybrid LLM-ML architectures. Representative contributions include named-entity-recognition-enhanced IOC extraction [20], multi-agent investigative systems for incident response [3,24,65], graph-augmented reasoning for threat-path reconstruction [131], autonomous containment agents [130], and domain-adapted intrusion detection architectures [103,106]. Notably, CTI platforms such as the **RAG-CDI framework** [55] and the real-time detection system **WatchOverGPT** [56] demonstrate the practical integration of LLMs for intelligence fusion. Additional system-level studies propose **customisable LLM frameworks** for vulnerability detection and repair, alert prioritisation mechanisms such as **TailorAlert**, and interpretable frameworks for cybersecurity risk assessment [60].

While these proposals demonstrate substantial potential for applied LLM cybersecurity, they also expose systemic fragilities. Recurrent issues include dependence on brittle prompt patterns, susceptibility to prompt injection in multi-component pipelines, hallucination propagation across chained reasoning steps, and over-reliance on curated retrieval corpora. Only a small fraction of systems are evaluated in live SOC environments or under sustained adversarial pressure; most rely on offline logs, synthetic scenarios, or limited testbeds. As a result, although systems research outlines plausible pathways toward deployment, the maturity and reliability of existing prototypes remain insufficient for high-assurance operational use.

### *8.3 Conceptual and Analytical Studies*

Conceptual and analytical studies contribute theoretical depth by examining socio-technical implications, risk taxonomies, lifecycle vulnerabilities, and governance structures associated with LLM deployment. Prominent analyses investigate AI-enabled persuasion and its behavioural foundations [13,123], lifecycle propagation of prompt injection in retrieval-augmented and agentic ecosystems [27], and multi-dimensional

risk structures governing safety and misuse [3]. Other contributions address **protecting code with LLMs** [134], the use of **generative AI for automated security testing** [62], the fragility of stylometric and watermark-based detection [149], and ethical considerations surrounding automated reasoning and decision support.

These studies excel in conceptual clarity and systemic perspective, yet frequently lack empirical grounding or validation under realistic adversarial conditions. Risk intensities are rarely quantified, formal evaluation protocols are often absent, and integration with behavioural experiments is limited. Nonetheless, conceptual and analytical works provide indispensable frameworks for interpreting systemic vulnerabilities and long-term risks that empirical studies alone cannot reveal.

### *8.4 Survey and Meta-Analytic Contributions*

Survey and meta-analytic studies synthesise longitudinal developments, identify methodological gaps, and map the evolving landscape of LLM-centric cybersecurity research. Representative examples include comprehensive surveys of AIGC-driven cyber threats [33–36,38], analyses of LLM safety and robustness [16], systematic reviews of adversarial prompting strategies [15], and surveys focusing on LLM-enabled intrusion detection and CTI processing [18,48,50,132].

However, methodological transparency varies considerably across these works. Many surveys rely heavily on preprints, lack PRISMA alignment, or aggregate heterogeneous results without normalising evaluation conditions. Only a limited subset employs explicit inclusion criteria, reproducible search strategies, or structured comparative synthesis. The present survey addresses these shortcomings by adhering to PRISMA guidelines, restricting analysis to peer-reviewed studies, and integrating cross-dimensional evidence spanning threats, defences, and evaluation frameworks.

### *8.5 Methodological Synthesis and Implications*

The interaction among these methodological orientations reveals a research field advancing rapidly yet unevenly. Empirical studies offer essential behavioural evidence but require more realistic datasets, multimodal evaluation, and longitudinal replication. System proposals demonstrate operational promise but remain constrained by hallucination risks, dependency fragilities, and unresolved security vulnerabilities. Conceptual analyses articulate systemic risks but require empirical validation to inform actionable guidance. Survey studies provide structural clarity yet must adopt more rigorous and standardised methodologies to reduce interpretive variance.

Integrating these orientations within the unified taxonomy proposed in this work enables a coherent assessment of research maturity. It highlights the need for stronger methodological coupling, including empirical validation of conceptual frameworks, system-level evaluation using adversarial benchmarks, and surveys incorporating meta-analytic quantification. Progress toward high-assurance deployment requires reproducible, operationally grounded, and longitudinal methodologies capable of capturing both dynamic adversarial behaviour and the rapid evolution of LLM architectures.

### *8.6 Interpretation of Methodological Mapping*

The distribution of studies in Table 18 reveals pronounced methodological asymmetries across LLM-cybersecurity domains. Research on AIGC-driven threats is predominantly empirical, reflecting the feasibility of controlled phishing, malware, jailbreak, and misinformation experiments, albeit with limited ecological and longitudinal validation. Defensive research exhibits a more balanced mix of empirical and system-oriented approaches, yet many proposed architectures remain at the prototype stage and lack rigorous

evaluation in operational environments. Evaluation and red-teaming research is the most diverse but also the most fragmented, underscoring the absence of unified standards, reproducible protocols, and cross-model longitudinal tracking.

**Table 18:** Methodological mapping of studies across threat, defence, and evaluation dimensions.

| Study | A | B | C | D | E |
|---|---|---|---|---|---|
| **A. AIGC-Driven Threats** | | | | | |
| Opara et al. [6] | Threats | Phishing | ✓ | – | – |
| Schmitt and Flechais [13] | Threats | Influence Ops | ✓ | – | ✓ |
| Aseeri and Bohacek [7] | Threats | Spear-Phishing | ✓ | – | – |
| Pham et al. [98] | Threats | Targeted Phishing | ✓ | – | – |
| Koide et al. [99] | Threats | Spam Bypass | ✓ | – | – |
| Li et al. [170] | Threats | AI-Phishing Survey | – | – | ✓ |
| Shrestha et al. [102] | Threats | Adv. Messaging | ✓ | – | – |
| Weinz et al. [101] | Threats | Human Factors | ✓ | – | ✓ |
| Iturbe et al. [8] | Threats | Polymorphic Malware | ✓ | – | – |
| Yamin et al. [9] | Threats | Malware Generation | ✓ | – | – |
| Rondanini et al. [160] | Threats | Edge Malware | ✓ | – | – |
| Devadiga et al. [107] | Threats | Exploit Derivation | ✓ | – | – |
| Maity and Arora [116] | Threats | Jailbreaking | ✓ | – | – |
| Villa et al. [32] | Threats | Jailbreak Survey | ✓ | – | – |
| Pearce et al. [117] | Threats | Prompt Injection | ✓ | – | – |
| Greshake et al. [166] | Threats | Indirect Injection | ✓ | – | – |
| De Maio et al. [27] | Threats | Lifecycle Risks | – | – | ✓ |
| Doumanas et al. [149] | Threats | Watermark Weakness | – | – | ✓ |
| Harris et al. [123] | Threats | Fake News Gen. | ✓ | – | – |
| Kuntur et al. [122] | Threats | Disinfo | ✓ | – | – |
| Valdez et al. [14] | Threats | Cross-Cultural Misinfo | ✓ | – | – |
| **B. LLM-Based Defensive Capabilities** | | | | | |
| Elouardi et al. [18] | Defence | IDS Survey | – | – | ✓ |
| Koide et al. [63] | Defence | Phishing Detection | ✓ | ✓ | – |
| Li and Gong [64] | Defence | NIDS Prompting | ✓ | ✓ | – |
| Rondanini et al. [105] | Defence | Edge Malware Det. | ✓ | – | – |
| Huynh et al. [19] | Defence | Behavioural Malware | ✓ | – | – |
| Che et al. [106] | Defence | Domain-Adapted IDS | ✓ | – | – |
| Hmimou et al. [103] | Defence | Fusion IDS | ✓ | – | – |
| Zou et al. [161] | Defence | NIDS Pipelines | ✓ | – | – |
| Sarker [50] | Defence | CTI Review | – | – | ✓ |
| Liu [48] | Defence | CTI Review | – | – | ✓ |
| Sarker [132] | Defence | CTI Review | – | – | ✓ |
| Loumachi et al. [20] | Defence | CTI Extraction | ✓ | ✓ | – |
| Gandhi [55] | Defence | RAG-CTI (RAG-CDI) | ✓ | ✓ | – |
| Shahid et al. [56] | Defence | Real-time CTI | ✓ | ✓ | – |

(Continued)

