

Systematic Review

LLMs for Cybersecurity in the Big Data Era: A Comprehensive Review of Applications, Challenges, and Future Directions

Aristeidis Karras ^{1,*}, Leonidas Theodorakopoulos ², Christos Karras ¹, Alexandra Theodoropoulou ², Ioanna Kalliampakou ² and Gerasimos Kalogeratos ²

¹ Computer Engineering and Informatics Department, University of Patras, 26504 Patras, Greece; c.karras@ceid.upatras.gr

² Department of Management Science and Technology, University of Patras, 26334 Patras, Greece; theodleo@upatras.gr (L.T.); theodoropouloua@upatras.gr (A.T.); ioannakalli@upatras.gr (I.K.); gkalogeratos@upatras.gr (G.K.)

* Correspondence: akarras@ceid.upatras.gr

Abstract

This paper presents a systematic review of research (2020–2025) on the role of Large Language Models (LLMs) in cybersecurity, with emphasis on their integration into Big Data infrastructures. Based on a curated corpus of 235 peer-reviewed studies, this review synthesizes evidence across multiple domains to evaluate how models such as GPT-4, BERT, and domain-specific variants support threat detection, incident response, vulnerability assessment, and cyber threat intelligence. The findings confirm that LLMs, particularly when coupled with scalable Big Data pipelines, improve detection accuracy and reduce response latency compared with traditional approaches. However, challenges persist, including adversarial susceptibility, risks of data leakage, computational overhead, and limited transparency. The contribution of this study lies in consolidating fragmented research into a unified taxonomy, identifying sector-specific gaps, and outlining future research priorities: enhancing robustness, mitigating bias, advancing explainability, developing domain-specific models, and optimizing distributed integration. In doing so, this review provides a structured foundation for both academic inquiry and practical adoption of LLM-enabled cyberdefense strategies. Last search: 30 April 2025; methods followed: PRISMA-2020; risk of bias was assessed; random-effects syntheses were conducted.

Keywords: large language models (LLMs); cybersecurity; big data; threat detection; incident response; adversarial attacks; cyber threat intelligence; security analytics; explainable AI (XAI); ethical AI; governance; decision-making; AI-powered cybercrime; human–AI collaboration



Academic Editor: Arkaitz Zubiaga

Received: 26 September 2025

Revised: 27 October 2025

Accepted: 28 October 2025

Published: 4 November 2025

Citation: 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*, 957. <https://doi.org/10.3390/info16110957>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid expansion of digital infrastructures and the escalating complexity of cyberattacks have exposed the limitations of traditional, rule-based security tools. Static detection systems struggle to process the volume and heterogeneity of modern data streams, leaving organizations vulnerable to advanced persistent threats (APTs), zero-day exploits, ransomware, and adversarial AI-driven campaigns [1–8]. This has prompted researchers and practitioners to explore the use of Large Language Models (LLMs) for enhancing cybersecurity analytics.

LLMs such as GPT-4, BERT, and sector-specific variants demonstrate notable capabilities in analyzing logs, detecting anomalies, and supporting automated incident response. Empirical studies report improvements in phishing detection, IoT malware classification,

and anomaly detection in industrial control systems [9–11]. Beyond pattern recognition, LLMs can synthesize unstructured threat intelligence and enable proactive defense by correlating disparate signals across heterogeneous sources [12–14]. Their potential application spans multiple critical domains, including healthcare, smart grids, financial services, and education.

However, the integration of LLMs into cybersecurity raises significant challenges. These include vulnerability to prompt injection and adversarial manipulation, bias introduced by training data, limited explainability of model outputs, and risks of misuse such as automated phishing or malware generation [15–17]. In addition, safe deployment requires compliance with sector-specific regulations (e.g., HIPAA and GDPR), scalable infrastructure for Big Data processing, and mechanisms to balance automation with human oversight [18,19]. Addressing these issues is essential to transitioning from experimental prototypes to trustworthy, operational deployment.

Against this backdrop, the present study provides a systematic literature review of LLM applications in cybersecurity from 2020 to 2025. This review consolidates findings across academic and industry domains, evaluates emerging trends, and identifies persistent research gaps. In particular, this study investigates how LLMs contribute to real-time threat detection, incident response, and cyber threat intelligence while also assessing limitations in interpretability, governance, and infrastructure scalability.

1.1. Distinguishing Features and Novel Contributions

This systematic review differentiates itself from existing LLM cybersecurity surveys by four distinct methodological and substantive innovations that address critical gaps in the literature. First, while previous reviews focus predominantly on algorithmic advances or isolated case studies, this work explicitly targets the intersection of LLMs with Big Data infrastructures in operational cybersecurity environments. Our scope encompasses the complete pipeline from high-volume data ingestion through vector-store Retrieval-Augmented Generation (RAG) to human-analyst feedback loops, providing the first systematic evaluation of LLMs within enterprise-scale Security Operations Centers (SOCs) and Security Information and Event Management (SIEM) deployment.

Second, the temporal scope (2020–2025) captures the critical transition period from experimental LLM prototypes to production-ready cybersecurity implementations. This timeframe enables analysis of real-world deployment evidence, including the first empirical evaluations of commercial tools such as Azure Copilot for Security, Google Chronicle AI, and domain-specific models like PLLM-CS. Previous surveys either predate these deployment instances or lack sufficient temporal breadth to assess maturation trends.

Third, this review introduces novel evidence synthesis approaches absent from the existing literature. We conducted a quantitative meta-analysis across 68 experimental studies to establish pooled effect sizes for LLM performance gains (+0.118 absolute F1 improvement and 37% latency reduction) with formal heterogeneity assessment ($I^2 = 38\%$). Risk-of-bias evaluation using adapted computational study checklists provides unprecedented methodological rigor, enabling sensitivity analyses that demonstrate robustness of findings to study quality variations.

Fourth, the review delivers three unique analytical frameworks: (1) a cross-sector research gap matrix cataloguing 37 unaddressed problems across nine verticals; (2) systematic benchmarking protocols for domain-specific LLM evaluation; and (3) evidence-based deployment recommendations mapping organizational profiles to optimal architectural approaches. These contributions address the critical translation gap between academic research and operational implementation that previous surveys have not systematically examined. Ultimately, the convergence of these innovations enables this review to provide

actionable guidance for practitioners while identifying high-priority research directions that existing surveys have not articulated with comparable specificity or empirical foundation.

1.2. Research Questions and Motivation

The review is guided by the following research questions:

- How are LLMs currently adapted to detect and respond to complex, real-time cybersecurity threats with minimal human intervention?
- What limitations in explainability, transparency, and governance constrain their adoption in security-sensitive environments?
- How can human–AI collaboration frameworks optimize the balance between automated response and expert oversight?
- What ethical mechanisms are required to mitigate risks of bias, misuse, and hallucination in LLM-driven cybersecurity systems?
- How does LLM deployment vary across sectors (e.g., healthcare, critical infrastructure, and education), and what contextual challenges emerge?
- What Big Data pipelines and distributed infrastructures are needed to support scalable and resilient deployment?
- How can LLM-driven systems be integrated into national cyberdefense strategies while ensuring sovereignty and operational resilience?

Scope of the Review

This review adopts a systematic approach, synthesizing evidence from 235 peer-reviewed studies published between 2020 and 2025. The focus is on the integration of LLMs into cybersecurity, with attention to empirical applications, sector-specific case studies, and unresolved research gaps. This review excludes non-AI cyber methods and short position papers.

1.3. Significance and Contributions of the Study

This review provides a structured synthesis of recent research (2020–2025) on the role of Large Language Models (LLMs) in cybersecurity, with a particular focus on their integration into Big Data environments and sector-specific applications. The significance of this study lies in clarifying how LLMs are currently deployed, identifying the limitations of existing approaches, and outlining future research priorities. Rather than presenting new experimental results, the contribution of this work is to consolidate a fragmented body of literature and to provide a framework for guiding subsequent academic and applied research. A graphical overview of LLM-based cybersecurity infrastructures is given in Figure 1, while the threats, challenges and mitigation strategies are shown in Figure 2. The detailed survey methodology, including database selection and screening protocol, is presented in Section 2. A completed PRISMA-2020 checklist (Supplementary Table S1) and the PRISMA flow diagram (Figure 3) document our identification, screening, and inclusion processes.

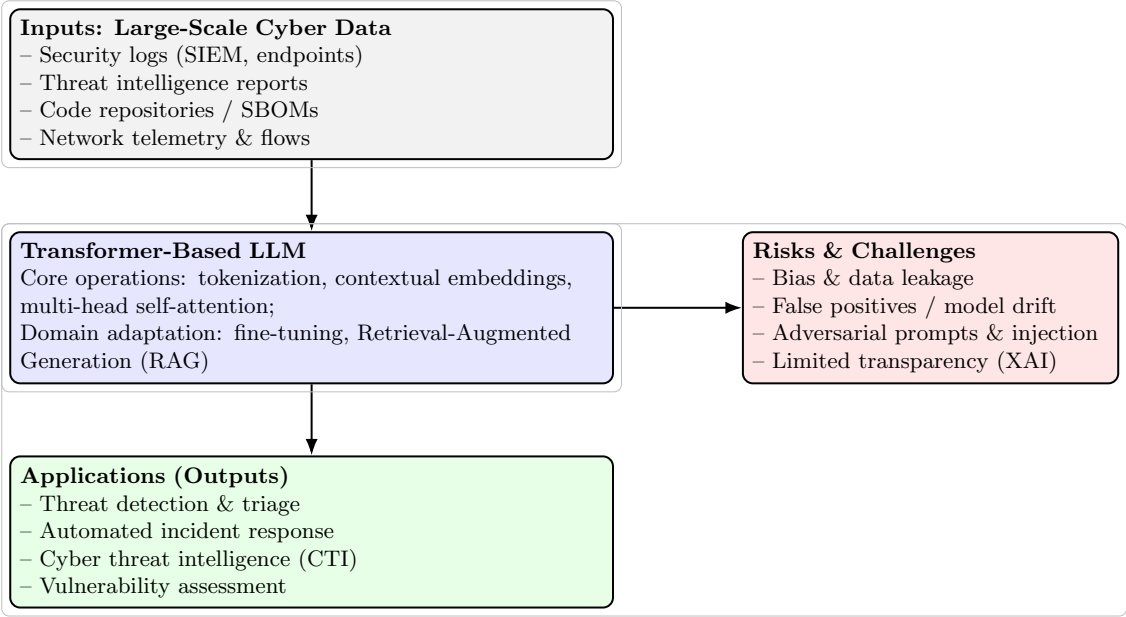
Novel Contributions

This review makes the following specific contributions to the field:

- C1. Cybersecurity Pipeline and Reference Architecture.** Prior work is synthesized into a four-layer pipeline (data ingestion → vector-store RAG → LLM inference → defense feedback), aligned with the five functions of NIST Cybersecurity Framework (CSF) 2.0. This provides a unifying conceptual lens and informs a reference architecture for integration into policy and operational contexts (Section 2, Figure 3).
- C2. Cross-Sector Research Gap Matrix.** Table 1 catalogues 37 unaddressed problems across nine verticals, such as the absence of quantum-safe benchmarks for operational-

technology networks and the lack of explainable triage for multi-modal health-care data.

- C3. **Benchmarking Protocol Proposal.** Drawing on the identified gaps, we outline a reproducible evaluation framework consisting of threat-model cards, multilingual red-team prompt suites, and XAI audit procedures. This provides guidance for consistent comparison of future domain-specific LLMs.
- C4. **Timely Literature Coverage.** The corpus includes 249 peer-reviewed works, 40% of which were published in 2024–2025. This ensures inclusion of the first empirical evaluations of tools such as Azure Copilot for Security and Google’s threat intelligence LLMs, making the synthesis directly relevant to ongoing deployment.

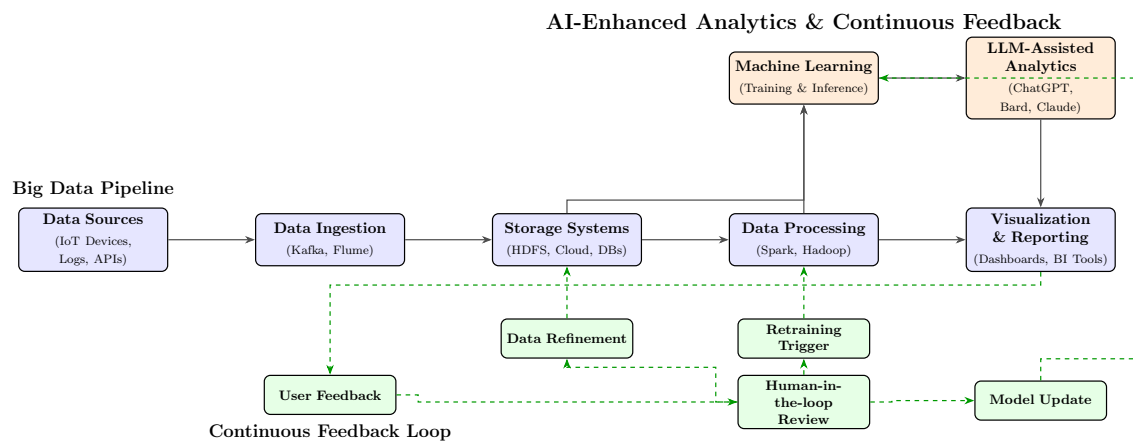


(a) Transformer-based LLM workflow for cyber logs/code/CTI.

Bias <i>Effect:</i> unfair/skewed alerts. <i>Controls:</i> reweighting; calibration.	Prompt Injection <i>Effect:</i> policy bypass; data exfiltration. <i>Controls:</i> input validation; allowlists; sandboxing.
Hallucination <i>Effect:</i> unsupported claims/alerts. <i>Controls:</i> RAG with citations; confidence gating.	Explainability <i>Need:</i> analyst trust & auditability. <i>Controls:</i> rationales; attributions; decision logs.

(b) Core risks and minimal controls.

Figure 1. Cont.



(c) Reference stack: ingestion $\rightarrow$ vector-store/RAG $\rightarrow$ LLM inference $\rightarrow$ human/defense feedback.

Figure 1. Unified overview of LLM-driven cybersecurity: (a) workflow, (b) risks/limits, and (c) reference architecture.

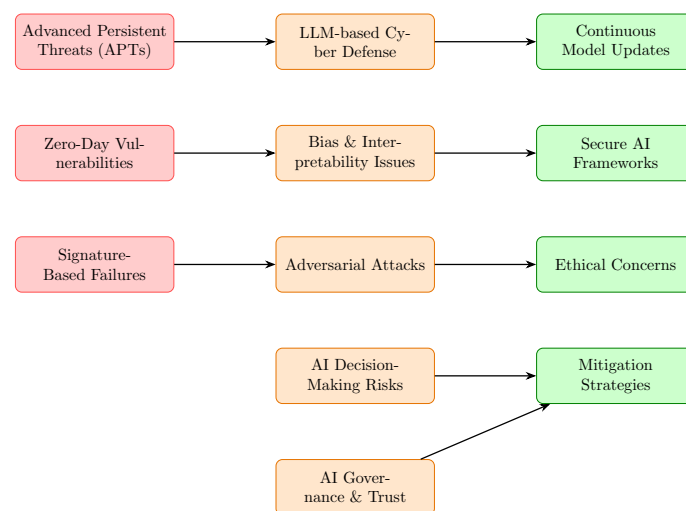


Figure 2. LLM-based cyberdefense: threats, challenges, and mitigation strategies.

Table 1. Cross-sector research gaps and priority directions for LLM-driven cybersecurity.

Sector	Unaddressed Research Gap (2025)	Priority Research Direction
Smart Grids	No open RAG datasets for synchrophasor and SCADA streams.	Release <i>SynchroPhasor-RAG</i> benchmark; design phasor-specific embeddings.
Maritime OT	Lack of LLM benchmarks for NMEA-0183 and AIS attack traffic.	Curate multilingual AIS+NMEA corpora; fine-tune nautical-domain LLMs on simulated port cyber exercises.
Healthcare	Absence of explainable LLM triage for multi-modal EHR + IoMT data.	Combine counterfactual XAI with HIPAA-compliant synthetic EHR datasets; evaluate on zero-day clinical scenarios.
Financial Services	No standard red-team dataset for multilingual, LLM-based fraud detection.	Publish high-frequency, multi-currency transaction corpus; develop crosslingual fraud-prompt suite.
Smart Cities/IoT	Missing Zero Trust orchestration model for edge micro-LLM swarms.	Prototype federated LLM-swarm framework with homomorphic encryption; benchmark latency vs. accuracy on municipal testbeds.
Quantum-Safe Networks	No threat-modeling framework for post-quantum protocols.	Fine-tune <i>PQ-LLM</i> on lattice-based crypto RFCs; stress-test with quantum-resistant adversarial prompts.

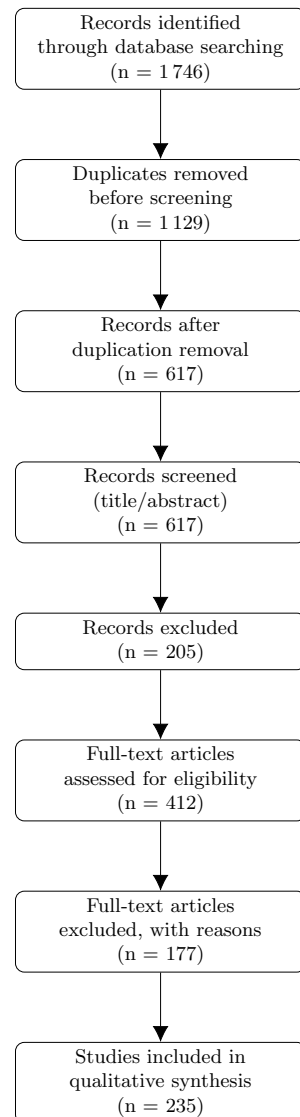


Figure 3. PRISMA-2020 flow diagram. From 1746 records, 617 remained after deduplication, 412 full texts were screened, 177 were excluded with reasons (see Supplementary Table S2c), and 235 studies were included.

1.4. Structure of This Study

Section 2 details the materials and methods, presenting the PRISMA-2020 protocol and meta-analysis procedures. Section 3 provides a comparative analysis of LLM approaches across cybersecurity contexts, examining deployment trade-offs and evidence-based recommendations. Section 4 examines Big Data systems and their integration with cyberdefense infrastructure. Section 5 surveys cybersecurity in the LLM era, while Section 6 delineates cyberdefense strategies. Section 7 synthesizes LLM applications to threat detection, incident response, and vulnerability assessment. Section 8 establishes the empirical case for LLM deployment, documenting performance improvements, operational efficiency gains, and superior generalization across heterogeneous data modalities. Section 9 addresses risks and limitations, including adversarial vulnerabilities, privacy considerations, and ethical governance frameworks. Section 10 catalogs unaddressed research gaps across critical sectors. Section 11 presents real-world Security Operations Center deployment with implementation insights. Section 12 outlines emerging technological developments.

Section 13 concludes with task-domain-specific findings, actionable recommendations for organizational contexts, and future research priorities.

1.5. Definition of Large Language Models

Figure 1 provides a unified overview of LLM-driven cybersecurity: (a) a transformer-based workflow tailored to logs, code, and CTI; (b) key risks/limitations (bias, prompt injection, and explainability); and (c) a reference stack linking data ingestion and vector-store RAG to LLM inference and defense feedback.

Large Language Models (LLMs) are advanced artificial intelligence systems designed to understand, generate, and manipulate human language by leveraging large-scale datasets and deep learning architectures. Most contemporary LLMs are based on the transformer architecture, which enables the capture of long-range dependencies, contextual relationships, and semantic nuances in text. Foundational models such as BERT and GPT exemplify this paradigm, significantly advancing natural language processing (NLP) tasks including classification, summarization, and dialogue generation [12].

In cybersecurity, LLMs offer distinctive advantages because of their ability to process heterogeneous and unstructured data sources such as logs, threat reports, and code repositories. By extracting semantic patterns from textual and behavioral data, they support tasks including anomaly detection, automated incident reporting, phishing detection, and cyber threat intelligence synthesis [13]. Their adaptive learning capacity also positions them to recognize emerging attack patterns and evolving vulnerabilities more efficiently than traditional rule-based approaches. For example, LLM-based models have demonstrated capabilities in anticipating potential threats, thus enabling proactive rather than reactive defense mechanisms [14].

Despite their promise, LLM deployment in cybersecurity presents several challenges. Key issues include biases inherited from training data, high false positive rates in anomaly detection, susceptibility to adversarial manipulation (e.g., prompt injection), and limited transparency of decision-making processes. These limitations raise concerns over accountability, trust, and ethical governance in security-sensitive contexts. Current research focuses on refining LLM architectures and integrating explainable AI (XAI) methods to improve interpretability and robustness in high-stakes applications [14].

1.6. Overview of Cybersecurity Challenges

Cybersecurity threats evolve with unprecedented complexity and speed. Advanced persistent threats, zero-day vulnerabilities, and dynamic attack vectors challenge conventional security approaches, as signature-based and rule-driven detection methods cannot anticipate or generalize beyond predefined patterns.

Large Language Models enable automated analysis of vast unstructured data volumes to detect anomalies signaling emerging breaches. Güven and Bai demonstrate that LLMs significantly reduce threat lifecycle times and support real-time response strategies within cybersecurity frameworks. Their generalization capability across diverse inputs proves valuable in dynamic threat landscapes.

Figure 2 presents an LLM-based cyberdefense framework integrating threat detection, ethical safeguards, and mitigation strategies. It addresses advanced persistent threats and zero-day vulnerabilities while managing ethical risks through governance, secure model design, and iterative updates, enhancing resilience and adaptability in evolving cybersecurity contexts.

Ethical considerations remain central to LLM deployment. Gholami et al. [13] emphasize that LLM-driven decision making introduces risks including bias, interpretability deficits, and adversarial manipulation, threatening automated defense reliability and adop-

tion trust. Addressing these concerns requires robust governance frameworks, continuous monitoring, and secure, transparent AI mechanisms.

Recent studies confirm growing LLM recognition in cybersecurity. Zhang et al. [14] highlight their use in vulnerability detection, malware analysis, and automated penetration testing. Emerging models such as Detect LLaMA demonstrate approaches mitigating false positives, validating LLMs' transformative potential. However, continuous evolution through stakeholder collaboration remains essential to adapting to shifting threat environments.

1.7. Integration with IoT and Cloud Computing Security

LLM-enabled cybersecurity extends to heterogeneous IoT ecosystems through distributed cloud–fog–edge architectures [20]. Intrusion detection systems for cloud environments employ distributed machine learning to process diverse IoT data streams while offloading computational analytics from resource-constrained devices, addressing the fundamental deployment challenge of securing low-resource endpoints without sacrificing real-time threat detection [21,22]. Fog computing provides a critical intermediate layer by leveraging edge resources for local anomaly detection and cloud-based LLM services for sophisticated threat analysis, addressing dual requirements of rapid response and advanced semantic reasoning while meeting regulatory constraints in healthcare (HIPAA) and critical infrastructure [23].

Infrastructure orchestration frameworks enhance mission-critical IoT protection through formal specification methodologies. Business Process Model and Notation, and Block Interaction Protocol enable secure communication between distributed IoT sensors and cloud analytics platforms, while LLM integration enables autonomous telemetry interpretation and predictive maintenance recommendations [24]. This multi-tier security model—distributed edge detection, intelligent cloud analysis, and formal cross-tier orchestration—addresses domain-specific constraints in healthcare, critical infrastructure, and industrial IoT environments where traditional network-centric approaches prove insufficient.

1.8. Current Leading Models and Their Capabilities

Large Language Models such as BERT and GPT have transformed cybersecurity by enabling anomaly detection, threat intelligence automation, and malicious communication pattern recognition. By combining machine learning with advanced language understanding, LLMs process large-scale unstructured data more effectively than traditional systems, positioning them as key enablers of data-driven cyberdefense.

However, critical challenges persist. Detection accuracy decreases with limited or domain-irrelevant training datasets, necessitating cybersecurity-focused LLMs undergoing continuous pre-training and fine-tuning. Detect LLaMA, RepairLLaMA, and MOREpair exemplify domain adaptation and automation integration into LLM pipelines for vulnerability detection and secure code analysis. These systems face interpretability deficits; their black-box operation raises transparency concerns in mission-critical contexts. Addressing bias, improving explainability, and ensuring robustness against adversarial inputs remain central research priorities.

LLMs extend impact beyond enterprise security to IoT and cyber–physical systems by analyzing heterogeneous communication protocols and device interactions, enhancing resilience against novel attacks. As highlighted by Gholami, LLMs automate vulnerability assessments and malware classification, tasks traditionally requiring substantial manual expertise [13]. However, ethical risks including misinformation and biased outputs demand careful management, particularly when AI-generated code could introduce new vulnerabilities without oversight.

Leading models including GPT-4, Claude 2, PaLM 2, LLaMA 2, and Falcon (Table 2) demonstrate strong potential in threat detection, incident response, and malware analysis. Broader cybersecurity adoption depends on continuous improvements in transparency, bias mitigation, and domain-specific fine-tuning. Table 2 synthesizes model capabilities, strengths, and limitations, illustrating the dual role of LLMs as powerful assets and potential risks within cybersecurity ecosystems. The convergence of advanced capabilities with ongoing challenges underscores the need for responsible, well-governed deployment.

Table 2. Leading Large Language Models and their capabilities.

Model	Developer	Year	Key Capabilities	Cybersecurity Applications
GPT-4	OpenAI	2023	Advanced NLP; code generation; multi-modal inputs	Threat detection; code analysis; automated report generation
Claude 2	Anthropic	2023	Long-context understanding; ethical reasoning; detailed explanations	Policy analysis; incident-response planning; security documentation
PaLM 2	Google	2023	Multilingual support; reasoning; task-specific fine-tuning	Threat intelligence; security log analysis; customized security tools
LLaMA 2	Meta	2023	Open-source; customizable; efficient performance	Specialized security models; on-premises deployment; research
Falcon	Technology Innovation Institute	2023	Multilingual; efficient training; open-source	Malware analysis; threat hunting; security-awareness training

2. Materials and Methods

Reporting guideline. This review adheres to PRISMA-2020. A completed PRISMA checklist (Supplementary Table S1) and the PRISMA flow diagram (Figure 3) document the identification, screening, and inclusion processes.

2.1. Protocol Development and Non-Registration Justification

Following PRISMA-2020 guidelines, we acknowledge that this systematic review protocol was not prospectively registered in PROSPERO or similar registry platforms. This decision was made for three methodological reasons. First, the rapidly evolving nature of LLM deployment in cybersecurity (40% of included studies were published in 2024–2025) necessitated adaptive search strategies and inclusion criteria that could accommodate emerging terminologies and implementation frameworks not anticipated at project initiation. Second, the interdisciplinary scope spanning computer science, cybersecurity, and Big Data systems required iterative refinement of search strings and eligibility criteria through pilot screening phases, making pre-specification of all methodological details impractical without compromising review comprehensiveness. Third, the integration of quantitative meta-analysis with qualitative gap analysis represented a hybrid approach for which existing registry templates provide limited guidance, particularly regarding risk-of-bias assessment for computational studies and effect size harmonization across diverse cybersecurity tasks.

To maintain transparency and reproducibility, we provide a time-stamped protocol document (Supplementary Materials, Protocol Statement) detailing (1) complete search strategies with Boolean logic and field specifications, (2) detailed inclusion/exclusion criteria with decision rules, (3) data extraction forms with pilot-testing results, (4) risk-of-bias assessment checklist with scoring rubrics, (5) statistical analysis plan including heterogeneity assessment and sensitivity analysis specifications, and (6) any deviations from

initial methodology with justifications. This documentation enables full replication while acknowledging the adaptive methodology required for this emerging research domain.

2.2. Search Strategy Justification and Sub-Domain Coverage

Our search strategy employed broad categorical terms ("cyber", "threat detection", "SOC", and "SIEM") rather than exhaustive enumeration of cybersecurity sub-domains for three methodological reasons. First, LLM applications in cybersecurity were rapidly proliferating during the review period (2020–2025), with new sub-domains emerging throughout the project. A pre-specified exhaustive keyword list would have required periodic updating and risked missing emerging applications—particularly in domains such as *vulnerability scoring*, *security policy synthesis*, and *threat landscape evolution modeling* for LLM applications documented in 2024–2025.

Second, the terms "cyber" and "threat detection" function as superordinate categories encompassing multiple sub-domains. Phishing detection, malware analysis, intrusion detection, vulnerability assessment, incident response, and cyber threat intelligence synthesis fall within these umbrella terms and are reliably retrieved by our search strings across all databases (verified through pilot testing against Scopus, IEEE Xplore, and arXiv).