**Table 18 (continued)**

| Study | A | B | C | D | E |
|---|---|---|---|---|---|
| Xu et al. [128] | Defence | RAG-CTI Summ. | ✓ | ✓ | – |
| Srinivas et al. [23] | Defence | SOC Assistant | ✓ | ✓ | – |
| Alharthi and Yasaei [171] | Defence | CTI Automation | ✓ | ✓ | – |
| de Fitero-Dominguez et al. [22] | Defence | Patch Gen. | ✓ | – | – |
| Babaey and Ravindran [59] | Defence | SQLi Defense | ✓ | ✓ | – |
| Che et al. [133] | Defence | Code Analysis | ✓ | – | – |
| Moorthy et al. [120] | Defence | Vuln. Discovery | ✓ | – | ✓ |
| Chu et al. [134] | Defence | Code Protection | – | – | ✓ |
| Corchado et al. [62] | Defence | Automated Sec Test | – | – | ✓ |
| Andreoni-LN et al. [24] | Defence | Multi-Agent IR | – | ✓ | – |
| Krishnamurthy et al. [65] | Defence | Agent Teams for IR | – | ✓ | – |
| Seo et al. [60] | Defence | Risk Assessment | – | ✓ | – |
| Jaffal et al. [3] | Defence | SOC Agents | ✓ | ✓ | – |
| Ibrahim and Kashef [131] | Defence | Threat Graph Reasoning | ✓ | ✓ | – |
| Zhai et al. [130] | Defence | Response Agents | ✓ | ✓ | – |
| Yang et al. [129] | Defence | Policy Verification | ✓ | – | – |
| Alawida et al. [33] | Defence | General Review | – | – | ✓ |
| Lopez et al. [35] | Defence | General Review | – | – | ✓ |
| Uddin et al. [38] | Defence | General Review | – | – | ✓ |
| **C. Evaluation Frameworks, Benchmarks, and Red-Teaming** | | | | | |
| Zhang et al. [25] | Eval. | Sec. Benchmark | ✓ | – | – |
| Kulkarni et al. [164] | Eval. | Misinfo Benchmark | ✓ | – | – |
| Nguyen et al. [137] | Eval. | Harm/Vuln. Data | ✓ | – | – |
| Sezgin [138] | Eval. | Deception Bench. | ✓ | – | – |
| Yao et al. [15] | Eval. | Red-Teaming Survey | – | – | ✓ |
| Hilario et al. [26] | Eval. | Conv. Red-Teaming | ✓ | – | – |
| Xue et al. [111] | Eval. | Dual-Path Jailbreak | ✓ | – | – |
| Dharmendra et al. [167] | Eval. | Human-Guided RT | ✓ | – | – |
| Zaydi and Maleh [168] | Eval. | Adv. Stress-Testing | ✓ | – | – |
| Villa et al. [32] | Eval. | Jailbreak Testing | ✓ | – | – |
| De Maio et al. [27] | Eval. | Lifecycle Risk | – | – | ✓ |
| Jaffal et al. [3] | Eval. | Risk Taxonomy | – | – | ✓ |
| Zhang et al. [16] | Eval. | Alignment Metrics | – | – | ✓ |
| Chen et al. [28] | Eval. | Calibration Metrics | ✓ | – | – |
| Omar et al. [169] | Eval. | Synthetic Behaviour | ✓ | – | ✓ |
| Alawida et al. [33] | Eval. | Governance Review | – | – | ✓ |
| Karras et al. [4] | Eval. | Governance/Risk | – | – | ✓ |

Note: **A** = Research Dimension; **B** = Taxonomy Subcategory; **C** = Empirical Study (e.g., experiments, data collection, measurement); **D** = System or Framework Contribution (e.g., prototype, tool, concrete pipeline); **E** = Conceptual, Analytical, or Survey Contribution (e.g., risk model, systematic review, governance analysis).

These patterns expose both fragmentation and clear priorities for future work: the development of standardised evaluation pipelines, multimodal and multilingual datasets, field-tested system deployments, and integrated methodologies that bridge empirical, system-level, and conceptual insights. Only through such convergence can the field advance toward reliable, security-critical deployment of LLM-driven cybersecurity capabilities.

## 9 Cross-Cutting Challenges and Open Research Gaps

The integrated analysis of AIGC-driven threats, LLM-enabled defensive mechanisms, and security evaluation frameworks reveals a set of deeply interdependent challenges that constrain the secure, reliable, and accountable adoption of LLMs within cybersecurity ecosystems. These challenges do not arise from isolated technical deficiencies; rather, they emerge from the interaction of probabilistic language models with adversarial environments, complex system integrations, and incomplete evaluation and governance practices. Despite rapid advances in individual components, the field remains characterised by methodological fragmentation, uneven empirical grounding, limited operational realism, and a pace of adversarial innovation that consistently outstrips defensive maturity.

Fig. 3 conceptualises these weaknesses as a tightly coupled system of risks rather than independent failure modes. Hallucination and reliability failures interact with adversarial fragility, data governance and privacy risks, evaluation gaps, and regulatory constraints, collectively amplifying systemic vulnerability. Consequently, improvements in a single dimension—such as detection accuracy or automation efficiency—do not necessarily translate into safer or more trustworthy deployments when other dimensions remain underdeveloped.

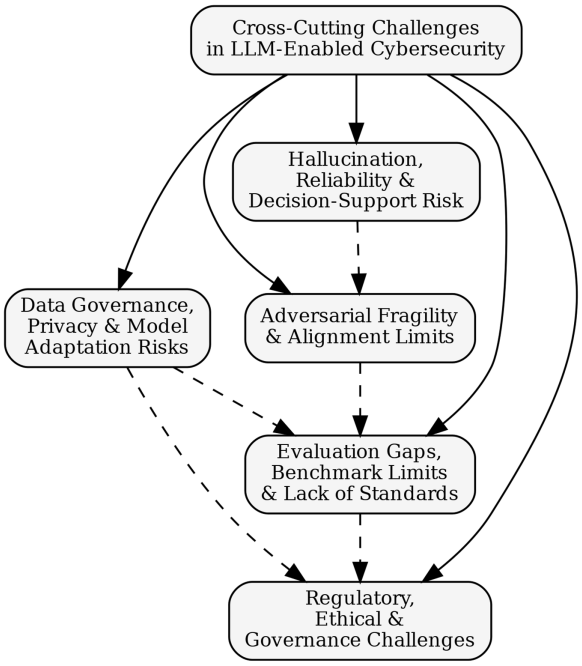


**Figure 3:** Conceptual overview of cross-cutting challenges in LLM-enabled cybersecurity, highlighting interrelated issues spanning hallucination and reliability, adversarial fragility, data governance and privacy, evaluation gaps, and governance constraints.

### *9.1 Scenario-Grounded Limitations and Open Challenges*

When examined under realistic deployment conditions, LLM-based cybersecurity systems exhibit persistent limitations that extend beyond model architecture or training strategy. Empirical evidence indicates that many shortcomings arise from the coupling of LLM reasoning with operational constraints, incomplete context, and adaptive adversarial behaviour [33,35,38]. Grounding these limitations in concrete scenarios—SOC automation, intrusion detection, and agentic response systems—clarifies why benchmark-level gains frequently fail to translate into dependable operational performance.

In SOC automation and threat intelligence pipelines, LLMs are increasingly employed for alert triage, IOC extraction, report generation, and cross-source correlation [20,23,132]. While these applications reduce analyst workload and accelerate situational awareness, they introduce operational risks that are largely invisible to accuracy-centric evaluations. A central concern is false-positive escalation: hallucinated or weakly supported indicators may be amplified through automated prioritisation workflows, increasing alert fatigue and diverting analyst attention from genuinely critical incidents [35,38]. As reflected in the comparative analyses in Table 13, retrieval-augmented generation pipelines are particularly sensitive to corpus quality, temporal relevance, and retrieval completeness [55,56]. In large-scale SOC environments characterised by heterogeneous, noisy, and evolving telemetry, even minor grounding errors can propagate across downstream automation, undermining trust and decision quality [36,50].

Comparable limitations arise in LLM-assisted intrusion and anomaly detection. Semantic reasoning over logs, traces, and traffic summaries improves detection of complex or low-and-slow attacks under controlled conditions [18,19,103], yet deployment-scale robustness remains limited. As shown in Tables 7 and 16, most IDS-oriented studies rely on curated or synthetic datasets that inadequately reflect class imbalance, noise, and concept drift in enterprise environments [35,38]. Systems optimised for benchmark performance therefore risk elevated false-positive rates and reduced sensitivity to novel or evolving attacks in production [39,105]. Moreover, prompt-based and instruction-tuned IDS components remain sensitive to input formatting and contextual framing, raising robustness concerns when exposed to malformed logs, partial traces, or adversarially manipulated inputs [64]. These constraints currently limit the viability of LLM-based IDS for high-throughput or real-time deployment without continuous human oversight.

The limitations are most acute in autonomous and agentic response systems. Multi-agent LLM frameworks aim to enable closed-loop defence through automated investigation, containment planning, and response execution [3,24]. However, comparative evidence in Tables 12 and 13 highlights substantial safety and verification challenges. Coordination failures between agents, propagation of hallucinated reasoning steps, and ambiguity in natural-language policy interpretation can lead to inappropriate or unsafe response actions [130,131]. Unlike rule-based automation, LLM-driven agents lack formal correctness and safety guarantees, complicating verification under adversarial or unforeseen conditions [35,38]. These risks are particularly severe in cyber-physical and critical infrastructure environments—such as smart grids and industrial control systems—where erroneous actions may cause service disruption or physical harm [47]. Consequently, current agentic defence architectures are best understood as decision-support systems rather than fully autonomous responders.

Beyond technical constraints, the literature reveals socio-technical barriers that directly constrain trustworthy deployment and are central to **RQ4**. Integration costs remain substantial, encompassing compute-intensive inference, continuous model adaptation, RAG corpus curation, and tooling integration within existing SOC infrastructures [18,35,38]. Analyst trust calibration further complicates adoption, as probabilistic outputs, non-deterministic explanations, and hallucination risk can lead to either over-reliance or systematic distrust [23,56]. Organisational resistance is particularly pronounced in regulated and safety-critical environments, where accountability, auditability, and liability concerns limit acceptance

of AI-assisted workflows lacking enforceable oversight [35,38]. Together, these frictions explain why many LLM-based defences remain confined to decision-support roles despite strong benchmark performance.

### *9.2 Hallucination, Reliability, and Decision-Support Risks*

Across nearly all LLM-enabled cybersecurity workflows, hallucination and output instability represent foundational risks. Fabricated indicators, incorrect remediation guidance, and inconsistent vulnerability explanations can mislead analysts when LLMs are embedded in SOC pipelines, CTI extraction, or multi-agent coordination. Empirical studies show that grounding failures in retrieval-augmented systems propagate confident yet incorrect conclusions [128], while surveys of generative AI for CTI report persistent reliability risks during data processing and report generation [132]. Similar issues arise in vulnerability repair [22] and automated SOC triage [23], with error rates increasing under incomplete telemetry, evolving malware semantics, and collaborative agent workflows [19,24,103,129,130]. Despite emerging work on calibration and uncertainty estimation, security-specific reliability frameworks capable of supporting high-assurance decision-making remain largely absent.