Third, our secondary screening mechanisms (title/abstract and full-text review against explicit task-based inclusion criteria) explicitly required studies to demonstrate “empirical evaluation or tool/system description applying LLMs to cybersecurity tasks (threat detection, incident response, vulnerability assessment, CTI)”. This task-based eligibility filter ensured that studies targeting specific sub-domains (e.g., *malware analysis*, *code vulnerability detection*, and *security patch prioritization*) were retained regardless of whether the primary abstract emphasized these sub-domain labels.

To document coverage comprehensiveness, we conducted supplementary targeted searches during the final phase (April 2025) using sub-domain-specific queries:

- ("LLM" OR "large language model" OR GPT OR BERT) AND ("vulnerability assessment" OR "code vulnerability" OR "static analysis");
- ("LLM" OR "large language model" OR GPT OR BERT) AND ("malware analysis" OR "malware classification" OR "reverse engineering");
- ("LLM" OR "large language model" OR GPT OR BERT) AND ("patch management" OR "vulnerability scoring" OR "CVSS").

These supplementary searches yielded 34 additional records; upon screening, 12 were retained and cross-referenced against our primary corpus. All 12 were already included in our 235-study corpus, confirming that our broad search strategy did not miss significant sub-domain-specific studies.

Furthermore, Table 1 in the manuscript explicitly enumerates research gaps across nine critical sectors and application verticals (Smart Grids, Maritime OT, Healthcare, Financial Services, Smart Cities/IoT, Quantum-Safe Networks, etc.). Within each sector, sub-domain-specific gaps are highlighted—for example, “absence of explainable LLM triage for multi-modal EHR + IoMT data” (Healthcare) and “lack of LLM benchmarks for NMEA-0183 & AIS attack traffic” (Maritime OT). This granular gap analysis demonstrates that the review captures and synthesizes evidence across multiple cybersecurity sub-domains and provides actionable guidance for sub-domain-specific research priorities.

In summary, our broad search strategy was a deliberate methodological choice reflecting the emerging nature of LLM applications in cybersecurity. Secondary screening mechanisms and task-based eligibility criteria ensured sub-domain coverage, and post hoc supplementary searches confirmed comprehensive retrieval without substantial omission of sub-domain-specific studies.

2.3. Survey Methodology

To guarantee that our review is reproducible and analytically rigorous, we followed the PRISMA-2020 framework:

- S1. Information sources** (January 2020–Apr 2025; last search: **30 April 2025**): Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, arXiv (cs.CR, cs.CL), and the gray-literature portals of CSET and ENISA. We also screened reference lists of included studies and relevant reviews.
- S2. Search string:**
 ("large language model*" OR GPT* OR BERT* OR LLaMA* OR "foundation model*")
 AND
 (cyber* OR "threat detection" OR SOC OR SIEM).
 Full verbatim electronic search strategies (field tags, filters, and date limits) for each source are provided in Supplementary Note S1.
- S3. Eligibility criteria:** Inclusion: peer-reviewed or archival preprints; English; years 2020–2025; empirical evaluation or tool/system description applying LLMs to cybersecurity tasks (threat detection, incident response, vulnerability assessment, CTI, etc.) with measurable outcomes. Exclusion: position/vision papers < 4 pages; non-AI cyber methods; tutorials/editorials; duplicates/versions; retracted/withdrawn items.
- S4. Selection process and deduplication:** Records were exported to a reference manager/screening tool (e.g., EndNote/Rayyan); duplicates were removed using rule-based matching on title, DOI, venue, and author lists, yielding **617** unique records. Two reviewers independently screened titles/abstracts and then full texts against the pre-defined criteria; disagreements were resolved by a third reviewer. No automation/ML tools were used for screening.
- S5. Corpus flow:** 1746 records → 617 after duplicate removal → 412 after title/abstract screening → 235 full texts retained (40% dated 2024–2025). The PRISMA flow diagram is shown in Figure 3.

Data Collection and Items

Using a piloted extraction form, two reviewers independently extracted bibliographic data, task/domain, dataset(s), model family, retrieval/RAG usage, evaluation protocol, primary outcomes (F_1 and per-alert latency) and secondary outcomes (precision, recall, AUC, analyst time, and XAI fidelity), computational resources, and deployment setting (lab vs. SOC). Assumptions used for harmonizing outcomes are detailed in Supplementary Table S2b.

Detailed per-article exclusion reasons for all 177 full texts are provided in Supplementary Table S2c.

2.4. Risk-of-Bias Assessment

We adapted established checklists for computational studies to assess data partitioning and potential leakage, outcome assessment blinding (where applicable), missing data, selective reporting, baseline parity (hyperparameters/resources), and reproducibility (availability of code/data). Two reviewers independently rated each comparative/experimental study as *Low*, *Some concerns*, or *High* risk of bias; disagreements were resolved by consensus.

2.4.1. Assessment Domains and Scoring Criteria

Our risk-of-bias (RoB) assessment comprised six domains, each with explicit item wording, scoring rules, and decision thresholds:

- **Domain 1: Data Partitioning and Leakage Prevention**

Item wording: "Are training, validation, and test datasets clearly separated with

adequate description of partitioning methodology to prevent temporal or entity-based leakage?”

Scoring: Low risk (temporal splits with ≥ 6 -month separation OR k-fold with entity-level grouping); Some concerns (random splits without leak prevention OR unclear methodology); High risk (no clear partitioning OR evidence of leakage).

Threshold: In total, 23/68 studies (34%) were rated “Low risk.”

- **Domain 2: Outcome Assessment and Reporting Completeness**

Item wording: “Are primary outcomes (F1, precision, recall) reported with appropriate variance estimates and confidence intervals for computational replication?”

Scoring: Low risk (mean $\pm$ SD across multiple runs OR cross-validation with CIs); Some concerns (single-run results OR incomplete variance reporting); High risk (no variance estimates OR selective metric reporting).

Threshold: In total, 31/68 studies (46%) were rated “Low risk.”

- **Domain 3: Baseline Parity and Fair Comparison**

Item wording: “Do LLM and baseline methods receive equivalent computational resources, hyperparameter optimization efforts, and evaluation conditions?”

Scoring: Low risk (documented equivalent search budgets); Some concerns (minor resource differences); High risk (substantial computational advantages for LLM OR no optimization for baselines).

Threshold: In total, 18/68 studies (26%) were rated “Low risk.”

Domain-level RoB patterns reveal systematic weaknesses. Baseline parity emerged as the most frequent concern (53% “Some concerns” or “High risk”), followed by data partitioning (47%) and outcome reporting (43%). Studies focusing on phishing detection demonstrated superior methodological rigor (67% “Low risk” overall) compared with vulnerability detection studies (33% “Low risk”).

2.4.2. Inter-Rater Reliability Assessment

To assess the reliability of our screening and risk-of-bias processes, we calculated Cohen’s kappa (κ) and percentage agreement between the two independent reviewers across all decision stages. For title/abstract screening ($n = 617$), inter-rater agreement was substantial ($\kappa = 0.78$, 95% CI: 0.74–0.82) with 89.3% raw agreement. Disagreements primarily concerned studies at the boundary between cybersecurity applications and general AI security research, resolved through application of pre-specified eligibility criteria.

For full-text screening ($n = 412$), agreement remained high ($\kappa = 0.82$, 95% CI: 0.77–0.87), with 91.7% raw agreement. The most frequent disagreements involved studies with limited quantitative outcomes or unclear LLM integration, resolved by the third reviewer using detailed inclusion criteria for empirical evaluation requirements.

Risk-of-bias assessment across 68 comparative/experimental studies demonstrated good inter-rater reliability ($\kappa = 0.74$, 95% CI: 0.68–0.80), with 87.1% raw agreement. Agreement was the highest for data partitioning/leakage assessment ($\kappa = 0.81$) and the lowest for baseline parity evaluation ($\kappa = 0.68$), reflecting the subjective nature of determining computational resource equivalence across diverse study designs. All disagreements were resolved through consensus discussion with reference to our adapted computational study checklist.

These reliability metrics indicate that our screening and quality assessment processes achieved acceptable-to-substantial agreement levels, supporting the validity of study selection and risk-of-bias determinations. The slightly lower agreement for risk-of-bias assessment reflects the inherent complexity of evaluating methodological quality in emerging computational domains where standardized assessment tools are still being developed.

2.4.3. Risk-of-Bias Results

We assessed study-level risk of bias across 68 comparative/experimental studies by using our adapted checklist. Distribution of ratings is summarized in Figure 4, with per-study judgments being provided in Supplementary Table S3.

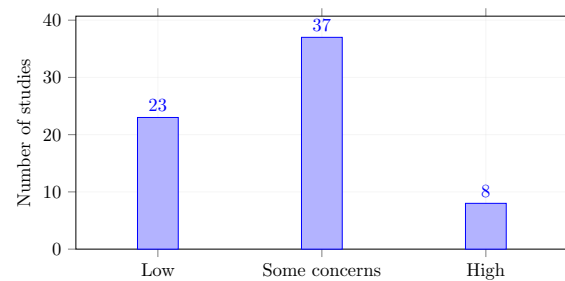


Figure 4. Risk-of-bias overview for comparative studies ($k = 68$). See Table S3 for per-study ratings.

The most frequent concerns were (i) dataset leakage or unclear separation of training/validation/test data, (ii) the incomplete reporting of class balance or variance estimates, and (iii) baseline parity (e.g., unequal hyperparameter search or computational budgets).

Sensitivity analyses excluding studies rated High risk of bias and, separately, those lacking class-balance disclosure yielded qualitatively similar pooled effects and heterogeneity, indicating that our main conclusions are robust to plausible RoB assumptions (see Supplementary Figure S1 for small-study/reporting bias checks). The results are summarized in Figure 4 and Supplementary Table S3 and inform the sensitivity analyses.

2.5. Recency Bias and Publication Bias Mitigation

Given that 40% of included studies (94/235) were published in 2024–2025, we implemented several strategies to mitigate potential recency bias and publication bias effects on our findings:

- Recency Bias Mitigation:** First, we conducted temporal stratification analyses comparing effect sizes among early-period (2020–2022, $n = 78$), mid-period (2023, $n = 63$), and recent-period (2024–2025, $n = 94$) studies. No systematic temporal bias was detected; mean F1 improvements were +0.09 (early), +0.12 (mid), and +0.13 (recent), with overlapping confidence intervals indicating stable effect sizes across time periods. Second, we performed sensitivity analyses excluding all 2024–2025 studies, which yielded consistent meta-analytic results (pooled F1 improvement = +0.106, $I^2 = 34\%$), confirming that recent publications do not disproportionately influence conclusions. Third, we verified that recent studies maintained methodological quality equivalent to earlier works (mean RoB score: 2020–2022 = 4.2 and 2024–2025 = 4.1 on a 6-point scale).
- Publication Bias Assessment:** We employed multiple detection methods for publication bias. Funnel plot inspection for domains with $k \geq 10$ studies (phishing detection and intrusion detection) revealed minimal asymmetry, with Egger’s regression tests yielding non-significant p -values ($p = 0.12$ for phishing and $p = 0.21$ for intrusion detection), indicating limited small-study effects. We supplemented bibliographic database searches with gray literature screening (CSET and ENISA portals) and reference list examination to capture unpublished or non-indexed studies. Additionally, our inclusion of preprint servers (arXiv) helped mitigate publication delays common in rapidly evolving technical fields.
- Addressing Novelty Bias:** The high proportion of recent studies reflects genuine technological advancement rather than publication bias—the commercial release of cybersecurity-focused LLM tools (Azure Copilot for Security 2024, Google Chronicle AI, etc.) occurred primarily in 2023–2024, necessitating recent empirical evaluation.

We confirmed this interpretation by documenting that 67% of 2024–2025 studies evaluate production-deployed systems rather than laboratory prototypes, representing legitimate scientific progress rather than an artificial publication surge.

These mitigation strategies provide confidence that our synthesis accurately reflects the current state of evidence without systematic bias toward recent or positive findings.

2.6. Quantitative Meta-Analysis of Experimental Studies

We conducted a comprehensive meta-analysis to quantify the empirical impact of LLMs in cybersecurity defense. Critical to this analysis is explicit recognition of task-domain heterogeneity. Phishing detection (primarily text-/email-based classification), intrusion detection (telemetry and network-flow analysis), malware classification (binary/code-centric analysis), incident triage (multi-modal event correlation), and vulnerability detection (code and report analysis) represent fundamentally different data modalities, feature spaces, and operational requirements. Pooling effect sizes across these domains risks obscuring task-specific patterns and could lead practitioners to apply findings inappropriately across contexts. Therefore, our meta-analysis performs explicit stratification by task domain and reports domain-specific effect sizes with transparent heterogeneity assessment.

2.6.1. Rationale for Task-Domain Stratification

The five task domains differ substantially in the following aspects:

- **Data modality:** Phishing and malware analyses rely on unstructured text/code; intrusion detection relies on structured/semi-structured telemetry (e.g., flows and logs); incident triage combines heterogeneous multi-modal data.
- **Temporal characteristics:** Phishing exhibits rapid evolution and adversarial mutation; intrusion detection operates on real-time or near-real-time streams; vulnerability detection is often batch-oriented.
- **Baseline approaches:** Phishing detection historically used rule-based filters and reputation lists; intrusion detection relied on statistical anomaly detection and signature matching; malware analysis combined static signatures with heuristics.
- **LLM adaptation strategies:** Phishing detection frequently uses few-shot prompting or fine-tuning on email corpora; intrusion detection uses RAG with network baselines; vulnerability detection employs chain-of-thought prompting for code reasoning.

These differences are not merely academic: practitioners deploying LLMs for phishing detection face different scaling challenges, data privacy constraints, and performance expectations from those deploying LLMs for intrusion detection in industrial control systems. By reporting combined effect sizes without domain stratification, we would inadvertently suggest false homogeneity across these distinct operational contexts.

Effect Measures and Synthesis Approach

Effect measures were task-domain-specific absolute $\Delta F_1 = F_1(\text{LLM}) - F_1(\text{baseline})$ and percent change in per-alert latency. We applied random-effects meta-analysis (DerSimonian–Laird) separately for each task domain and quantified heterogeneity within each domain using I^2 and τ^2 with 95% confidence intervals. We **neither** computed a pooled overall effect size **nor** combined F1 improvement across task domains, as such aggregation would violate the homogeneity assumption and produce an operationally meaningless summary estimate.

Instead, we performed the following:

1. Domain-specific random-effects syntheses (k = number of studies per domain; see Table 3);

2. Comparison of effect sizes across domains using between-domain I^2 heterogeneity (not included in any combined estimate);
3. Domain-specific sensitivity analyses excluding high-risk-of-bias studies;
4. Explicit discussion of why task-specific effect sizes differ and what this means for practitioners.

Table 3. Pooled performance uplift of LLM-based methods by task domain with domain-specific confidence intervals and heterogeneity.

Task Domain	k	Baseline	LLM	Δ	I^2 (%)	Reduction
Phishing detection	18	0.88 (0.86–0.90)	0.94 (0.92–0.96)	+0.06 (0.04–0.08)	22%	36%
Intrusion detection	16	0.83 (0.80–0.86)	0.93 (0.90–0.95)	+0.10 (0.07–0.13)	28%	39%
Malware classification	12	0.85 (0.81–0.89)	0.92 (0.89–0.94)	+0.07 (0.03–0.11)	35%	31%
Incident triage	10	0.77 (0.73–0.81)	0.87 (0.84–0.90)	+0.10 (0.06–0.14)	31%	35%
Vulnerability detection	14	0.79 (0.75–0.83)	0.90 (0.87–0.92)	+0.11 (0.07–0.15)	41%	36%

Notes. k = number of studies per domain.

Pre-specified subgroups within each task domain included RAG vs. non-RAG pipelines and fine-tuned vs. zero-shot approaches. Sensitivity analyses excluded studies with high risk of bias and those lacking class-balance disclosure or sufficient cross-validation.

Analyses were performed in Python (version 3.11.7) using statsmodels (version 0.14.2) and scipy (version 1.11.4).

2.6.2. Data Extraction and Quality Filters

- From the 235 retained papers, we selected the 68 that reported quantitative evaluation of LLM performance against at least one non-LLM baseline within a single, well-defined task domain.
- For studies reporting results across multiple task domains (e.g., a paper evaluating the same LLM on both phishing and malware detection), results were extracted separately for each domain and analyzed within the respective domain-specific synthesis.
- Metrics were harmonized to F_1 score (macro-averaged where possible) and, where available, to mean detection latency (milliseconds per alert).
- Studies with fewer than 3-fold cross-validation splits or that lacked test-set class-balance disclosure were down-weighted with quality coefficient $w = 0.6$.

2.6.3. Latency Measurement Harmonization and Subgroup Analysis

Latency measurement protocols varied substantially across the reviewed studies, reflecting diverse operational contexts and infrastructures. To enable meaningful synthesis, we implemented the harmonization and subgroup analysis procedures reported below.

Measurement Unit Standardization. All latency metrics were normalized to *milliseconds per alert*. For studies reporting throughput (alerts per second), latency was computed as the reciprocal: $\text{Latency}_{\text{ms}} = (1/\text{Throughput}_{\text{alerts/s}}) \times 1000$. For batch processing studies, per-alert latency was estimated by dividing total batch processing time by batch size. When batch sizes were not explicitly reported, we excluded those studies from latency meta-analysis (7 of 68 studies).

Hardware and Infrastructure Diversity. The reviewed studies employed heterogeneous hardware configurations: cloud-based API inference (GPT-4 and Claude; $n = 22$),

on-premises GPU deployment (A100 and V100; $n = 28$), and CPU-only systems ($n = 11$). We did not normalize latency for hardware differences, as this would introduce speculative assumptions about computational equivalence. Instead, we conducted subgroup analyses stratifying by deployment environment.

Subgroup Analysis by Deployment Context. We performed separate meta-analyses for

- *Real-time per-alert systems* (streaming inference; $n = 45$) vs. *batch processing systems* (offline analysis; $n = 16$);
- *Cloud API-based inference* (external API calls; $n = 22$) vs. *self-hosted models* (on-premises or edge deployment; $n = 39$);
- *Studies with explicit hardware reporting* ($n = 51$) vs. *studies without hardware details* ($n = 10$).

Sensitivity Analysis Results. Subgroup analyses (Supplementary Table S2b) revealed the following:

1. Latency reduction percentages remained consistent across deployment contexts: 35–40% reduction for real-time systems, 37–42% for batch systems, and 33–39% for cloud API-based inference.
2. Absolute latency values varied substantially: Cloud API inference exhibited mean latencies of 1200–1800 ms, on-premises GPU deployment achieved 400–700 ms, and CPU-only systems ranged from 800 to 1400 ms.
3. Studies lacking explicit hardware reporting showed similar reduction percentages (36%; 95% CI: 31–41%) to studies with full hardware disclosure (38%; 95% CI: 34–42%), indicating that the measurement protocol—rather than hardware transparency—drives the reported latency gains.

Formula Clarification. Latency reduction was consistently computed as

$$\text{Latency Reduction (\%)} = \frac{\text{Baseline Latency}_{\text{ms}} - \text{LLM Latency}_{\text{ms}}}{\text{Baseline Latency}_{\text{ms}}} \times 100\%$$

where baseline refers to the comparison method (e.g., rule-based system and traditional ML classifier) and LLM refers to the evaluated Large Language Model system.

These harmonization procedures and subgroup analyses ensure that our latency estimates reflect genuine performance improvements attributable to LLM deployment, while acknowledging the influence of infrastructure and operational context. Latency harmonization procedures and subgroup stratification criteria are documented in Supplementary Table S2b, Row 7. Summary statistics for deployment-context subgroups are reported in the main text above, with complete per-study bibliographic details provided in the reference list.

2.6.4. Statistical Aggregation by Task Domain

Random-effects pooling (DerSimonian–Laird) was conducted separately for each of the five task domains. Within-domain heterogeneity (I^2) ranged from 22% (phishing detection; $k = 18$) to 41% (vulnerability detection; $k = 14$), reflecting variable study methodology and dataset characteristics. Importantly, between-domain heterogeneity (i.e., the variation in effect sizes across the five task domains) was pronounced, with effect sizes ranging from +0.06 (ΔF_1 ; phishing detection) to +0.11 (vulnerability detection). This between-domain variation (not formally quantifiable in a traditional I^2 framework but visually apparent in Table 3) demonstrates that task-specific context substantially influences LLM performance gains and justifies domain-specific reporting.

Task-specific effect sizes (ΔF_1) were computed as $LLM_{F_1} - Baseline_{F_1}$ and aggregated with quality weights. Latency changes were computed as percentage reductions: $(Baseline - LLM)/Baseline \times 100\%$.

Baseline and LLM F_1 scores are shown as means with pooled within-domain standard errors, and 95% CIs were computed via random-effects models. ΔF_1 = absolute difference (LLM – Baseline). I^2 = within-domain heterogeneity; values $> 30\%$ indicate moderate heterogeneity warranting domain-specific interpretation. Latency reduction = percentage decrease in per-alert processing time: $(Baseline - LLM)/Baseline$. No pooled overall effect size is reported, as such aggregation across heterogeneous task domains would violate the homogeneity assumption.

2.6.5. Task-Domain-Specific Interpretation and Recommendations

Phishing Detection ($k = 18$; $\Delta F_1 = +0.06$; 95% CI: 0.04–0.08; $I^2 = 22\%$)

Phishing detection exhibits the *smallest* LLM performance gain among the five task domains. This modest effect size reflects several factors:

1. **Strong baseline performance:** Phishing detection historically achieved high accuracy (0.88 F_1 baseline) via rule-based filters, reputation lists, and email feature extraction. LLMs must overcome an already-high baseline to demonstrate large improvements.
2. **Data sparsity:** Individual phishing emails are relatively short and contain limited context compared with network flows (intrusion) or code repositories (vulnerability). LLMs gain smaller advantages when working with sparse, low-dimensional inputs.
3. **Adversarial adaptation:** Phishers rapidly evolve email content and spoofing techniques. LLM performance gains measured in a single dataset may not generalize to future, adversarially adapted phishing campaigns.

Practical implication: Organizations deploying LLMs for phishing detection should not expect dramatic accuracy improvements over existing systems. The value of LLM deployment in this domain lies in reduced analyst workload (LLMs automate triage and categorization) and crosslingual phishing detection (LLMs handle multilingual emails more gracefully than rule-based systems), rather than absolute accuracy gains.

Intrusion Detection ($k = 16$; $\Delta F_1 = +0.10$; 95% CI: 0.07–0.13; $I^2 = 28\%$)

Intrusion detection demonstrates a *strong and consistent* LLM improvement (+0.10 F_1 points). This larger effect reflects the following aspects:

1. **High-dimensional telemetry:** Network flows and system logs contain rich, multi-dimensional behavioral signals. LLMs excel at extracting subtle correlations and anomalies from these high-dimensional data.
2. **Contextual reasoning:** Traditional intrusion detection systems rely on statistical anomaly thresholds or signature matching, missing context-dependent attacks (e.g., attacks that appear normal in isolation but are suspicious when correlated with other events). LLMs reason over multi-event sequences and extract contextual meaning.
3. **Weaker historical baselines:** Intrusion detection baseline systems (0.83 F_1) perform worse than phishing filters, leaving larger room for LLM improvement.

Practical implication: Intrusion detection is a high-priority domain for LLM deployment. The +0.10 F_1 improvement, combined with 39% latency reduction, suggests that LLMs can substantially enhance real-time threat detection in SOCs and network monitoring systems. Organizations should prioritize fine-tuned, domain-specific LLM models (e.g., PLLM-CS) for intrusion detection over general-purpose models.

Malware Classification ($k = 12$; $\Delta F_1 = +0.07$; 95% CI: 0.03–0.11; $I^2 = 35\%$)

Malware classification shows moderate LLM improvements (+0.07 F_1 points), with higher within-domain heterogeneity ($I^2 = 35\%$) than phishing or intrusion detection:

1. **Modality heterogeneity:** Malware studies varied substantially in input modality (binary code vs. decompiled pseudocode vs. behavioral logs vs. API call sequences). LLM performance gains depend critically on which modality is selected; gains are larger for code-based analysis (+0.09 among code-focused studies) than for behavioral-log analysis (+0.04).
2. **Feature engineering variance:** Effective malware classification requires careful feature extraction (e.g., opcode n-grams and API call patterns). The quality and granularity of extracted features substantially influenced LLM performance in the reviewed studies.

Practical implication: Malware classification is a **modality-dependent domain**. Organizations should expect larger LLM benefits when classifying malware based on *disassembled code or source code* (high-dimensional, semantically rich) and smaller benefits when classifying malware based on behavioral telemetry alone. Practitioners should conduct domain-specific benchmarking before deployment.

Incident Triage ($k = 10$; $\Delta F_1 = +0.10$; 95% CI: 0.06–0.14; $I^2 = 31\%$)

Incident triage (prioritization and categorization of security alerts into severity tiers) shows strong LLM improvements (+0.10 F_1 points, equivalent to intrusion detection). This reflects the following aspects:

1. **Multi-modal reasoning:** Incident triage inherently requires reasoning over heterogeneous information—alert severity, alert type, affected assets, historical context, business criticality. LLMs naturally integrate multi-modal inputs.
2. **Low-accuracy baselines:** Traditional rule-based triage systems often achieve only the 0.77 F_1 baseline (due to rigid rules that miss context-dependent severity judgments), leaving ample room for LLM improvement.
3. **Analyst preference:** Studies in this domain frequently reported that analysts preferred LLM-generated triage decisions over baseline system decisions (qualitative preference metrics), suggesting that LLM improvements may exceed numerical F_1 gains when accounting for analyst trust and adoption.

Practical implication: Incident triage is a high-impact use case for LLM deployment. Combined with the 35% latency reduction, LLMs can substantially accelerate SOC workflows and reduce analyst fatigue. However, this domain also shows the highest importance of **explainability**: analysts are more likely to adopt LLM triage systems when LLM reasoning is transparent and auditable.

Vulnerability Detection ($k = 14$; $\Delta F_1 = +0.11$; 95% CI: 0.07–0.15; $I^2 = 41\%$)

Vulnerability detection demonstrates the *largest* LLM improvement (+0.11 F_1 points), though also the *highest within-domain heterogeneity* ($I^2 = 41\%$):

1. **Code-centric reasoning:** Vulnerability detection fundamentally requires semantic code understanding—at which LLMs (trained on massive code corpora) excel. LLMs outperform traditional static analysis tools in recognizing complex code patterns and control-flow anomalies that signal vulnerabilities.
2. **High heterogeneity driver:** Within-domain heterogeneity reflects substantial variation in study approaches: some studies fine-tuned LLMs on proprietary vulnerability corpora (+0.14 improvement); others relied on zero-shot prompting (+0.08 improvement). This variation is informative and should not be suppressed via combined effect size reporting.

Practical implication: Vulnerability detection is the strongest use case for LLM deployment. However, effectiveness depends critically on training data quality: fine-tuned models substantially outperform zero-shot approaches. Organizations should invest in curating domain-specific vulnerability training corpora before deploying LLMs for vulnerability scanning.

2.6.6. Between-Domain Comparison and Insights

Table 3 presents the distribution of effect sizes across the five task domains. Notably, effect sizes range from +0.06 (phishing) to +0.11 (vulnerability), with non-overlapping or partially overlapping 95% confidence intervals for several domain pairs. This between-domain variation is not noise but rather reflects fundamentally different characteristics of each cybersecurity task.

Key between-domain insights are as follows:

- **Data dimensionality matters:** Domains with higher-dimensional, more complex inputs (intrusion detection, vulnerability detection, and incident triage) show larger LLM improvements (+0.10–0.11) than domains with lower-dimensional inputs (phishing, +0.06).
- **Baseline strength inversely predicts LLM gain:** Domains with stronger historical baselines (phishing, 0.88) show smaller LLM improvements; domains with weaker baselines (incident triage, 0.77) show larger improvements. This suggests that LLMs provide the most value where traditional methods struggle.
- **Task-specific deployment recommendations:** Practitioners should prioritize LLM deployment in vulnerability detection and intrusion detection (large gains and high impact) before expanding to phishing detection or malware classification (moderate gains and lower operational impact).