### *9.3 Adversarial Fragility and Alignment Limits*

Adversarial prompting, jailbreaks, and prompt injection remain among the most consistently demonstrated vulnerabilities. Large-scale evaluations show that role-play manipulation, persona steering, and multi-turn coercion can bypass alignment safeguards even in heavily guarded models [32,111]. Obfuscation-based and syntactic jailbreaks exhibit high transferability across model sizes and training regimes [116]. Indirect prompt injection—via untrusted documents, code, or retrieved web content—remains particularly effective in RAG and tool-augmented systems [117,166]. Lifecycle-oriented analyses further demonstrate how injection vectors propagate across retrieval chains, agent communication layers, and automated planners [27]. Collectively, these findings indicate that model-level alignment is insufficient in isolation and must be complemented by system-level safeguards, including input validation, provenance tracking, sandboxing, and continuous red-teaming.

### *9.4 Data Governance, Privacy, and Model Adaptation Risks*

Data governance and privacy risks remain substantially under-addressed in LLM-based cybersecurity systems. SOC logs, malware traces, and proprietary source code frequently contain sensitive operational information, creating exposure risks through memorisation, cross-user leakage, and unintended disclosure [33]. Studies on organisational fine-tuning and domain adaptation show that incorporating sensitive datasets into LLM training pipelines introduces additional compliance and leakage risks [120]. Lifecycle analyses further demonstrate how privacy vulnerabilities propagate across chained RAG and multi-agent systems [27]. Despite these concerns, the surveyed literature reports limited adoption of differential privacy, secure inference, federated deployment, or redaction-aware training, leaving organisations exposed to unresolved regulatory and operational risks.

### *9.5 Evaluation Gaps and the Absence of Standards*

Security evaluation practices remain fragmented and insufficiently standardised. Existing benchmarks capture isolated threat vectors—such as phishing realism [6], harmful-query handling [137], or scenario-based deception [138]—but rarely incorporate multimodal, multilingual, or adaptive adversarial dynamics. Red-teaming methodologies vary widely, from manual probing to automated jailbreak testing and multi-step manipulation, yet lack shared coverage criteria or reporting standards [15,26,32]. Frequent vendor-side

model updates further invalidate prior evaluation results, underscoring the need for lifecycle-integrated and continuously updated assessment frameworks.

### *9.6 Regulatory, Ethical, and Governance Challenges*

Regulatory and governance frameworks lag behind the technical complexity of LLM deployment in cybersecurity contexts. Existing policies emphasise transparency and accountability but provide limited guidance on adversarial robustness, autonomous agent behaviour, or cross-organisational data governance. Analyses of digital deception and systemic risk highlight gaps in current governance models for managing interconnected LLM pipelines and agentic systems [13,33]. Ethical principles such as fairness, explainability, and auditability remain difficult to operationalise in security-critical environments, leaving unresolved questions around liability, oversight, and risk ownership.

### *9.7 Synthesis of Cross-Cutting Challenges*

Across the surveyed literature, a consistent pattern emerges: AIGC-enabled offensive capabilities evolve more rapidly than defensive maturity, evaluation frameworks lack operational breadth and standardisation, and governance mechanisms trail both technical innovation and adversarial adaptation. Reliability failures, fragmented red-teaming practices, and limited lifecycle-oriented risk modelling collectively amplify systemic fragility. Addressing these gaps requires an integrated, end-to-end perspective that combines robust threat modelling, provenance validation, privacy-preserving design, continuous adversarial evaluation, and standardised metrics aligned with real-world deployment constraints.

## 10 Future Research Directions

The synthesis of 167 peer-reviewed studies spanning AIGC-driven threats, LLM-enabled defensive capabilities, and security evaluation frameworks reveals a research landscape that is advancing rapidly yet remains structurally imbalanced. While offensive applications of LLMs—including phishing automation [6], polymorphic malware and exploit generation [8], jailbreak and alignment evasion [32,116], and large-scale misinformation and influence operations [13,14]—have progressed in sophistication, scalability, and autonomy, corresponding advances in defensive systems, evaluation methodologies, and governance mechanisms have not occurred at a comparable pace. As demonstrated throughout Sections 5–7, this imbalance manifests in brittle defensive deployments, fragmented and benchmark-centric evaluation practices, and governance discussions that remain weakly coupled to enforceable technical controls.

A central finding of this review is that many limitations observed in current LLM-based cybersecurity systems do not arise from isolated algorithmic deficiencies, but from systemic gaps spanning the model–system–evaluation–governance lifecycle. Evaluation practices remain difficult to compare due to inconsistent metrics, narrow benchmark assumptions, and limited consideration of adaptive adversarial behaviour. Concurrently, emerging deployment paradigms—such as retrieval-augmented generation (RAG), tool-augmented reasoning, and multi-agent defence architectures—introduce novel failure modes that are insufficiently addressed by existing assurance and governance frameworks [27]. These structural asymmetries amplify systemic risk and motivate the need for coordinated, evidence-driven research priorities that integrate technical, organisational, and regulatory perspectives.

To address these challenges in a principled and actionable manner, this review advances a structured future research agenda that explicitly links empirically observed weaknesses to targeted research interventions. Table 19 consolidates these priorities into a unified roadmap organised around five interdependent research pillars, capturing the technical, operational, evaluative, and governance dimensions of LLM-enabled cybersecurity. Although the pillars are framed in technical terms, their realisation inherently depends

on interdisciplinary enablers—most notably human–computer interaction (HCI), cognitive ergonomics, policy studies, and regulatory science—particularly for analyst-facing systems and compliance-critical deployments.

**Table 19:** Roadmap of key research directions and solutions for LLM-enabled cybersecurity.

| Research Pillar | Key Challenges | Research Directions |
|---|---|---|
| **1. Domain-Specialised & Security-Aligned LLMs** | Hallucinated exploit reasoning and inconsistent remediation in generic models [16]; limited grounding in security ontologies (e.g., CVE, ATT&CK, SIGMA); brittleness of RAG pipelines in CTI extraction [55,128]; need for customisable vulnerability detection and repair. | Neuro-symbolic architectures integrating attack graphs and security taxonomies; security-aware pre-training and retrieval-anchored verification; policy-constrained decoding enforcing cyber-specific rules; memory-scoped and isolation-aware reasoning; auditable reasoning traces for forensic analysis. |
| **2. Robust Multi-Agent & Tool-Augmented Systems** | Error and hallucination propagation across agent teams [24]; tool misuse and privilege escalation [117]; lack of verifiable inter-agent communication protocols [65]; fragility in alert prioritisation and risk assessment [60]. | Formal agent-role and privilege specifications; symbolic verification and model checking of workflows; sandboxed tool execution with runtime invariants; confidence-aware human-in-the-loop escalation for high-impact decisions. |
| **3. Next-Generation Evaluation & Red-Teaming** | Static, text-centric benchmarks; lack of adversary-adaptive and longitudinal robustness assessment; fragmented reporting of robustness, refusal fidelity, and hallucination metrics. | Multilingual and multimodal adversary-adaptive benchmarks; automated red-teaming agents with evolving attack strategies; longitudinal audits across model updates; shared repositories of failure cases and safety regressions. |
| **4. Privacy, Data Governance & Lifecycle Management** | Memorisation and leakage of sensitive SOC data [33]; vulnerability propagation across RAG layers and agent pipelines [27]. | Differential privacy for SOC telemetry; secure enclaves and encrypted inference; redaction-aware training; tamper-evident provenance and audit logs; lifecycle-aware data governance policies. |
| **5. Governance, Regulation & Accountability** | Misalignment between technical capabilities and legal expectations [13]; absence of certification standards; limited mechanisms for auditing autonomous decisions. | Certification and conformity standards for LLM-based defence tools; risk-scoring aligned with regulatory harm categories; traceable audit interfaces; multi-stakeholder governance frameworks. |

Crucially, the proposed agenda emphasises that meaningful progress must extend across the full LLM lifecycle. Early-stage advances in domain-specialised and security-aligned model design must

be complemented by adversary-aware evaluation, system-level verification, runtime monitoring, and compliance-ready governance mechanisms. As evidenced by the scenario-grounded limitations discussed in Section 6.9, failure to address any single layer—such as evaluation realism, human trust calibration, or policy enforcement—can undermine otherwise strong technical performance. By grounding each research direction in recurring limitations identified across the evidence base, the roadmap promotes methodologically rigorous advances while encouraging deployment practices that integrate technical safeguards with procedural, organisational, and human-centred controls.

Fig. 4 visualises this agenda as a vertically structured roadmap organised around five interconnected thematic pillars: (i) domain-specialised and security-aligned LLMs, (ii) robust multi-agent and tool-augmented defence architectures, (iii) next-generation evaluation and red-teaming ecosystems, (iv) privacy-preserving data governance and lifecycle management, and (v) governance- and regulation-aligned assurance frameworks. Notably, several pillars explicitly require interdisciplinary collaboration: HCI and cognitive ergonomics are essential for explanation design, trust calibration, and analyst–LLM interaction within SOC environments, while policy studies and regulatory science underpin the translation of technical safeguards into auditable and enforceable compliance mechanisms.