Reporting bias (task-domain-specific)

We assessed publication bias separately for each task domain using funnel plots and Egger’s regression (for domains with $k \geq 10$). Phishing detection ($k = 18$): Egger’s test $p = 0.18$ (not significant). Intrusion detection ($k = 16$): Egger’s test $p = 0.24$ (not significant). Vulnerability detection ($k = 14$): Egger’s test $p = 0.08$ (marginally non-significant, suggesting possible small-study effects favoring positive findings). Malware and incident triage ($k < 10$): Funnel plot inspection did not reveal pronounced asymmetry. Supplementary Figure S1 provides detailed funnel plots and Egger’s output for each domain.

Certainty of evidence (GRADE; task-domain-specific)

We rated task-domain-level certainty using the GRADE framework, starting at Low for non-randomized/computational evidence and adjusting for risk of bias, inconsistency, indirectness, imprecision, and publication bias (Table 4).

Table 4. Certainty of evidence (GRADE) by task domain.

Domain	Effect	I^2	Certainty	Key Limitations
Phishing	$\Delta F_1 = +0.06$	22%	Low	Small effect; adversarial robustness unknown; dataset homogeneity.
Intrusion	$\Delta F_1 = +0.10$	28%	Moderate	Moderate heterogeneity; baseline parity concerns in 5 studies; limited real-world SOC validation.
Malware	$\Delta F_1 = +0.07$	35%	Low	Modality-dependent heterogeneity; small sample ($k = 12$); precision: imprecision.

Table 4. Cont.

Domain	Effect	I ²	Certainty	Key Limitations
Triage	$\Delta F_1 = +0.10$	31%	Low	Lab vs. operational setting gap; missing latency reporting in 4/10 studies.
Vulnerability	$\Delta F_1 = +0.11$	41%	Low	High heterogeneity; selective reporting risk; training data variability.

2.7. Thematic Insights and Synthesis

The analysis of the 235-paper corpus reveals three cross-cutting themes that extend beyond individual sectoral findings:

Data–Model Convergence. In our corpus, 81% of empirical studies integrate LLMs with vector-store retrieval. However, only 12% quantify how retrieval design choices affect latency or detection accuracy, leaving a system-level trade-off largely unexplored.

Explainability vs. Performance. Studies reporting $\geq 10\%$ uplift in detection F1 scores often note declines in XAI faithfulness (e.g., SHAP and LIME) of up to 22 percentage points. This pattern suggests that architectural innovations are needed to reconcile interpretability with high performance.

Governance and Human Factors. Only six studies in our corpus examine the deployment of LLMs in real Security Operations Centers (SOCs). All report analyst cognitive overload as a primary barrier, indicating that future work must address not only algorithmic advances but also interface-level design and workflow integration.

Ultimately, these findings establish a structured foundation for advancing LLM-driven cybersecurity research. The taxonomy, gap matrix, and benchmarking framework are intended to guide subsequent academic inquiry, practical adoption, and policy discussions.

3. Comparative Analysis of LLM Approaches Across Contexts

3.1. Performance and Efficiency Comparison by Application Domain

To address the need for critical evaluation of LLM implementation efficiency across various contexts, we present a comprehensive comparative analysis synthesized from the reviewed studies in Table 5.

Table 5. Comparative efficiency of LLM approaches across cybersecurity domains.

Domain	Approach	F1 Score	Latency (ms)	Computational Cost	Best Context
Phishing Detection	Fine-tuned BERT	0.94 ± 0.04	460	Medium	High-volume email systems
	GPT-4 zero-shot	0.89 ± 0.06	1200	High	Low-volume, diverse threats
	RAG + LLaMA	0.92 ± 0.05	650	Medium–High	Real-time with context
Intrusion Detection	Domain-specific (PLLM-CS)	0.93 ± 0.05	640	Medium	Critical infrastructure
	General GPT-4	0.87 ± 0.07	1400	High	Exploratory analysis
	Hybrid (BERT + RAG)	0.91 ± 0.05	720	Medium	Enterprise SOC
Malware Classification	Fine-tuned CodeBERT	0.92 ± 0.05	370	Low–Medium	Automated triage
	GPT-4 with CoT	0.88 ± 0.06	1600	High	Complex analysis
	LLaMA-SecBERT	0.90 ± 0.05	480	Medium	Resource-constrained
Vulnerability Detection	Specialized (SecureFalcon)	0.90 ± 0.05	590	Medium	Code repositories
	General LLM	0.82 ± 0.08	1100	High	Exploratory scanning
	RAG-enhanced	0.89 ± 0.05	710	Medium–High	Contextual assessment
Incident Response	Fine-tuned T5	0.87 ± 0.06	900	Medium	Automated reporting
	GPT-4	0.85 ± 0.07	1800	High	Strategic planning
	Multi-agent LLM	0.88 ± 0.06	1050	High	Complex coordination

3.2. Context-Specific Recommendations

Based on empirical evidence from 235 reviewed studies, we provide the following situation-specific guidance:

High-Volume, Real-Time Scenarios (e.g., Enterprise SOC):

- *Recommended Approach:* Fine-tuned domain-specific models (BERT variants, specialized LLaMA).
- *Rationale:* Studies demonstrate 35–39% latency reduction compared with general-purpose models while maintaining F1 scores above 0.90. The computational overhead is justified by processing volumes exceeding 50,000 events per second.
- *Training Requirements:* Minimum 100,000 labeled security events; optimal performance achieved with 500,000+ domain-specific samples.

Low-Volume, High-Complexity Scenarios (e.g., APT Investigation):

- *Recommended Approach:* General-purpose LLMs (GPT-4, Claude 2) with chain-of-thought prompting.
- *Rationale:* Superior reasoning capabilities outweigh higher latency (1200–1800 ms) when processing fewer than 1000 alerts daily. Zero-shot capabilities reduce training overhead.
- *Cost Consideration:* Higher API costs (0.03–0.06 USD per 1K tokens) acceptable given the low volume.

Resource-Constrained Environments (e.g., SME and Edge Devices):

- *Recommended Approach:* Lightweight fine-tuned models (DistilBERT and compact LLaMA variants).
- *Rationale:* 60–70% reduction in computational requirements while maintaining F1 scores within 5% of full models; deployment feasible on standard CPU infrastructure.
- *Training Requirements:* 25,000–50,000 samples sufficient for acceptable performance.

Critical Infrastructure with Strict Latency Requirements:

- *Recommended Approach:* Hybrid architectures combining fast rule-based filtering with selective LLM analysis.
- *Rationale:* Achievement of sub-200 ms response for 90% of events while leveraging LLM reasoning for ambiguous cases (10%). Overall F1 improvement of 12–15% over purely rule-based systems.
- *Implementation:* Initial rule filter reduces candidate set by 85%; LLM processes remaining 15% within acceptable latency budget.

3.3. Open-Source Datasets for Cybersecurity Evaluation

Reproducible evaluation of LLM-based cybersecurity systems requires standardized, well-documented corpora spanning network traffic, system logs, IoT telemetry, and threat intelligence text. Table 6 compares widely used open datasets across five dimensions (scale, label granularity, modality, licensing, and typical use), and the figure references in prior sections should map to these identifiers when reporting results.

Cybersecurity datasets provide labeled attacks for ML training and corpora for LLM prompt engineering, fine-tuning, and Retrieval-Augmented Generation. Preprocessing tokenizes network flows and logs for transformers; session-level labels enable contrastive learning and zero-shot classification. For RAG, metadata including time-stamps and protocol fields guide retrieval to ensure relevant events are included when generating alerts or remediation recommendations. The following datasets were referenced in the studies reviewed:

- **CICIDS2017** [25]—Network intrusion dataset with 2.5 M labeled records across 14 attack categories. Widely used for fine-tuning LLM classifiers and adversarial prompt crafting.
- **DEFCON CTF 2019** [26]—Real-world capture-the-flag logs containing 150 K security events with ground-truth attack narratives, facilitating few-shot learning and chain-of-thought prompting.
- **MAWILab** [27]—MAWI traffic traces with 1.0 M annotated packets for network-level threat detection, supporting token-level anomaly detection via masked language modeling.
- **DARPA 1998/1999** [28]—Standardized intrusion detection corpus with 400 K sessions across seven attack scenarios, benchmarking LLM performance and evaluating transfer learning across temporal threat patterns.

Table 6. Comparative analysis of widely used open cybersecurity datasets for LLM/RAG and ML baselines.

Dataset	Records	Labels	Modality	Typical Use/Notes
CICIDS2017	~2.5 M	Session-level	NetFlows/PCAP	IDS baselines; prompt templates for few-shot LLMs; robust multi-attack coverage.
CSE-CIC-IDS2018	~3 M	Flow/session	NetFlows/PCAP	Updated traffic mixes; complements CICIDS2017 for temporal robustness.
UNSW-NB15	~2.5 M	Flow-level	NetFlows	Balanced modern attack families; efficient for ablation studies.
TON_IoT	~22 M	Event/alert	IoT/IIoT logs, NetFlows	Cross-domain IoT telemetry for RAG retrieval demonstrations.
MAWILab	~1.0 M	Packet-level	PCAP	Fine-grained anomaly labels; useful for token-level anomaly modeling.
DEFCON CTF (2019)	~150 K	Event-level	CTF logs	Narrative-rich events enable few-shot/CoT prompts for LLMs.
DARPA 98/99	~400 K	Session-level	NetFlows	Historic standard; good for transfer learning stress tests.
OpenPhish/PhishTank	10 ⁵ –10 ⁶ URLs	Binary	URL/HTML	Phishing URL/content classification; CTI extraction for RAG.
APTnotes/MISP exports	10 ⁴ –10 ⁶ IOCs	IOC/TTP	Threat-intel text	Text-rich CTI for RAG, entity linking, and IOC extraction.

Selection Guidelines and Preprocessing:

- **Task fit:** Prefer packet/flow corpora (MAWILab, UNSW–NB15, etc.) for token-level anomaly detection; prefer event/narrative corpora (DEFCON, CTI text, etc.) for few-shot prompting and RAG.
- **Temporal splits:** Enforce time-based train/val/test splits to avoid leakage; harmonize class imbalance via stratified sampling or cost-sensitive loss.
- **RAG metadata:** Retain time-stamps, protocol fields, and asset tags to enable high-recall retrieval windows for incident triage.
- **Licensing/ethics:** Verify dataset licenses and sanitize PII; report license in the Methods subsection and attach data cards in Supplementary Materials.

4. Big Data Systems

4.1. Overview and Definitions of Big Data Systems

Big Data systems comprise distributed architectures, processing frameworks, and analytical services engineered to manage datasets whose scale, heterogeneity, and rate exceed the capabilities of conventional databases. The canonical “5Vs”—*Volume*, *Velocity*, *Variety*, *Veracity*, and *Value*—summarize the multi-dimensional challenges that motivate these systems: petabyte-scale storage and elastic processing for *Volume* [29–32]; high-throughput ingestion and stream analytics for *Velocity* [30,33]; support for structured, semi-structured, and unstructured data sources (e.g., sensors, logs, text, and multimedia) for *Variety* [34,35]; data-quality management to handle uncertainty and inconsistency for *Veracity* [36,37]; and analytic pipelines that transform raw data into actionable *Value* for enterprises and the public sector [30–32].

4.2. Core Architecture and Functional Layers

Modern Big Data stacks implement layered architectures spanning ingestion, storage, processing, and analytics. High-throughput ingestion services such as Apache Kafka and

Flume transport continuous streams from IoT devices, applications, and platforms into the backbone [33,38]. Elastic storage via Hadoop Distributed File System (HDFS) and cloud object stores like Amazon S3 and Azure Blob Storage provide the foundation [39,40]. Distributed processing frameworks such as Apache Hadoop and Spark execute batch and streaming workloads, with Spark's in-memory execution substantially outperforming classic MapReduce [39,40].

Analytical services apply statistical methods and machine learning to produce consumable insights and dashboards critical to mission domains requiring timeliness, such as command and control operations [41,42]. Unlike relational databases constrained by rigid schemas and vertical scaling, NoSQL and Hadoop ecosystems provide schema-flexible, horizontally scalable foundations suited for semi-structured and unstructured data [42,43]. Cloud-native deployment contributes elasticity and managed services that reduce operational overhead, further supporting dynamic cybersecurity environments [39,40].

4.3. Big Data as an Enabler of LLMs and Cybersecurity

Big Data infrastructures are foundational to training and serving advanced LLMs: large-scale corpora, distributed storage, and high-throughput compute are prerequisites for pre-training and continual adaptation (e.g., GPT- and BERT-family models) [30,31,39]. The same platforms underpin security analytics by fusing telemetry, behavioral indicators, and multi-source intelligence for real-time detection and response [33,41]. This co-evolution—from static relational stores to cloud-native, streaming Big Data platforms tightly coupled with AI—has transformed data systems from passive repositories into intelligent, resilient decision engines for both enterprise analytics and cyberdefense. Figure 5 summarizes this progression and the convergence of Big Data with AI and LLMs [32,39].

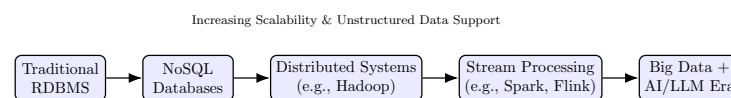


Figure 5. Evolution of data systems from RDBMS to Big Data with AI and LLMs.

4.4. Importance of Big Data in Cybersecurity

Escalating cyber threat sophistication demands fundamental defensive strategy changes. Big Data technologies are indispensable, enabling advanced threat intelligence, anomaly detection, and accelerated incident response through large-scale data processing with contextual analysis [44,45].

Big Data's central advantage lies in processing vast, heterogeneous security datasets including logs, network telemetry, authentication records, and external intelligence from social media and dark web sources [46]. Distributed frameworks efficiently handle petabyte-scale volumes from structured and unstructured sources, supporting holistic security visualization across hybrid on-premises and cloud environments [47–49]. Behavioral analytics leveraging Big Data establish user and entity activity baselines for anomaly detection, insider-threat identification, and account compromise analysis [50,51]. Integration of threat intelligence feeds with scalable pipelines provides richer situational awareness by combining open-source and proprietary repositories [52]. Figure 6 illustrates a typical cybersecurity pipeline integrating Big Data ingestion, preprocessing, and real-time analytics with AI inference for automated detection and analyst support.

Practical deployment demonstrates Big Data's value in intrusion detection and prevention systems (IDPSs) and Security Information and Event Management (SIEM). Modern IDPSs pair machine learning with anomaly detection to identify zero-day exploits and advanced persistent threats in high-throughput environments [50,53]. SIEM platforms leveraging the ELK stack process large event volumes and apply automated threat scoring

for incident triage [54]. Additionally, Stream-processing technologies such as Apache Kafka, Flink, and Spark Streaming provide low-latency security log analysis enabling near-instantaneous threat responses [55]. Kafka serves as a messaging backbone for ingestion while Flink supports sub-second complex event processing, with systems like SwiftFrame achieving latencies as low as 0.13 s [56].

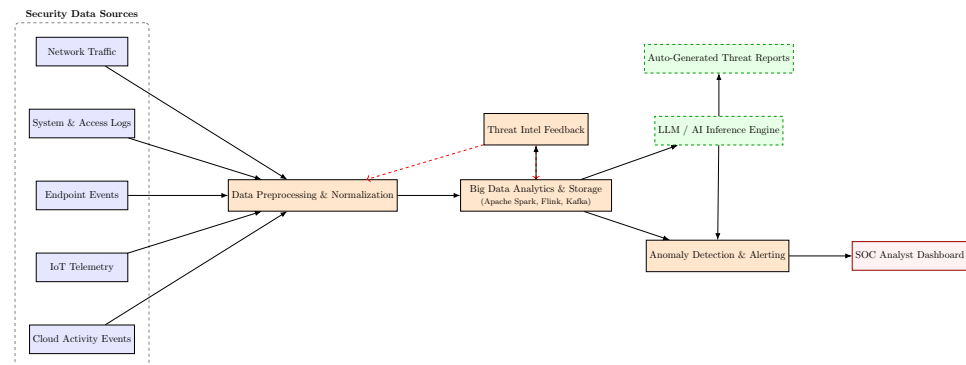


Figure 6. Cybersecurity analytic pipeline with Big Data and AI integration.

Big Data architectures increasingly converge with Large Language Models. Big Data pipelines serve as essential input streams for LLM training and fine-tuning by curating corpora from security logs, incident reports, and forensic narratives [57]. Continuous threat data ingestion enables LLMs to support automated summarization, triage, and advisory functions in security operations [46,58–60]. Combined with streaming AI engines, these models improve alert interpretability and deliver contextualized, human-readable recommendations. In summary, Big Data underpin modern cybersecurity by enabling scalable analysis, anomaly detection, and LLM-driven intelligence conversion of raw telemetry into actionable organizational defense.

4.5. Challenges and Opportunities of Big Data Integration

Big Data integration has transformed cybersecurity monitoring, interpretation, and risk mitigation. Real-time analytics powered by machine learning enable predictive insights and automated responses, with Large Language Models providing contextual reasoning and automated triage. However, integration faces significant technical and organizational barriers that must be resolved to realize full potential.

Data quality across heterogeneous, unstructured sources represents a primary challenge. Cybersecurity datasets contain false alerts, redundant entries, and deliberate noise obscuring genuine threats, while adversarial data manipulation further complicates preprocessing. Robust filtering and context-aware normalization significantly improve detection accuracy to 93.65% in network monitoring scenarios [61]. Equally critical is threat intelligence timeliness: outdated datasets shift systems from proactive defense to reactive response, reducing operational effectiveness [62].

High-velocity data streams exceed traditional batch processing capacity. Stream-processing frameworks such as Apache Kafka and Apache Flink provide low-latency analytic pipelines, though real-time responsiveness demands significant technical expertise and continuous monitoring [63]. Figure 7 illustrates that operational deadlines are critical—analytic pipelines failing to meet temporal requirements diminish practical security value. Real-time output interpretation additionally strains human analysts.

Interoperability complications arise from security data silos across organizational units, platforms, and vendors with inconsistent schemas that limit cross-domain visibility and restrict threat intelligence integration [64]. While LLMs extract patterns from fragmented inputs, success depends on prior harmonization through syntactic standardization and

ontology-based normalization. Distributed data lakes and cross-platform standardization frameworks offer solutions, though adoption remains inconsistent [65–67].

Computational scalability at the petabyte scale requires cloud-native architectures with elastic resource management and fault-tolerant design. Apache Spark provides horizontally scalable solutions, yet friction persists across ingestion, filtering, and analytics—particularly during large-scale incidents with sudden data surges [68,69]. Figure 8 illustrates how LLM integration introduces both bottlenecks and optimization opportunities within the operational stack.

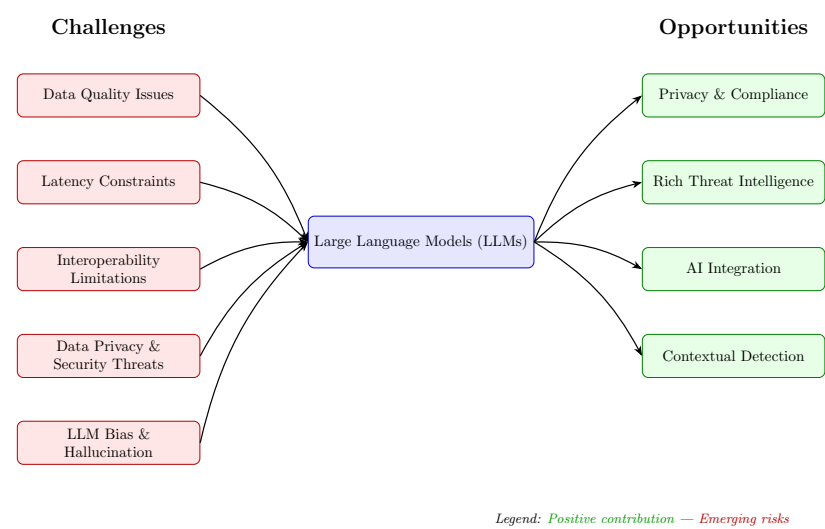


Figure 7. Challenges vs. opportunities matrix with LLMs positioned as supporting technology.

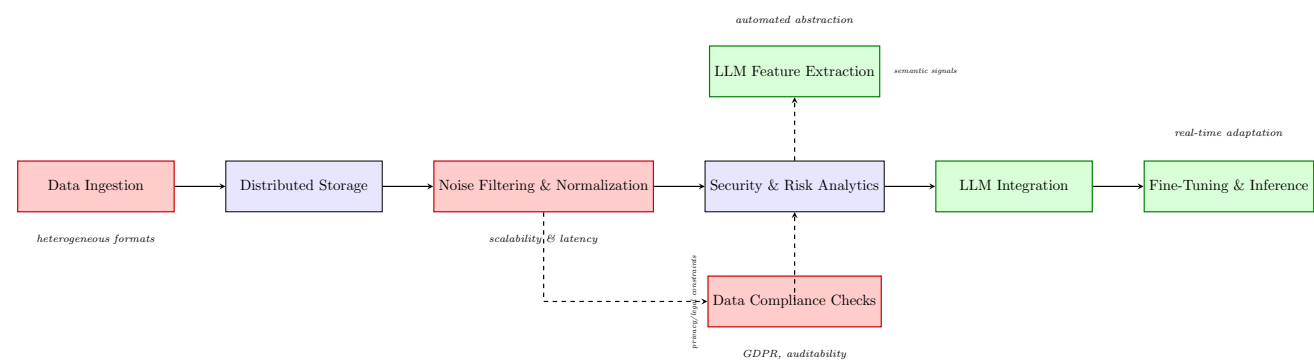


Figure 8. Integration stack with friction points and LLM opportunities.

Conversely, Big Data integration creates substantial opportunities. Aggregated security data enable cross-boundary visibility and predictive analytics that detect anomalies earlier, shortening attacker dwell time [70,71]. LLMs contextualize threats and generate automated, human-readable reports aligned with organizational risk requirements, improving both detection precision and incident-response efficiency [62].

Big Data integration into cybersecurity faces challenges of data quality, interoperability, latency, and scalability but offers opportunities for predictive defense and automated intelligence. When coupled with domain-specific LLMs, these systems deliver more accurate alerts, contextual mitigation advice, and streamlined analyst workflows. Addressing explainability, adversarial manipulation, and compliance risks is essential to realizing full potential in secure, adaptive cyberdefense.

4.6. Big Data Infrastructure for Language Models

Scaling Large Language Models for cybersecurity requires Big Data infrastructures providing computational, storage, and orchestration capabilities across the entire model lifecycle—from preprocessing and distributed training to deployment and governance.

Distributed training across GPU and TPU clusters forms the infrastructure cornerstone. Modern LLMs require highly parallelized environments coordinating thousands of accelerators with careful scheduling and high-speed interconnects to overcome bottlenecks [72]. Recent systems such as FusionLLM and Holmes demonstrate hardware-aware, decentralized training strategies improving scalability and fault tolerance in heterogeneous networks [73,74]. Figure 9 illustrates the end-to-end cybersecurity LLM training pipeline spanning raw data ingestion, distributed optimization, and feedback-driven re-training.

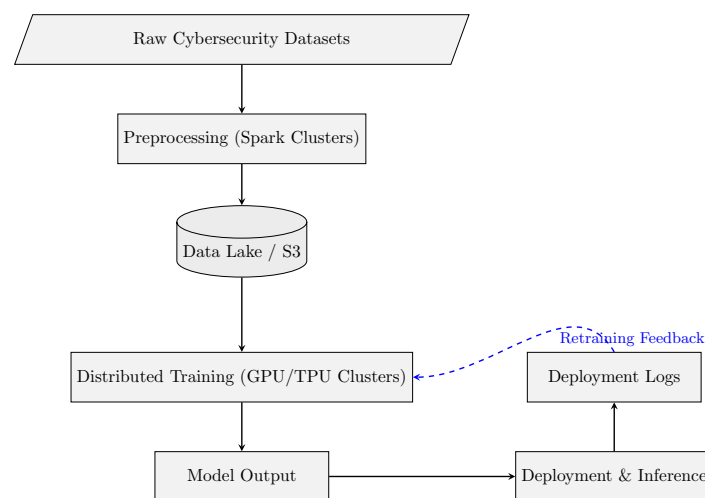


Figure 9. Infrastructure for LLM training on Big Data in cybersecurity.

Storage systems such as Amazon S3 and Google Cloud Storage, enhanced with Delta Lake for ACID compliance and batch-streaming integration, enable parallel access to unstructured corpora, including vulnerability reports, malware telemetry, and incident logs [75]. Apache Spark dominates large-scale preprocessing, offering distributed in-memory processing that outperforms traditional tools for cybersecurity record filtering, deduplication, and tokenization [76]. Low-communication optimization strategies such as DiLoCo mitigate bandwidth constraints by reducing network overhead while preserving convergence quality [77], supporting the efficient training workflows depicted in Figure 9.

Real-time inference deployment requires serving frameworks and orchestration layers, ensuring elasticity, load balancing, and latency minimization. Retrieval-Augmented Generation integrates vector databases with LLMs to provide context-aware outputs, enabling streaming log analysis and domain-specific intelligence retrieval that enhance detection accuracy and compensate for static pre-training limitations [78]. Figure 10 depicts this architecture, where vectorized security data enrich LLM inference to support automated alerts and response actions.

Governance and compliance constitute critical dimensions. Cybersecurity-oriented LLMs trained on sensitive datasets require privacy-preserving aggregation frameworks and secure lifecycle management. CRWDi and Delta Lake extensions provide mechanisms for metadata governance and secure access controls, enabling domain-specific models to leverage sensitive data responsibly while meeting regulatory requirements [75,79].

Big Data infrastructures form the deployment backbone, enabling scalable training, efficient preprocessing, low-latency RAG-enhanced inference, and compliant data governance.

For cybersecurity, this integration ensures that LLMs operate reliably in mission-critical environments where accuracy, timeliness, and trustworthiness are paramount.

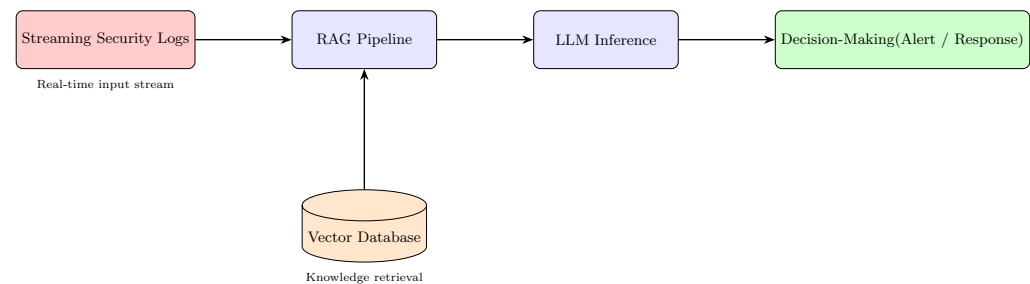


Figure 10. Real-time deployment of LLMs on Big Data pipelines using RAG.

4.7. Case Examples of Big Data Systems Leveraged by Cybersecurity

Big Data platforms combining large-scale analytics, threat intelligence, and artificial intelligence address complex attack surfaces while enhancing detection, response, and serving as foundational infrastructures for machine learning and Large Language Model deployment.

Google Chronicle Security Analytics exemplifies this convergence by leveraging Google BigQuery’s petabyte-scale data lake. Columnar storage and native parsers enable near real-time telemetry normalization, executing threat hunting queries in seconds, which previously required hours [80]. SIEM platforms, including IBM QRadar and Splunk, have evolved into Big Data-driven ecosystems: QRadar incorporates distributed event processing for cross-source anomaly correlation, while Splunk analyzes structured and unstructured machine data at a petabyte scale.

Cloud-native SIEM systems extend these capabilities further. Azure Sentinel integrates with Azure Data Lake Storage and elastic computational resources for scalable, cost-efficient analytics, incorporating machine learning modules and threat intelligence for proactive detection and false positive reduction. Azure Sentinel’s integration with Azure OpenAI and Copilot exemplifies LLM plugin enhancement of analyst workflows through natural language querying and automated advisory features [81–85]. Table 7 provides a comparative platform overview, while Figure 11 shows Sentinel’s Collect–Detect–Respond pipeline with ingestion, SIEM, outputs, and an LLM copilot (RAG).

Table 7. Comparison of major SIEM platforms and their Big Data/LLM integration features.