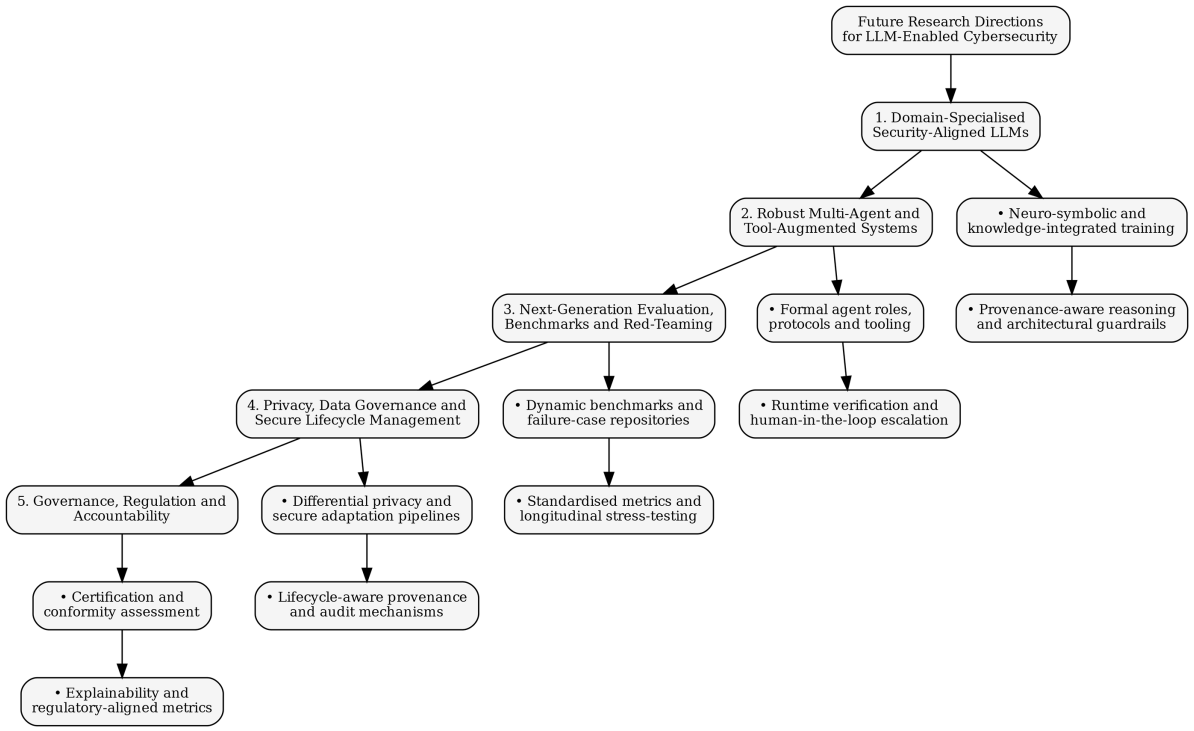


**Figure 4:** Research roadmap for LLM-enabled cybersecurity, outlining future directions across domain-specialised security-aligned models, robust multi-agent and tool-augmented systems, adversary-aware evaluation and red-teaming, privacy- and lifecycle-aware data governance, and governance- and regulation-aligned assurance frameworks.

Table 19 operationalises this vision by mapping each research pillar to the corresponding empirical challenges and prioritised research directions. A key insight emerging from this synthesis is that resilience cannot be achieved through progress along a single dimension. Instead, robust and trustworthy LLM-enabled cybersecurity will depend on coordinated advances across modelling, system integration, evaluation realism, human–system interaction design, privacy protection, and governance alignment—an insight that directly informs both near-term research investment and longer-term standardisation and policy efforts.

Beyond this high-level roadmap, the literature consistently points to the need for prioritisation across time horizons. Near-term research should focus on operational realism, evaluation standardisation, and safe hybrid deployment models that incorporate explicit human oversight. Mid-term efforts must address systemic robustness in multi-agent and tool-augmented systems, alongside privacy-preserving lifecycle management. Long-term investments are required to develop domain-specialised LLMs with formal reasoning guarantees and to establish certification and assurance regimes suitable for safety-critical and national infrastructure contexts.

To make this prioritisation explicit, Table 20 aligns the proposed research directions with the four research questions (RQ1-RQ4) and categorises them by expected time horizon and impact. This staged alignment ensures that future work not only advances the state of the art but also systematically closes the methodological, operational, and governance gaps identified throughout the review.

**Table 20:** Prioritised future research roadmap explicitly aligned with research questions (RQ1–RQ4).

| Time Horizon | Research Priority | RQ Link | Rationale and Expected Impact |
|---|---|---|---|
| **Near-Term (1–3 years)** | Operationally grounded evaluation and red-teaming ecosystems | RQ3 | Improves robustness assessment and reproducibility through adversary-adaptive, multilingual, and SOC-aligned benchmarks, directly addressing fragmentation in current evaluation practices. |
| | Hybrid IDS architectures with explicit human-in-the-loop controls | RQ2 | Reflects evidence that LLMs are most effective as post-detection reasoning layers, mitigating hallucination and false-positive escalation while enabling safe deployment. |
| | Governance-aligned deployment safeguards | RQ4 | Operationalises ethical and regulatory principles via audit logs, escalation workflows, and compliance-ready controls. |
| **Mid-Term (3–5 years)** | Verified multi-agent and tool-augmented defence systems | RQ2, RQ3 | Targets systemic failure modes such as error propagation and tool misuse, advancing verification, runtime monitoring, and controlled autonomy. |
| | Privacy-preserving lifecycle and data governance mechanisms | RQ4 | Addresses data leakage and compliance risks, enabling trustworthy deployment in regulated and long-lived environments. |
| **Long-Term (5+ years)** | Domain-specialised, security-aligned LLMs with formal reasoning guarantees | RQ1, RQ2 | Moves beyond prompt-level mitigation toward foundational solutions for hallucination, exploit reasoning errors, and misuse risk. |

(Continued)

**Table 20 (continued)**

| Time Horizon | Research Priority | RQ Link | Rationale and Expected Impact |
|---|---|---|---|
| | Standardised certification and assurance frameworks | RQ4 | Enables adoption in safety-critical contexts by linking technical robustness to governance and accountability structures. |

Note: RQ1 addresses AIGC-enabled threats, RQ2 defensive capabilities, RQ3 evaluation and benchmarking, and RQ4 governance and systemic challenges.

## 11 Conclusion

This PRISMA-guided systematic review synthesised 167 peer-reviewed studies published between 2022 and 2025 to examine the transformative impact of Large Language Models (LLMs) on cybersecurity across four tightly coupled dimensions: threat generation, defensive capabilities, security evaluation, and governance. Anchored in a unified threat–defence–evaluation taxonomy and supported by transparent screening, coding, and synthesis procedures, the review provides an evidence-driven response to the four research questions (RQ1–RQ4) that motivated this study.

In addressing **RQ1**—how LLMs enable or amplify cybersecurity threats—the review demonstrates that AIGC-driven attacks represent a qualitative shift rather than an incremental extension of prior techniques. Across phishing and social engineering, malware and exploit development, jailbreaks and prompt injection, and large-scale misinformation and influence operations, LLMs introduce unprecedented scalability, contextual awareness, and adaptability. These properties substantially lower the technical barrier for adversaries while eroding the effectiveness of traditional signature-based detection, stylometric attribution, and static defence mechanisms. The convergence of multilingual generation, semantic reasoning, and rapid iteration fundamentally reshapes the attacker–defender balance.

In response to **RQ2**, which examined the defensive role of LLMs, the evidence indicates that LLMs deliver their greatest value when deployed as *cognitive and reasoning layers* within hybrid security architectures. Meaningful advances are observed in threat intelligence extraction, SOC automation, alert correlation, anomaly interpretation, vulnerability analysis, and multi-agent investigation workflows. At the same time, the review identifies persistent and non-trivial limitations, including hallucinated indicators, false-positive escalation, sensitivity to noisy or adversarial inputs, and significant computational overhead. Crucially, LLMs rarely function effectively as primary detectors; instead, they augment conventional IDS and security tools through semantic abstraction, contextual explanation, and decision support. These findings reinforce the necessity of defence-in-depth designs and robust human-in-the-loop oversight for operational deployment.

With respect to **RQ3**, which focused on security evaluation frameworks and red-teaming methodologies, the review exposes a fragmented and methodologically uneven evaluation landscape. Existing benchmarks are predominantly text-centric, static, and weakly aligned with real-world SOC, IDS, and OT/ICS operating conditions. Red-teaming practices vary widely in scope, reproducibility, and reporting rigor, while frequent model updates undermine longitudinal robustness assessment. Comparative synthesis demonstrates that accuracy-centric metrics alone are insufficient; operational indicators such as inference latency, false-positive amplification, reasoning stability, and system-level failure propagation provide more meaningful insight into deployment readiness. The absence of standardised, adversary-aware, and lifecycle-oriented evaluation pipelines remains a critical barrier to trustworthy assessment.

Addressing **RQ4**, which examined methodological and governance challenges, the review identifies systemic gaps that constrain the safe and accountable deployment of LLM-based cybersecurity systems. These include overreliance on synthetic or laboratory datasets, limited analysis of real-world deployment dynamics, insufficient reporting of operational cost and energy impact, and weak integration between technical design decisions and regulatory obligations. Although ethical considerations are frequently discussed, they are often decoupled from enforceable controls. By explicitly mapping observed failure modes to auditable safeguards and regulatory frameworks—such as the EU NIS-2 Directive and guidance from CISA—the review demonstrates that governance must be operationalised through accountability, transparency, human oversight, auditability, and change-management mechanisms rather than treated as an abstract overlay.

Taken together, the answers to RQ1–RQ4 reveal that LLMs constitute a dual-use inflection point in cybersecurity. They simultaneously accelerate offensive sophistication and enable higher-level defensive reasoning, shifting security practice from isolated signal-level detection toward context-aware, explainable, and decision-oriented defence systems. The central contribution of this review lies in systematically integrating these dimensions through a unified taxonomy and critical synthesis, clarifying the trade-offs among capability, robustness, cost, and governance. By connecting empirical evidence with evaluation realism and regulatory alignment, this work provides a structured foundation for future research and practice aimed at developing secure, accountable, and resilient LLM-enabled cybersecurity systems.

**Acknowledgement:** The authors acknowledge support from their respective institutions.

**Funding Statement:** The author, Hamed Alqahtani extends his appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group under grant number (GRP.2/663/46).