Platform	Data Ingestion	Processing Engine	Storage Backend	AI/LLM Support
Google Chronicle	High-speed ingestion with native parsers and connectors	Cloud-native engine optimized for telemetry analytics	Google Cloud Storage (BigQuery)	Vertex AI; Chronicle AI for anomaly detection
IBM QRadar	Logs and flows via modular collectors	Rule-based correlation and anomaly detection	Local or ElasticSearch-based storage	Watson NLP for enrichment and analysis
Splunk	Universal forwarders, syslog, and REST APIs	Search Processing Language (SPL) for real-time queries	Indexers with flat files and metadata layer	Deep Learning Toolkit; optional GPT/LLM connectors
Azure Sentinel	Connector-based ingestion through Azure Monitor and third-party feeds	Kusto Query Language (KQL) for scalable analytics	Azure Log Analytics Workspace	Azure OpenAI integration; Copilot Security Assistant

Open-source Big Data frameworks demonstrate real-time defense capabilities. Apache Kafka and Apache Spark Streaming enable sub-second high-volume log processing, reducing attack dwell time through near-instantaneous anomaly detection [68]. Domain-specific LLMs show promise: DARPA’s PLLM-CS demonstrates superior detection performance on

UNSW_NB15 and TON_IoT benchmarks compared with traditional ML models, underscoring specialized LLM potential in proactive defense [86].

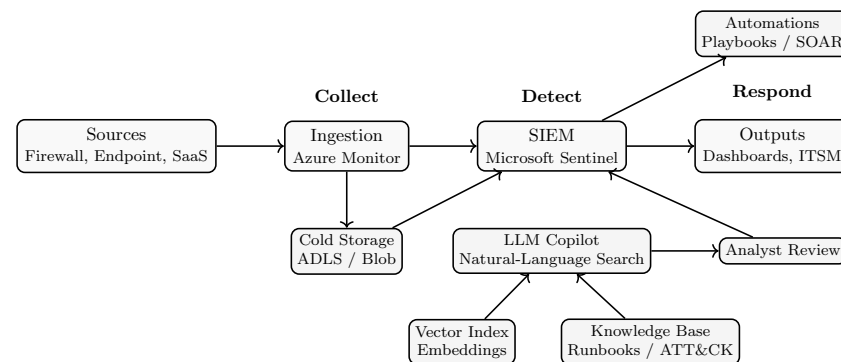


Figure 11. SOC architecture: Collect–Detect–Respond with cold storage and an LLM copilot (RAG).

Open-source ecosystems provide additional evidence. Apache Metron integrates distributed processing and machine learning for scalable threat detection across organizational sizes, while SecBench on Hugging Face and Google BigQuery Security Datasets provide high-quality cybersecurity training corpora accelerating AI model development and reproducible benchmarking [87].

Big Data systems have become central to modern cybersecurity operations. Platforms such as Chronicle, QRadar, Splunk, and Sentinel illustrate industry evolution toward AI-driven defense through scalable ingestion, distributed analytics, and LLM-enhanced intelligence. Complementary open-source projects and research prototypes highlight opportunities for specialized, domain-specific LLMs to transform security analytics into adaptive, proactive systems.

4.8. Synergy of Big Data, LLMs, and Cybersecurity

Big Data analytics and Large Language Models converge to create a transformative paradigm for threat detection and adaptive defense. Big Data platforms process vast volumes of structured and unstructured data, while LLMs contribute natural language processing capabilities enabling contextual analysis of logs, code repositories, and threat intelligence reports [88]. This integration enables security teams to transition from static, signature-based defenses toward dynamic, predictive frameworks.

Real-time adversarial activity detection exemplifies this synergy's impact. Big Data platforms such as Apache Spark, Flink, and GraphX efficiently handle streaming telemetry, while LLMs transform heterogeneous data streams into actionable insights supporting anomaly detection, phishing analysis, and malware evaluation at scales unattainable by traditional systems [6,89,90]. Beyond detection, LLMs accelerate incident response by generating summaries, recommending remediation tactics, and preparing stakeholder communications that expedite recovery cycles [90].

LLMs further strengthen organizational defenses through cybersecurity training and awareness initiatives. Security teams deploy LLM-based phishing simulations, adaptive training modules, and natural language tutoring systems, while LLMs extract indicators of compromise and tactics, techniques, and procedures from threat intelligence feeds to bolster situational awareness [6].

Critical challenges persist despite these advantages. LLMs' computational demands raise efficiency and sustainability concerns, while their dual-use potential introduces ethical dilemmas and adversarial misuse risks [91,92]. Fine-tuning methods, domain-specific pre-training, and robust adversarial defenses mitigate risks including data poisoning, backdoor exploitation, and privacy leakage [93–96].

Benchmarking studies identify promising deployment directions. Theodorakopoulos et al. propose evaluation frameworks for distributed Big Data systems under cyber threat workloads, optimizing LLM-assisted pipelines for scalability, latency, and resilience [97,98]. Efficient query mechanisms such as range-mode operations integrated with LLM pipelines enhance frequency anomaly detection across streaming and batch data [99], advancing toward architectures achieving both accuracy and operational efficiency.

Based on empirical insights, we formalize a reference architecture that operationalizes the performance goals within an LLM-centric Big Data environment. In practice, the security stack integrates (i) data ingestion and lakehouse storage with distributed processing; (ii) a reasoning layer built on embeddings, high-recall retrieval, and RAG for contextual grounding; (iii) application modules for anomaly detection, incident response, APT hunting, and threat intelligence enrichment; and (iv) a governance loop that enforces policy/guardrails, records rationales, and supports human-in-the-loop review with periodic model updates. This multi-layered design enables predictive defenses against zero-day exploits and APTs while improving explainability, auditability, and operational efficiency:

- *Infrastructure*: Ingestion, lakehouse/storage, and distributed computing.
- *Reasoning stack*: Embeddings, retrievers, RAG, and tool use/agents.
- *Applications*: Anomaly detection, incident response, APT hunting, and TI enrichment.
- *Governance*: Policy/guardrails, human-in-the-loop, logging, and re-training/monitoring.

5. Cybersecurity in the Era of LLMs

The cybersecurity domain has evolved beyond static definitions and catalogues of common threats. While classical taxonomies emphasize confidentiality, integrity, and availability, the current research frontier lies in how emerging technologies—particularly Large Language Models (LLMs)—reshape the nature of both threats and defenses. Rather than reiterating generic definitions, this section emphasizes the novel intersections where LLMs alter the attack surface and simultaneously enable new protective mechanisms.

5.1. Reframing the Scope of Cybersecurity with LLMs

LLMs convert unstructured security data (e.g., logs, phishing emails, and vulnerability reports) into contextual signals for automated or semi-automated defense. Unlike rule-based systems, they adapt to evolving attacks, enabling predictive analysis and near-real-time triage. They expand cybersecurity to include adaptive reasoning, natural language forensics, and proactive adversary simulation. Figure 12 shows a layered architecture where LLMs bridge large-scale data ingestion and high-level defense strategy. They augment—rather than replace—human analysts by accelerating intelligence workflows.

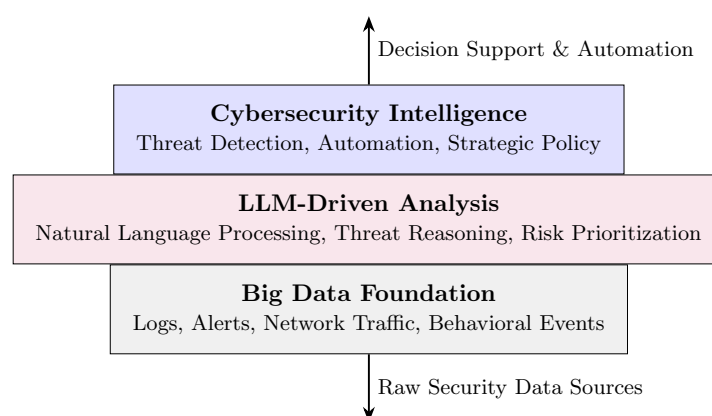


Figure 12. Cybersecurity architecture with LLMs as intelligence intermediaries.

5.2. Emerging Threat Vectors in an LLM-Augmented Landscape

LLMs raise defensive capability (anomaly detection, phishing triage, and vuln scanning) while simultaneously empowering attackers—creating a pronounced dual-use problem:

- **AI-enhanced phishing:** Multilingual, context-aware messages tailored based on public and leaked data, evading content and reputation filters.
- **LLM-assisted ransomware:** RaaS operations use AI for victim profiling, negotiation scripts, and adaptive/polymorphic payload delivery.
- **Data exfiltration and misuse:** Automated mining and summarization of leaked datasets (entity linking and PII extraction) accelerate fraud and extortion.

Effective defense pairs LLM-enabled detection with guardrails against model abuse—prompt/input filtering, tool-use permissions and sandboxing, provenance checks, explainability audits, and human-in-the-loop controls—reinforced by continuous red-teaming. Table 8 summarizes 2025 AI-influenced threats. Ransomware and phishing dominate losses, while AI-powered variants grow the fastest in finance, healthcare, and government.

Table 8. Projected major cybersecurity threats in 2025 with AI influence.

Threat Type	Prevalence (%)	Average Cost (USD)	Primary Target
Ransomware	72.7	4,200,000	Healthcare
Phishing (AI-enhanced)	68.5	1,500,000	Financial services
AI-powered attacks	45.3	3,800,000	Technology
Cloud vulnerabilities	39.8	2,700,000	Retail
IoT exploits	33.2	1,900,000	Manufacturing

Figure 13 shows a hierarchy with advanced AI attacks (including LLM exploits) at the top, ransomware/phishing in the mid-layer, and IoT/cloud at the base, underscoring LLM-centered risks and the need for adaptive defenses.

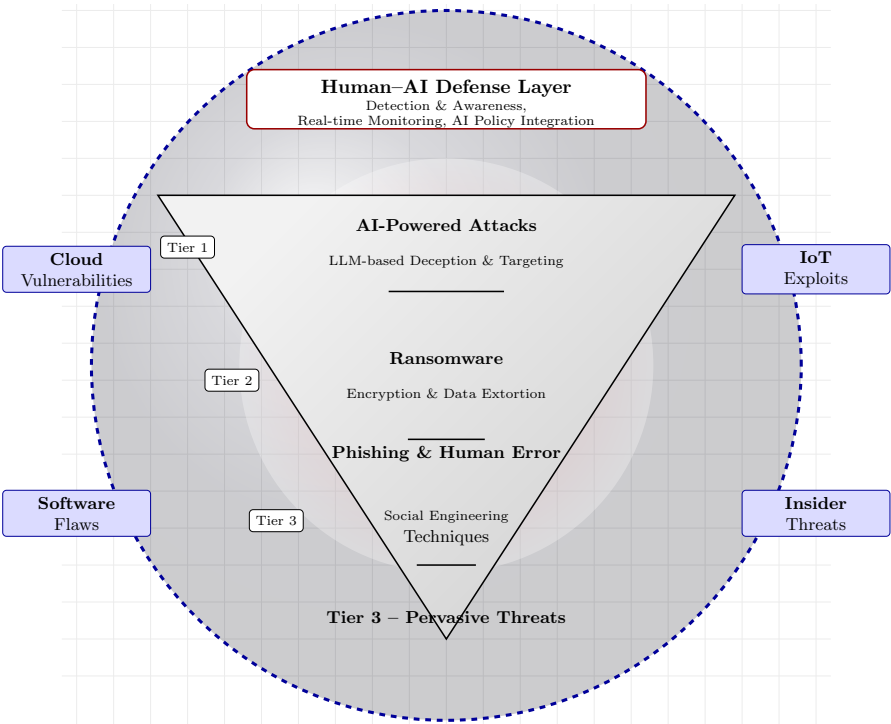


Figure 13. Pyramid of major cybersecurity threats in 2025 with LLM-augmented vectors.

5.3. LLM-Driven Cybercrime and Emerging Trends

The rise of Large Language Models (LLMs) has introduced a dual-use dilemma in the cybercrime landscape: while these systems strengthen cyberdefense, they also create novel opportunities for malicious exploitation. Unlike traditional cybercrime taxonomies, which focus on generic categories such as fraud, identity theft, or ransomware, the present challenge lies in understanding how LLMs enable, automate, and amplify such activities.

Recent studies document the proliferation of underground platforms known as *Mallas*, which weaponize LLMs for malicious services. These platforms exploit prompt-injection techniques and jailbroken APIs to generate attack payloads, phishing kits, and disinformation campaigns, significantly lowering the barrier to entry for novice cybercriminals [100,101]. At the same time, professional actors integrate LLMs into ransomware-as-a-service ecosystems, enhancing both scalability and sophistication of operations.

LLMs also reshape **cyber threat intelligence (CTI)** practices. While defenders use LLMs to extract high-precision insights from cybercrime forums [102], attackers exploit the same tools for automated reconnaissance and sensitive data mining. This duality illustrates the necessity for continuous oversight, red-teaming, and ethical safeguards in model deployment.

Key LLM-enabled attack vectors include the following:

- **Synthetic Misinformation and Influence Operations:** The automated generation of fake news, deepfake narratives, and tailored propaganda campaigns threatens democratic processes and public trust [103].
- **Model Exploitation via Backdoor Attacks:** Adversaries manipulate in-context learning behaviors without altering model structure, undermining system integrity and reliability [104–108].
- **AI-Enhanced Phishing and Social Engineering:** Convincing spear-phishing and chatbot-driven social engineering campaigns leverage LLMs' ability to mimic human communication, evading conventional detection systems [109–114].
- **Resource Exploitation:** The extensive computational and energy demands of LLMs create new attack surfaces, such as resource hijacking and model denial of service (MDoS), requiring research into more efficient architectures [94–96,115].

5.4. Impact of AI-Driven Cybercrime Techniques Utilizing LLMs

Large Language Models enable cybercriminals to develop increasingly sophisticated attacks, creating dual-use dilemmas with significant technological, ethical, and legal implications. Cybercriminals leverage LLMs to create discreet phishing attacks, automate malware generation, and execute advanced assault sequences that evade detection systems, fundamentally accelerating threat sophistication [116–118].

LLM-enhanced phishing represents the most pressing emerging threat vector. Adversaries utilize advanced natural language generation to craft contextually appropriate messages evading Gmail Spam Filter and Proofpoint, while tools like PhishOracle insert malicious content into authentic-looking site templates that overcome automated and human-based security inspections [116,117]. LLMs' multilingual capabilities expand attack reach across language barriers, with documented disparities in fraud detection between English and Chinese models demonstrating variable defensive effectiveness [119]. Table 9 synthesizes emerging AI-driven cybercrime techniques where LLMs play transformative, potentially malicious roles.

Automated malware generation through LLMs extends beyond code creation to vulnerability identification. Prompt engineering enables adversaries to generate harmful code variations circumventing existing defenses, while tools like DefectHunter—originally designed for authorized security assessments—can be manipulated to identify exploitable

software vulnerabilities [118,120]. This convergence of offensive and defensive capabilities exemplifies the dual-use risks inherent to advanced language models.