**Author Contributions:** The authors confirm contribution to the paper as follows: Conceptualization, Hamed Alqahtani and Gulshan Kumar; methodology, Hamed Alqahtani and Gulshan Kumar; validation, Hamed Alqahtani and Gulshan Kumar; formal analysis, Hamed Alqahtani; investigation, Hamed Alqahtani; data curation, Hamed Alqahtani; writing—original draft preparation, Hamed Alqahtani; writing—review and editing, Hamed Alqahtani and Gulshan Kumar; visualization, Hamed Alqahtani; supervision, Gulshan Kumar; project administration, Gulshan Kumar. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** This article does not involve data availability, and this section is not applicable.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inform Process Syst. 2020;33:1877–901. doi:10.18653/v1/2021.mrl-1.1.
2. Alqahtani H, Kumar G. Large language models for effective detection of algorithmically generated domains: a comprehensive review. Comput Model Eng Sci. 2025;144(2):1439–79. doi:10.32604/cmes.2025.067738.
3. Jaffal NO, Alkhanafseh M, Mohaisen D. Large language models in cybersecurity: a survey of applications, vulnerabilities, and defense techniques. AI. 2025;6(9):216. doi:10.3390/ai6090216.
4. Karras A, Theodorakopoulos L, Karras C, Theodoropoulou A, Kalliampakou I, Kalogeratos G. LLMs for cybersecurity in the big data era: a comprehensive review of applications, challenges, and future directions. Information. 2025;16(11):957. doi:10.3390/info16110957.
5. Agarwal A, Trivedi A, Sharma P, Bhardwaj A. Use cases of ChatGPT and other AI tools with security concerns. In: Applications, challenges, and the future of ChatGPT. Hershey, PA, USA: IGI Global Scientific Publishing; 2024. p. 216–25. doi:10.4018/979-8-3693-6824-4.ch012.

6. Opara C, Modesti P, Golightly L. Evaluating spam filters and Stylometric Detection of AI-generated phishing emails. Expert Syst Appl. 2025;276:127044. doi:10.1016/j.eswa.2025.127044.
7. Aseeri AM, Bohacek S. Using ensembles of LLMs to detect phishing emails. In: International Conference on Advanced Information Networking and Applications. Cham, Switzerland: Springer; 2025. p. 21–34.
8. Iturbe E, Llorente-Vazquez O, Rego A, Rios E, Toledo N. Unleashing offensive artificial intelligence: automated attack technique code generation. Comput Secur. 2024;147:104077. doi:10.1016/j.cose.2024.104077.
9. Yamin MM, Hashmi E, Katt B. Combining uncensored and censored LLMS for ransomware generation. In: International Conference on Web Information Systems Engineering. Cham, Switzerland: Springer; 2024. p. 189–202.
10. Shandilya SK, Prharsha G, Datta A, Choudhary G, Park H, You I. GPT based malware: unveiling vulnerabilities and creating a way forward in digital space. In: 2023 International Conference on Data Security and Privacy Protection (DSPP). Piscataway, NJ, USA: IEEE; 2023. p. 164–73.
11. Huang K, Huang G, Duan Y, Hyun J. Utilizing prompt engineering to operationalize cybersecurity. In: Generative AI security: theories and practices. Cham, Switzerland: Springer; 2024. p. 271–303. doi:10.1007/978-3-031-54252-7_9.
12. Alauthman M, Almomani A, Aoudi S, Al-Qerem A, Aldweesh A. Automated vulnerability discovery generative AI in offensive security. In: Examining cybersecurity risks produced by generative AI. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 309–28. doi:10.4018/979-8-3373-0832-6.ch013.
13. Schmitt M, Flechais I. Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev. 2024;57(12):324.
14. Valdez HPD, Abri F, Webb J, Austin TH. Exploring the use and misuse of large language models. Information. 2025;16(9):758. doi:10.3390/info16090758.
15. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confid Comput. 2024;4(2):100211. doi:10.1016/j.hcc.2024.100211.
16. Zhang J, Bu H, Wen H, Liu Y, Fei H, Xi R, et al. When LLMS meet cybersecurity: a systematic literature review. Cybersecurity. 2025;8(1):55. doi:10.1186/s42400-025-00361-w.
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.31222/osf.io/v7gm2_v1.
18. Elouardi S, Motii A, Jouhari M, Amadou ANH, Hedabou M. A survey on Hybrid-CNN and LLMs for intrusion detection systems: recent IoT datasets. IEEE Access. 2024;12(1):180009–33. doi:10.1109/access.2024.3506604.
19. Huynh L, Zhang Y, Jayasundera D, Jeon W, Kim H, Bi T, et al. Detecting code vulnerabilities using LLMs. In: 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Piscataway, NJ, USA: IEEE; 2025. p. 401–14.
20. Loumachi FY, Ghanem MC, Ferrag MA. Advancing cyber incident timeline analysis through retrieval-augmented generation and large language models. Computers. 2025;14(67):1–42. doi:10.3390/computers14020067.
21. Sarker IH. Introduction to AI-driven cybersecurity and threat intelligence. In: AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability. Cham, Switzerland: Springer; 2024. p. 3–19.
22. de Fitero-Dominguez D, Garcia-Lopez E, Garcia-Cabot A, Martinez-Herraiz JJ. Enhanced automated code vulnerability repair using large language models. Eng Appl Artif Intell. 2024;138(1):109291. doi:10.1016/j.engappai.2024.109291.
23. Srinivas S, Kirk B, Zendejas J, Espino M, Boskovich M, Bari A, et al. AI-augmented SOC: a survey of LLMs and agents for security automation. J Cybersecur Priv. 2025;5(4):95.
24. Andreoni M, Lunardi WT, Lawton G, Thakkar S. Enhancing autonomous system security and resilience with generative AI: a comprehensive survey. IEEE Access. 2024;12(1):109470–93. doi:10.1109/access.2024.3439363.
25. Zhang J, Wu P, London J, Tenney D. Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: a comprehensive analysis. IEEE Access. 2025;13(2):28335–52. doi:10.1109/access.2025.3540075.

26. Hilario E, Azam S, Sundaram J, Imran Mohammed K, Shanmugam B. Generative AI for pentesting: the good, the bad, the ugly. Int J Inform Secur. 2024;23(3):2075–97. doi:10.1007/s10207-024-00835-x.
27. De Maio C, Di Gisi M, Fenza G, Gallo M, Loia V. A lifecycle-oriented survey of emerging threats and vulnerabilities in large language models. IEEE Access. 2025;13:176482–500. doi:10.1109/access.2025.3619764.
28. Chen C, Zhou D, Ye Y, TJj Li, Yao Y. Clear: towards contextual LLM-empowered privacy policy analysis and risk generation for large language model applications. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM; 2025. p. 277–97.
29. Campanile L, De Fazio R, Di Giovanni M, Marulli F. Beyond the hype: toward a concrete adoption of the fair and responsible use of AI. In: Ital-IA 2024: 4th National Conference on Artificial Intelligence; 2024 May 29–30; Naples, Italy; 2024. p. 60–5.
30. Humphreys D, Koay A, Desmond D, Mealy E. AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI Ethics. 2024;4(3):791–804.
31. Liu F, Jiang J, Lu Y, Huang Z, Jiang J. The ethical security of large language models: a systematic review. Front Eng Manag. 2025;12(1):128–40. doi:10.1007/s42524-025-4082-6.
32. Villa C, Mirza S, Pöpper C. Exposing the guardrails: reverse-engineering and jailbreaking safety filters in DALL·E text-to-image pipelines. In: 34th USENIX Security Symposium (USENIX Security 25). Berkeley, CA, USA: USENIX Association; 2025. p. 897–916.
33. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information. 2023;14(8):462. doi:10.3390/info14080462.
34. Esmradi A, Yip DW, Chan CF. A comprehensive survey of attack techniques, implementation, and mitigation strategies in large language models. In: International Conference on Ubiquitous Security. Cham, Switzerland: Springer; 2023. p. 76–95.
35. Lopez Delgado JL, Lopez Ramos JA. A comprehensive survey on generative AI solutions in IoT security. Electronics. 2024;13(24):4965. doi:10.3390/electronics13244965.
36. Nawara D, Kashef R. A comprehensive survey on LLM-powered recommender systems: from discriminative, generative to multi-modal paradigms. IEEE Access. 2025;13:145772–98. doi:10.1109/access.2025.3599832.
37. Cheng P, Wu Z, Du W, Zhao H, Lu W, Liu G. Backdoor attacks and countermeasures in natural language processing models: a comprehensive security review. IEEE Trans Neural Netw Learn Syst. 2025;36(8):13628–48. doi:10.1109/TNNLS.2025.3540303.
38. Uddin M, Arfeen SU, Alanazi F, Hussain S, Mazhar T, Arafatur Rahman M. A critical analysis of generative AI: challenges, opportunities, and future research directions. Arch Comput Methods Eng. 2025. doi:10.1007/s11831-025-10355-z.
39. Ferrag MA, Ndhlovu M, Tihanyi N, Cordeiro LC, Debbah M, Lestable T, et al. Revolutionizing cyber threat detection with large language models: a privacy-preserving bert-based lightweight model for IoT/IIoT devices. IEEE Access. 2024;12:23733–50. doi:10.1109/ACCESS.2024.3363469.
40. Hasanov I, Virtanen S, Hakkala A, Isoaho J. Application of large language models in cybersecurity: a systematic literature review. IEEE Access. 2024;12(1):176751–78. doi:10.1109/access.2024.3505983.
41. Alqahtani H, Kumar G. A comprehensive review of generative AI techniques and their impact on cybersecurity. Soft Comput. 2025;29(13):4945–82. doi:10.1007/s00500-025-10702-z.
42. Gunda M, Manda V, Naradasu P, Mekala S, Bhattacharya S. ChatGPT in cyber onslaught and fortification: past, present, and future. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). Piscataway, NJ, USA: IEEE; 2024. p. 1–4.
43. Lebed S, Namiot D, Zubareva E, Khenkin P, Vorobeva A, Svichkar D. Large language models in cyberattacks. In: Doklady mathematics. Vol. 110. Cham, Switzerland: Springer; 2024. p. S510–20.
44. Siemerink A, Jansen S, Labunets K. The dual-edged sword of large language models in phishing. In: Nordic Conference on Secure IT Systems. Cham, Switzerland: Springer; 2024. p. 258–79.
45. De Queiroz HJDS. Phishing and social engineering attack prevention with LLMs. In: Revolutionizing cybersecurity with deep learning and large language models. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 133–64.

46. Li Z, Zhang J, Han W. A survey of cybersecurity knowledge base and its automatic labeling. In: International Conference on Network Simulation and Evaluation. Cham, Switzerland: Springer; 2023. p. 53–70.
47. Imanifard A, Majidi B, Shamisa A. SmartGridAgent: an educational framework for reliable digital twin-based smart grid workforce training with locally hosted LLMs. Smart Grids Sustain Energy. 2025;10(2):41. doi:10.1007/s40866-025-00274-0.
48. Liu Z. A review of advancements and applications of pre-trained language models in cybersecurity. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS). Piscataway, NJ, USA: IEEE; 2024. p. 1–10.
49. Motlagh FN, Hajizadeh M, Majd M, Najafi P, Cheng F, Meinel C. Large language models in cybersecurity: state-of-the-art. arXiv:2402.00891. 2024.
50. Sarker IH. CyberAI: a comprehensive summary of AI variants, explainable and responsible AI for cybersecurity. In: AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability. Cham, Switzerland: Springer; 2024. p. 173–200.
51. Palma G, Cecchi G, Caronna M, Rizzo A. Leveraging large language models for scalable and explainable cybersecurity log analysis. J Cybersecur Priv. 2025;5(3):55. doi:10.3390/jcp5030055.
52. Khan R, Gupta N, Sinhababu A, Chakravarty R. Impact of conversational and generative AI systems on libraries: a use case large language model (LLM). Sci Technol Librar. 2024;43(4):319–33.
53. Mohsin A, Janicke H, Nepal S, Holmes D. Digital twins and the future of their use enabling shift left and shift right cybersecurity operations. In: 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). Piscataway, NJ, USA: IEEE; 2023. p. 277–86.
54. Luntovskyy A, Winkler U. Advanced software technology paradigms and AI deployment. In: Intelligent Systems Conference. Cham, Switzerland: Springer; 2024. p. 590–605.
55. Gandhi ST. RAG-driven cybersecurity intelligence: leveraging semantic search for improved threat detection. Int J Res Appl Innov. 2023;6(3):8889–97.
56. Shahid AR, Hasan SM, Kankanamge MW, Hossain MZ, Imteaj A. WatchOverGPT: a framework for real-time crime detection and response using wearable camera and large language model. In: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC). Piscataway, NJ, USA: IEEE; 2024. p. 2189–94.
57. Moongela H, Mayayise T. The impact of large language models on cybersecurity. In: Annual Conference of South African Institute of Computer Scientists and Information Technologists. Cham, Switzerland: Springer; 2026. p. 129–46.
58. Muzammal SM, Mahadevappa P, Tayyab M. Exploring security challenges in generative AI for web engineering. In: Generative AI for web engineering models. Hershey, PA, USA: IGI Global; 2025. p. 331–60. doi:10.4018/979-8-3693-3703-5.ch016.
59. Babaey V, Ravindran A. Gensqli: a generative artificial intelligence framework for automatically securing web application firewalls against structured query language injection attacks. Future Internet. 2024;17(1):8.
60. Seo J, Zhang N, Rong C. Flexible and secure code deployment in federated learning using large language models: prompt engineering to enhance malicious code detection. In: 2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). Piscataway, NJ, USA: IEEE; 2023. p. 341–9.
61. Martell MJ, Baweja JA, Dreslin BD. Mitigative strategies for recovering from large language model trust violations. J Cognit Eng Decis Mak. 2025;19(1):76–95. doi:10.1177/15553434241303577.
62. Corchado JM, López S, Garcia R, Chamoso P, Juan M, Nunez V. Generative artificial intelligence: fundamentals. ADCAIJ. 2023;12(1):e31704–4. doi:10.14201/adcaij.31704.
63. Koide T, Nakano H, Chiba D. Chatphishdetector: detecting phishing sites using large language models. IEEE Access. 2024;12:154381–400. doi:10.1109/ACCESS.2024.3483905.
64. Li L, Gong B. Prompting large language models for malicious webpage detection. In: 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML). Piscataway, NJ, USA: IEEE; 2023. p. 393–400.
65. Krishnamurthy O. Enhancing cyber security enhancement through generative AI. Int J Univ Sci Eng. 2023;9(1):35–50.

66. Chowdhury MM, Rifat N, Ahsan M, Latif S, Gomes R, Rahman MS. ChatGPT: a threat against the CIA triad of cyber security. In: 2023 IEEE International Conference on Electro Information Technology (eIT). Piscataway, NJ, USA: IEEE; 2023. p. 1–6.
67. Iyengar A, Kundu A. Large language models and computer security. In: 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). Piscataway, NJ, USA: IEEE; 2023. p. 307–13.
68. Zhang S, Li H, Sun K, Chen H, Wang Y, Li S. Security and privacy challenges of AIGC in metaverse: a comprehensive survey. ACM Comput Surv. 2025;57(10):1–37. doi:10.1145/3729419.
69. McIntosh TR, Susnjak T, Liu T, Watters P, Xu D, Liu D, et al. From google gemini to openai q* (q-star): a survey on reshaping the generative artificial intelligence (AI) research landscape. Technologies. 2025;13(2):51. doi:10.3390/technologies13020051.
70. Pahuja S, Kukreja S, Singh A. Comprehensive review of generative artificial intelligence: mechanisms, models, and applications. Procedia Comput Sci. 2025;258(8):3731–40. doi:10.1016/j.procs.2025.04.628.
71. Bengesi S, El-Sayed H, Sarker MK, Houkpati Y, Irungu J, Oladunni T. Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Access. 2024;12:69812–37. doi:10.1109/ACCESS.2024.3397775.
72. Mohammed D, MacLennan H. Secure authentication and identity management with AI. In: Revolutionizing cybersecurity with deep learning and large language models. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 271–306.
73. Zhang T, Kong F, Deng D, Tang X, Wu X, Xu C, et al. Moving target defense meets artificial intelligence-driven network: a comprehensive survey. IEEE Int Things J. 2025;12(10):13384–97. doi:10.1109/JIOT.2025.3533016.
74. Sakib MN, Islam MA, Pathak R, Arifin MM. Risks, causes, and mitigations of widespread deployments of large language models (LLMS): a survey. In: 2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). Piscataway, NJ, USA: IEEE; 2024. p. 1–7.
75. Xu D, Gondal I, Yi X, Susnjak T, Watters P, McIntosh TR. The erosion of cybersecurity zero-trust principles through generative AI: a survey on the challenges and future directions. J Cybersecur Priv. 2025;5(4):87. doi:10.3390/jcp5040087.
76. Lv P, Sun M, Wang H, Wang X, Zhang S, Chen Y, et al. Rag-WM: an efficient black-box watermarking approach for retrieval-augmented generation of large language models. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: ACM; 2025. p. 1709–23.
77. Mudgal P, Wouhaybi R. An assessment of ChatGPT on log data. In: International Conference on AI-generated Content. Cham, Switzerland: Springer; 2023. p. 148–69.
78. Hazra S. Review on social and ethical concerns of generative AI and IoT. In: Generative AI: current trends and applications. Cham, Switzerland: Springer; 2024. p. 257–85. doi:10.1007/978-981-97-8460-8_13.
79. Zeb A, Rehman FU, Bin Othayman M, Rabnawaz M. Artificial intelligence and ChatGPT are fostering knowledge sharing, ethics, academia and libraries. Int J Inform Learn Technol. 2025;42(1):67–83. doi:10.1108/ijilt-03-2024-0046.
80. Wang X, Wu YC. Balancing innovation and regulation in the age of generative artificial intelligence. J Inform Policy. 2024;14:385–416. doi:10.5325/jinfopoli.14.2024.0012.
81. Carvalko JR. Generative AI, ingenuity, and law. IEEE Trans Technol Soc. 2024;5(2):169–82. doi:10.1109/tts.2024.3413591.
82. Iyengar S, Nabavirazavi S, Hariprasad Y, P. HB, Mohan CK. The convergence of AI/ML and cybersecurity: advancing digital forensic techniques. In: Artificial intelligence in practice: theory and application for cyber security and forensics. Cham, Switzerland: Springer; 2025. p. 139–59. doi:10.1007/978-3-031-89327-8_4.
83. Sleiman JP. Generative artificial intelligence and large language models for digital banking: first outlook and perspectives. J Digital Bank. 2023;8(2):102–17. doi:10.69554/cnmi7720.
84. Lambert J, Stevens M. ChatGPT and generative AI technology: a mixed bag of concerns and new opportunities. Comput Schools. 2024;41(4):559–83. doi:10.1080/07380569.2023.2256710.

85. Tihanyi N, Bisztray T, Ferrag MA, Jain R, Cordeiro LC. How secure is AI-generated code: a large-scale comparison of large language models. Empir Softw Eng. 2025;30(2):47. doi:10.1007/s10664-024-10590-1.
86. Liu H, Xue J, Zhao S, Liu Y, Lu Z. The dual role of large language models in network security: survey and research trends. In: Proceedings of the 2025 ACM Workshop on Wireless Security and Machine Learning. New York, NY, USA: ACM; 2025. p. 20–5.
87. Atlam HF. LLMs in cyber security: bridging practice and education. Big Data Cogn Comput. 2025;9(7):184. doi:10.3390/bdcc9070184.
88. Celik A, Eltawil AM. At the dawn of generative AI era: a tutorial-cum-survey on new frontiers in 6G wireless intelligence. IEEE Open J Communicat Soc. 2024;5:2433–89. doi:10.1109/OJCOMS.2024.3362271.
89. Zheng Y, Chang CH, Huang SH, Chen PY, Picek S. An overview of trustworthy AI: advances in IP protection, privacy-preserving federated learning, security verification, and GAI safety alignment. IEEE J Emerg Select Topics Circ Syst. 2024;14(4):582–607. doi:10.1109/jetcas.2024.3477348.
90. Alwahedi F, Aldhaheri A, Ferrag MA, Battah A, Tihanyi N. Machine learning techniques for IoT security: current research and future vision with generative AI and large language models. Internet Things Cyber-Phys Syst. 2024;4:167–85. doi:10.1016/j.iotcps.2023.12.003.
91. Arokodare O, Wimmer H. Large language models for phishing and spam detection: a BERT approach. In: 1st Annual Fall National Conference on Creativity, Innovation, and Technology Conference (NCCIT); 2023 Nov 15–16; Online.
92. Şentürk MA, Bahtiyar Ş. A survey on large language models in phishing detection. In: 2025 12th IFIP International Conference on New Technologies, Mobility and Security (NTMS). Piscataway, NJ, USA: IEEE; 2025. p. 196–204.
93. Tan XW, See K, Kok S. Anticipate, simulate, reason (ASR): a comprehensive generative AI framework for combating messaging scams. arXiv:2507.17543. 2025.
94. Shamoo Y. Cybercrime investigation and fraud detection with AI. In: Digital forensics in the age of AI. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 83–114. doi:10.4018/979-8-3373-0857-9.ch004.
95. Cheng Y, Zhang W, Zhang Z, Zhang C, Wang S, Mao S. Towards federated large language models: motivations, methods, and future directions. IEEE Communicat Surv Tutor. 2025;27(4):2733–64. doi:10.1109/COMST.2024.3503680.
96. Novelo R, Silva RR, Bernardino J. A literature review of personalized large language models for email generation and automation. Future Internet. 2025;17(12):536. doi:10.3390/fi17120536.
97. Reti D, Becker N, Angeli T, Chattopadhyay A, Schneider D, Vollmer S, et al. Act as a honeytoken generator! an investigation into honeytoken generation with large language models. In: Proceedings of the 11th ACM Workshop on Adaptive and Autonomous Cyber Defense. New York, NY, USA: ACM; 2024. p. 1–12.
98. Pham E, Bui T, Chen H. LLM-based browser extension for phishing detection. In: 2025 Silicon Valley Cybersecurity Conference (SVCC). Piscataway, NJ, USA: IEEE; 2025. p. 1–3.
99. Koide T, Fukushi N, Nakano H, Chiba D. Chatspamdetector: leveraging large language models for effective phishing email detection. In: International Conference on Security and Privacy in Communication Systems. Cham, Switzerland: Springer; 2024. p. 297–319.
100. Heiding F, Lermen S, Kao A, Schneier B, Vishwanath A. Evaluating large language models' capability to launch fully automated spear phishing campaigns: validated on human subjects. arXiv:2412.00586. 2024.
101. Weinz M, Zannone N, Allodi L, Apruzzese G. The impact of emerging phishing threats: assessing quishing and LLM-generated phishing emails against organizations. In: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security. New York, NY, USA: ACM; 2025. p. 1550–66.
102. Shrestha L, Balogun H, Khan S. AI-driven phishing: techniques, threats, and defence strategies. In: International Conference on Global Security, Safety, and Sustainability. Cham, Switzerland: Springer; 2023. p. 121–43.
103. Hmimou Y, Tabaa M, Khiat A, Hidila Z. A multi-agent system for cybersecurity threat detection and correlation using large language models. IEEE Access. 2025;13(5):150199–215. doi:10.1109/access.2025.3602681.
104. Deshpande AS, Gupta S. GenAI in the cyber kill chain: a comprehensive review of risks, threat operative strategies and adaptive defense approaches. In: 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG). Piscataway, NJ, USA: IEEE; 2023. p. 1–5.

105. Rondanini C, Carminati B, Ferrari E, Kundu A, Gaudiano A. Malware detection at the edge with lightweight LLMs: a performance evaluation. ACM Trans Internet Technol. 2026;26(1):15. doi:10.1145/3769681.
106. Che X, Zheng Y, Zhu M, Li Q, Dong X. A domain-adaptive large language model with refinement framework for IoT cybersecurity. In: 2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics. Piscataway, NJ, USA: IEEE; 2024. p. 224–9.
107. Devadiga D, Jin G, Potdar B, Koo H, Han A, Shringi A, et al. Gleam: GAN and LLM for evasive adversarial malware. In: 2023 14th International Conference on Information and Communication Technology Convergence (ICTC). Piscataway, NJ, USA: IEEE; 2023. p. 53–8.
108. Sikos LF. Defensive generative AI. In: Generative AI in cybersecurity. Cham, Switzerland: Springer; 2025. p. 1–24. doi:10.1007/978-3-032-05250-6_1.
109. Anglano C. A review of mobile surveillanceware: capabilities, countermeasures, and research challenges. Electronics. 2025;14(14):2763. doi:10.3390/electronics14142763.
110. Al Maqousi A, Basu K. Cybersecurity in smart cities generative AI as threat and solution. In: Examining cybersecurity risks produced by generative AI. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 357–80. doi:10.4018/979-8-3373-0832-6.ch015.
111. Xue Y, Wang J, Yin Z, Ma Y, Qin H, Tao R, et al. Dual intention escape: penetrating and toxic jailbreak attack against large language models. In: Proceedings of the ACM on Web Conference 2025. New York, NY, USA: ACM; 2025. p. 863–71.
112. Farghaly MS, Aslan HK, Abdel Halim IT. A hybrid human-AI model for enhanced automated vulnerability scoring in modern vehicle sensor systems. Future Internet. 2025;17(8):339. doi:10.3390/fi17080339.
113. Katukam R. AI-driven log summarization for security operations centers: a web-based approach using Gemini API. Int J Emerg Res Eng Technol. 2025;6(3):136–45. doi:10.63282/3050-922X.IJERET-V6I3P117.
114. Harrath Y, Adohinzin O, Kaabi J, Saathoff M. Bridging domains: advances in explainable, automated, and privacy-preserving AI for computer science and cybersecurity. Computers. 2025;14(9):374. doi:10.3390/computers14090374.
115. Grov G, Halvorsen J, Eckhoff MW, Hansen BJ, Eian M, Mavroeidis V. On the use of neurosymbolic AI for defending against cyber attacks. In: International Conference on Neural-Symbolic Learning and Reasoning. Cham, Switzerland: Springer; 2024. p. 119–40.
116. Maity S, Arora J. The colossal defense: security challenges of large language models. In: 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON). Piscataway, NJ, USA: IEEE; 2024. p. 1–5. doi:10.1109/delcon64804.2024.10866433.
117. Pearce H, Tan B, Ahmad B, Karri R, Dolan-Gavitt B. Examining zero-shot vulnerability repair with large language models. In: 2023 IEEE Symposium on Security and Privacy (SP). Piscataway, NJ, USA: IEEE; 2023. p. 2339–56.
118. AlOmar B, Trabelsi Z. Integrating generative AI in cybersecurity curricula. In: 2025 IEEE Global Engineering Education Conference (EDUCON). Piscataway, NJ, USA: IEEE; 2025. p. 1–9.
119. Huang H, Meng T, Jia W. Joint optimization of prompt security and system performance in edge-cloud LLM systems. In: IEEE INFOCOM 2025-IEEE Conference on Computer Communications. Piscataway, NJ, USA: IEEE; 2025. p. 1–10.
120. Moorthy V, Rupen, Chopra D, Jain A. Uncovering hidden risks: a vulnerability analysis of advanced conversational AI models. In: International Conference on Smart Trends for Information Technology and Computer Communications. Cham, Switzerland: Springer; 2025. p. 377–88.
121. Sai S, Yashvardhan U, Chamola V, Sikdar B. Generative AI for cyber security: analyzing the potential of ChatGPT, DALL-E, and other models for enhancing the security space. IEEE Access. 2024;12:53497–516. doi:10.1109/ACCESS.2024.3385107.
122. Kuntur S, Wróblewska A, Paprzycki M, Ganzha M. Under the influence: a survey of large language models in fake news detection. IEEE Trans Artif Intell. 2025;6(2):458–76. doi:10.1109/TAI.2024.3471735.

123. Harris S, Hadi HJ, Ahmad N, Alshara MA. Fake news detection revisited: an extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas. Technologies. 2024;12(11):222. doi:10.3390/technologies12110222.
124. Huang H, Sun N, Tani M, Zhang Y, Jiang J, Jha S. Can LLM-generated misinformation be detected: a study on Cyber Threat Intelligence. Future Gener Comput Syst. 2025;173:107877. doi:10.1016/j.future.2025.107877.
125. Zhou W, Zhu X, Han QL, Li L, Chen X, Wen S, et al. The security of using large language models: a survey with emphasis on ChatGPT. IEEE/CAA J Automat Sinica. 2025;12(1):1–26. doi:10.1109/JAS.2024.124983.
126. Golec J, Hachaj T. Ten natural language processing tasks with generative artificial intelligence. Appl Sci. 2025;15(16):9057. doi:10.3390/app15169057.
127. Okey OD, Udo EU, Rosa RL, Rodríguez DZ, Kleinschmidt JH. Investigating ChatGPT and cybersecurity: a perspective on topic modeling and sentiment analysis. Comput Secur. 2023;135:103476. doi:10.1016/j.cose.2023.103476.
128. Xu H, Wang S, Li N, Wang K, Zhao Y, Chen K, et al. Large language models for cyber security: a systematic literature review. ACM Trans Softw Eng Methodol. 2024. doi:10.1145/3769676.
129. Yang Z, Zhang Y, Zeng J, Yang Y, Jia Y, Song H, et al. AI-driven safety and security for UAVs: from machine learning to large language models. Drones. 2025;9(6):392. doi:10.3390/drones9060392.
130. Zhai J, Li Z, Miao H, Li Z, Zhou X, Yang H. Automatic generation of cybersecurity teaching cases using large language models. Comput Appl Eng Educ. 2025;33(5):e70081. doi:10.1002/cae.70081.
131. Ibrahim N, Kashef R. Exploring the emerging role of large language models in smart grid cybersecurity: a survey of attacks, detection mechanisms, and mitigation strategies. Front Energy Res. 2025;13:1531655. doi:10.3389/fenrg.2025.1531655.
132. Sarker IH. Generative AI and large language modeling in cybersecurity. In: AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability. Cham, Switzerland: Springer; 2024. p. 79–99.
133. Che X, Gao P, Zhu M, Gu H, Li Q, Li M. Text-enhanced method for LLMs domain adaptation in cybersecurity. J Comput Inf Syst. 2025. doi:10.1080/08874417.2025.2554853.
134. Chu T, Song Z, Yang C. How to protect copyright data in optimization of large language models?. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38; Palo Alto, CA, USA: AAAI Press; 2024. p. 17871–9.
135. Kaushik A, Kaushik A. Decoding potential of ChatGPT: a comprehensive exploration of AI generated contents and challenges. In: Textual intelligence: large language models and their real-world applications. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2025. p. 177–99.
136. Usman Y, Oladipupo H, During AD, Robert A, Chataut R. AI, ML, and LLM integration in 5G/6G networks: a comprehensive survey of architectures, challenges, and future directions. IEEE Access. 2025;13:168914–50.
137. Nguyen TT, Vu HT, Nguyen HN. Security, privacy, and ethical challenges of artificial intelligence in large language model scope: a comprehensive survey. In: 2024 1st International Conference on Cryptography and Information Security (VCRIS). Piscataway, NJ, USA: IEEE; 2024. p. 1–6.
138. Sezgin A. Scenario-driven evaluation of autonomous agents: integrating large language model for UAV mission reliability. Drones. 2025;9(3):213. doi:10.3390/drones9030213.
139. Huang L, Dave D, Cody T, Beling PA, Jin M. From capabilities to performance: evaluating key functional properties of LLM architectures in penetration testing. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025 Nov 4–9; Suzhou, China. p. 15890–916.
140. Narula S, Ghasemigol M, Carnerero-Cano J, Minnich A, Lupu E, Takabi D. Exploring research and tools in AI security: a systematic mapping study. IEEE Access. 2025;13(2):84057–80. doi:10.1109/access.2025.3567195.
141. Moenks N, Penava P, Buettner R. A systematic literature review of large language model applications in industry. IEEE Access. 2025;13(4):160010–33. doi:10.1109/access.2025.3608650.
142. Mohawesh R, Ottom MA, Salameh HB. A data-driven risk assessment of cybersecurity challenges posed by generative AI. Decis Anal J. 2025;15(6):100580. doi:10.1016/j.dajour.2025.100580.
143. Ihekweazu C, Zhou B, Adelowo EA. Ethics-driven education: integrating AI responsibly for academic excellence. Inform Syst Educat J. 2024;22(3):36–46. doi:10.62273/JWXX9525.

144. Khan MU, Ullah MF, Khan SU, Kong W. Ethical principles of integrating ChatGPT into IoT-based software wearables: a fuzzy-TOPSIS ranking and analysis approach. Int J Intell Syst. 2025;2025(1):6660868. doi:10.1155/int/6660868.
145. Czaja P, Gdowski B, Niemiec M, Mees W, Stoianov N, Votis K, et al. Cybersecurity challenges and opportunities of machine learning-based artificial intelligence. Neural Comput Appl. 2025;37:27931–56. doi:10.1007/s00521-025-11604-9.
146. Xiong H, Bian J, Li Y, Li X, Du M, Wang S, et al. When search engine services meet large language models: visions and challenges. IEEE Trans Serv Comput. 2024;17(6):4558–77. doi:10.1109/tsc.2024.3451185.
147. Sun N, Miao Y, Mo X, Zhang J. Large language models for cybersecurity education: a survey of current practices and future directions. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham, Switzerland: Springer; 2025. p. 3–20.
148. Mateo Sanguino TDJ. Enhancing security in industrial application development: case study on self-generating artificial intelligence. Appl Sci. 2024;14(9):3780. doi:10.3390/app14093780.
149. Doumanas D, Karakikes A, Soularidis A, Mainas E, Kotis K. Emerging threat vectors: how malicious actors exploit LLMs to undermine border security. AI. 2025;6(9):232.
150. Truong TC, Diep QB, Zelinka I. Artificial intelligence in the cyber domain: offense and defense. Symmetry. 2020;12(3):410. doi:10.3390/sym12030410.
151. Divakaran DM, Peddinti ST. Large language models for cybersecurity: new opportunities. IEEE Secur Priv. 2024;23:38–45.
152. Kasri W, Himeur Y, Alkhazaleh HA, Tarapiah S, Atalla S, Mansoor W, et al. From vulnerability to defense: the role of large language models in enhancing cybersecurity. Computation. 2025;13(2):30. doi:10.3390/computation13020030.
153. Friha O, Ferrag MA, Kantarci B, Cakmak B, Ozgun A, Ghoualmi-Zine N. LLM-based edge intelligence: a comprehensive survey on architectures, applications, security and trustworthiness. IEEE Open J Communicat Soc. 2024;5:5799–856. doi:10.1109/OJCOMS.2024.3456549.
154. Javaid S, Fahim H, He B, Saeed N. Large language models for UAVS: current state and pathways to the future. IEEE Open J Vehic Technol. 2024;5:1166–92. doi:10.1109/OJVT.2024.3446799.
155. Huang K, Wang Y, Goertzel B, Li Y, Wright S, Ponnapalli J. Generative AI security: theories and practices (future of business and finance). Berlin, Germany: Springer; 2025.
156. Ismail, Kurnia R, Widyatama F, Wibawa IM, Brata ZA, Ukasyah, et al. Enhancing security operations center: wazuh security event response with retrieval-augmented-generation-driven copilot. Sensors. 2025;25(3):870. doi:10.3390/s25030870.
157. Uddin M, Irshad MS, Kandhro IA, Alanazi F, Ahmed F, Maaz M, et al. Generative AI revolution in cybersecurity: a comprehensive review of threat intelligence and operations. Artif Intell Rev. 2025;58(8):236. doi:10.1007/s10462-025-11219-5.
158. Karkuzhali S, Senthilkumar S. LLM-powered security solutions in healthcare, government, and industrial cybersecurity. In: Revolutionizing cybersecurity with deep learning and large language models. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 97–132.
159. Haque MA, Siddique S, Rahman MM, Hasan AR, Das LR, Kamal M, et al. SOK: exploring hallucinations and security risks in AI-assisted software development with insights for LLM deployment. In: 2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA). Piscataway, NJ, USA: IEEE; 2025. p. 57–64.
160. Rondanini C, Carminati B, Ferrari E, Kundu A, Jajoo A. Large language models to enhance malware detection in edge computing. In: 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA). Piscataway, NJ, USA: IEEE; 2024. p. 1–10.
161. Zou Q, Singhal A, Sun X, Liu P. Deep learning for detecting logic-flaw-exploiting network attacks: an end-to-end approach. J Comput Secur. 2022;30(4):541–70. doi:10.3233/jcs-210101.
162. Lai H. Intrusion detection technology based on large language models. In: 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT). Piscataway, NJ, USA: IEEE; 2023. p. 1–5.

163. Usman Y, Ihejirika CJ, Offor SN, Robert A, Chataut R. Green cybersecurity: leveraging AI, ML, and LLMs to optimize energy, threat detection, and sustainability Frameworks. IEEE Access. 2025;13(1):159345–79. doi:10.1109/access.2025.3602451.
164. Kulkarni A, Balachandran V, Divakaran DM, Das T. From ML to LLM: evaluating the robustness of phishing web page detection models against adversarial attacks. Digital Threats Res Pract. 2025;6(2):1–25. doi:10.1145/3737295.
165. Toth R, Bisztray T, Erdodi L. LLMS in web development: evaluating LLM-generated PHP code unveiling vulnerabilities and limitations. In: International Conference on Computer Safety, Reliability, and Security. Cham, Switzerland: Springer; 2024. p. 425–37.
166. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. 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. New York, NY, USA: ACM; 2023. p. 79–90.
167. Dharmendra H, Raghunandan G, Sindhu A, Samanvitha C, Nethravathi N, Elango D. Human evaluation in large language model testing: assessing the quality of AI model output. In: Advancements in intelligent process automation. Hershey, PA, USA: IGI Global; 2025. p. 553–74. doi:10.4018/979-8-3693-5380-6.ch022.
168. Zaydi M, Maleh Y. Empowering red teams with generative AI: transforming penetration testing through adaptive intelligence. EDPACS. 2025;70(2):41–66. doi:10.1080/07366981.2024.2439628.
169. Omar KO, Zraqou J, Gomez JM. From synthetic text to real threats: unraveling the security risks of generative AI. In: Examining cybersecurity risks produced by generative AI. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 1–20. doi:10.4018/979-8-3373-0832-6.ch001.
170. Li W, Manickam S, Chong YW, Leng W, Nanda P. A state-of-the-art review on phishing website detection techniques. IEEE Access. 2024;12(4):187976–8012. doi:10.1109/access.2024.3514972.
171. Alharthi D, Yasaei R. Llm-powered automated cloud forensics: from log analysis to investigation. In: 2025 IEEE 18th International Conference on Cloud Computing (CLOUD). Piscataway, NJ, USA: IEEE; 2025. p. 12–22.