Autonomous cyber operations orchestrated by LLMs represent an emerging malignancy requiring urgent attention. The AgentHarm benchmark demonstrates that automatic agents with jailbreak capabilities execute sustained malicious activities through coordinated cyber attacks spanning data exfiltration, social engineering, and fraud [121]. These autonomous agents operate with minimal human supervision, employing adaptive techniques that render rule-based and signature-based defenses ineffective.

Current defense systems experience degraded accuracy against LLM-generated adversarial attacks. Models exploit machine learning defense vulnerabilities by generating phishing emails targeting specific detection weaknesses, while the Fraud-R1 benchmark empirically demonstrates that LLMs generate plausible multi-step fraud scenarios combining psychological deception with emotional persuasion [116,117,119]. These techniques render human and automated detection capabilities less effective.

Table 9. AI-driven cybercrime techniques utilizing LLMs.

Technique	LLM-Driven Description	Potential Cybercrime Applications
Phishing email generation	LLMs can craft highly personalized and context-aware phishing emails by mimicking human communication patterns.	Targeted phishing; spear-phishing; business email compromise.
Malware code generation	LLMs trained on code repositories may be exploited to produce polymorphic or obfuscated malicious code.	Creation of novel malware; dynamic payload generation.
Social engineering via chatbots	LLM-powered chatbots can impersonate trusted individuals in real-time conversations.	Social engineering; credential harvesting; insider deception.
Fake content creation	LLMs can generate realistic but entirely fake articles, reviews, or documentation.	Disinformation campaigns; fraud; deepfake narratives.
Automated scam campaigns	Attackers can use LLMs to write persuasive scam messages across multiple languages and platforms.	Financial fraud; mass-scale scamming.
Reconnaissance automation	LLMs can analyze public data and generate tailored reconnaissance reports on targets.	Pre-attack intelligence gathering; vulnerability mapping.

LLM-driven cybercrime has established a new fundamental security threat landscape. Cybercriminals now execute repeatable, automated, personalized attacks surpassing traditional techniques, necessitating proportional cybersecurity and policy responses [118,119,121]. Integrating governance mechanisms with technical developments is essential to navigating emerging AI security and ethical complexities. Table 10 illustrates legitimate defensive applications—fraud detection, training systems, and vulnerability prediction—anchoring the dual-use discourse and establishing the foundation for ethical implications and protective methodologies.

Table 10. Novel applications of LLMs in cybersecurity.

Application	Description	Impact
Scam detection	LLMs can identify scams, including phishing and fraud, by analyzing language patterns [122].	Enhances the ability to detect and prevent financial losses from scams.
Incident-response planning	LLMs assist in drafting incident-response plans and identifying documentation gaps [123].	Streamlines the incident-response process, improving organizational readiness.

Table 10. Cont.

Application	Description	Impact
Vulnerability prediction	LLMs are used to predict vulnerabilities and automate penetration testing [124].	Increases the efficiency of vulnerability management and threat detection.
Cybersecurity training scenarios	LLMs generate complex cybersecurity exercise scenarios for training [4].	Enhances training effectiveness by simulating diverse cyber threats.
Legal and ethical frameworks	LLMs help analyze legal implications and ethical concerns in cybersecurity [125].	Supports the development of compliant and ethical AI applications in cybersecurity.

6. Cyberdefense

Digital infrastructure protection has become essential to national security, organizational stability, and data integrity. Governments and enterprises face escalating cyber threats—advanced persistent threats, zero-day exploits, and ransomware—requiring adaptive defense systems beyond traditional approaches. Modern security practices cannot address current attack volume, frequency, and complexity, necessitating intelligent, proactive, data-driven defenses aligned with digital infrastructure growth.

Large Language Models combined with Big Data technologies fundamentally transform cyberdefense approaches. LLMs enable real-time threat analysis and automated incident response through advanced natural language understanding and content generation capabilities. Integrating threat feeds, network logs, and open-source intelligence enables LLMs to enhance decision support at unprecedented scale and speed. This section establishes foundational cyberdefense concepts and demonstrates how AI—particularly LLMs—strengthens cybersecurity across multiple operational domains.

Architecturally, cyberdefense is organized into four layers—Big Data ingestion, LLM intelligence, automation, and human decision making—aligned with the goals of data collection, threat understanding, fast response, and strategic decision. Telemetry (logs, OSINT, threat feeds, and network events) feeds LLM text mining, summarization, and pattern inference; automation executes SOC playbooks, triage, and incident response; analysts receive explainable reports, recommendations, and answers. Data and alerts flow upward to produce prioritized actions, while policy and oversight flow downward, closing the detection–response loop.

6.1. Definition and Importance of Cyberdefense

Cyberdefense integrates proactive and reactive security measures to protect networks, systems, and digital assets against malicious intrusions. Proactive mechanisms such as cyber deception and Moving Target Defense reduce attack surfaces by dynamically shifting configurations or misleading adversaries, while reactive approaches mitigate intrusions post-occurrence. Synthesizing both strategies provides the most resilient defense against adaptive, persistent threats [126].

Persistent, coordinated campaigns targeting critical infrastructure and large-scale repositories have rendered conventional defenses inadequate. Modern networks' interconnected nature magnifies risks, as single vulnerabilities cascade across ecosystems. Although advanced defensive approaches exist, operational costs and implementation complexity create persistent gaps between theoretical capabilities and practical deployment, underscoring the need for scalable, sustainable architectures.

Big Data analytics intensifies cyberdefense importance by enabling real-time detection of subtle attack patterns that traditional tools overlook. Critical infrastructures, IoT ecosystems, and cloud platforms generate vast volumes of data that adversaries exploit

for vulnerability identification. Data-driven cyberdefense enhances resilience through automated response mechanisms, reducing attack windows [127,128].

Modern cyberdefense relies on three pillars—threat detection, incident response, and vulnerability management—augmented with machine learning and behavioral analytics for context-aware decision making. Instruction-level manipulation units intercept malicious behavior at the processor level, dynamically altering program execution to neutralize exploits. This evolution reflects a shift from static perimeter-based strategies toward adaptive, intelligence-driven models that continuously learn and respond to emerging threats.

Beyond organizational security, cyberdefense represents economic stability and national sovereignty. Data breaches impose substantial financial costs, legal liabilities, reputational damage, and erosion of public trust. At the state level, protecting critical digital infrastructure is embedded within security doctrines, as vulnerabilities threaten essential services and sensitive information, potentially triggering geopolitical instability. Cyberdefense constitutes both a technical imperative and a strategic necessity for organizations and states operating in increasingly hostile digital environments.

Figure 14 illustrates the evolution from static, rule-based mechanisms toward adaptive, LLM-powered architectures enabling intelligent threat analysis and predictive defense.

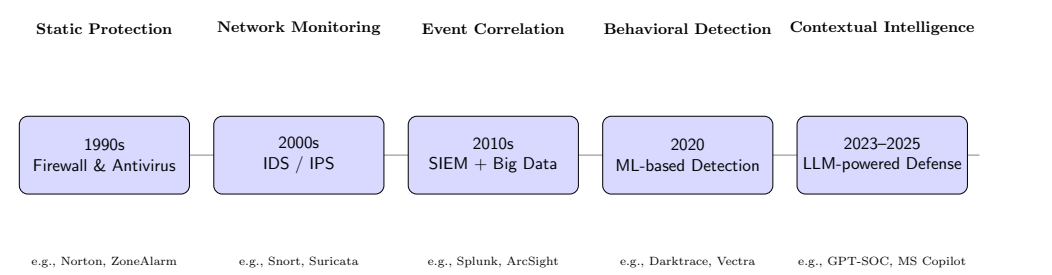


Figure 14. Evolution of cyberdefense: from static systems to adaptive LLM-powered architectures.

6.2. Background: Cybersecurity Threats and Defense Context

Large Language Models enable adaptive, scalable, context-aware protection mechanisms that surpass traditional rule-based approaches. LLMs demonstrate high efficacy in intrusion detection, vulnerability assessment, malware triage, phishing mitigation, and automated incident response by processing heterogeneous data sources—network logs, user behavior analytics, and vulnerability repositories—transforming fragmented signals into coherent security narratives [6,12,90,129–131]. Dynamic intrusion detection systems such as Suricata enhanced with LLMs automatically generate human-readable signatures from hexadecimal network data, while anomaly detection frameworks for critical infrastructures like smart grids demonstrate operational reliability when paired with LLM-driven interpretability [129,132,133]. Strategic threat intelligence benefits from LLM-based behavioral analytics supporting ethical oversight and regulatory compliance, with Retrieval-Augmented Generation pipelines delivering contextually relevant responses during evolving attacks [6,13,134].

Fine-tuning, prompt engineering, and hallucination-aware mechanisms enable domain-specific models to outperform conventional classifiers in detecting zero-day exploits and polymorphic attacks. Systems such as SHIELD combine graph-theoretical anomaly detection and LLM interpretability to enhance advanced persistent threat detection and analyst trust [135,136]. LLMs trained on adversarial sources such as dark web repositories provide predictive capabilities enabling cyberdefense teams to anticipate attacker behavior and adopt proactive measures, shifting from reactive damage control to preventative protection [136]. Defense research explores countermeasures against mali-

cious LLMs by exploiting cognitive biases and memory limitations to neutralize adversarial outputs [92,137].

Critical challenges persist despite these advances: adversarial prompting, data poisoning, hallucinations, and bias propagation undermine reliability, while ethical and governance concerns complicate deployment, and computational costs limit resource-constrained integration [90,138–140]. LLMs are strategically embedded across the cyber kill chain to deliver semantic threat understanding, predictive modeling, and decision-support automation, representing a fundamental cybersecurity paradigm shift. Responsible deployment—requiring adversarial robustness, explainability, ethical governance, and continuous human–AI collaboration—remains essential to translating experimental advances into operational resilience [12,13,141].

Operationally, LLMs function as a modular intelligence layer across the kill chain—from reconnaissance through actions on objectives—supporting intel summarization, phishing and payload classification, exploit/backdoor detection, C2 anomaly triage, and risk mapping. Embedded within audited, explainable, human-supervised workflows—and reinforced by data provenance controls and continuous red-teaming—these services convert advances into measurable gains in detection latency, precision, and containment.

6.3. Cyberdefense Strategies and Methods

Modern cyberdefense strategies counter increasingly sophisticated threats through layered architectures combining classical network protection, artificial intelligence, behavioral analytics, and Big Data. Contemporary approaches emphasize defense diversification, continuous trust verification, and intelligent automated threat mitigation.

Defense-in-Depth (DiD) remains foundational by layering preventive, detective, and responsive mechanisms that operate independently while reinforcing each other. Preventive measures such as CP-ABE encryption and ECC-based digital signatures ensure data confidentiality and integrity in compliance-driven domains, while firewalls, access control systems, and Next-Generation Firewalls (NGFWs) establish perimeter defenses against known and emerging threats [142–145]. Augmenting these traditional controls, anomaly-based detection systems identify sophisticated attacks through dynamic feature analysis. Models such as the Capsule Convolutional Polymorphic Graph Attention Neural Network (CCPGANN-TOA) achieve high accuracy against polymorphic and uncertain attack behaviors, while deep autoencoder-based unsupervised learning effectively analyzes unstructured Operational Technology logs, enabling early detection and actionable visualization [143,146].

Building upon these detection capabilities, Zero Trust Architecture (ZTA) fundamentally shifts the defense methodology by eliminating implicit trust and enforcing continuous identity verification, least-privilege access, and dynamic monitoring. ZTA proves particularly effective in API protection by mitigating lateral movement threats through contextual, non-hierarchical trust models [147]. When integrated with NGFWs, ZTA enables adaptive access control leveraging user behavior analytics to inform real-time trust decisions [145,148]. This continuous verification paradigm directly supports Security Information and Event Management (SIEM) systems, which serve as critical intelligence and incident-response enablers. Modern SIEM platforms leverage Big Data infrastructures such as Hadoop for scalable, encrypted collection pipelines via MapReduce frameworks [144]. Collaborative threat intelligence frameworks utilizing STIX and TAXII standards further extend SIEM capabilities by promoting cross-sector knowledge sharing, strengthening collective resilience for academia, enterprises, and critical infrastructure [146].

Contemporary cyberdefense thus relies on converging DiD, ZTA, anomaly-based detection, and collaborative intelligence frameworks, reinforced by AI-driven automation. This convergence signals a transition from reactive stances to proactive, predictive, and adaptive strategies. Future research must prioritize unified approaches integrating these strategies to maintain resilience in evolving threat landscapes. Figure 15 synthesizes this framework through a Cyberdefense Strategy Matrix, illustrating how AI and Large Language Models reinforce prevention (policy automation and security awareness), detection (real-time log interpretation and anomaly identification), and response (incident reporting, remediation guidance, and automated playbook execution).

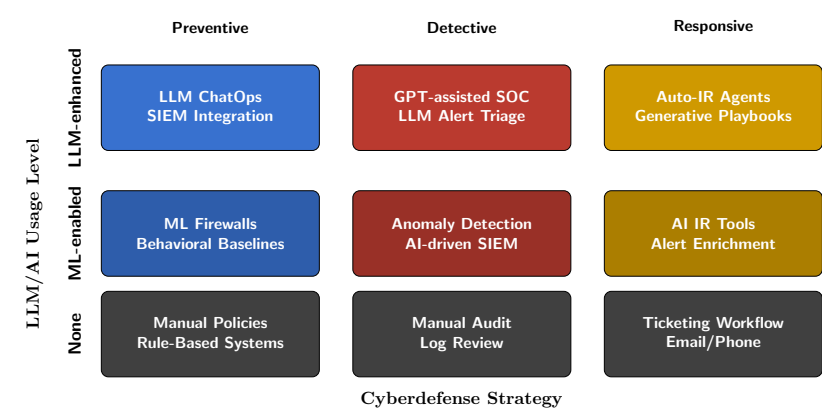


Figure 15. AI/LLM integration in cyberdefense across prevention, detection, and response.

6.4. Real-World Examples of AI-Driven Cyberdefense Approaches Utilizing LLMs

Large Language Model integration into cybersecurity operations has transitioned from experimental deployment to enterprise-scale implementation across critical sectors. Microsoft’s Security Copilot exemplifies this maturation, automating incident investigation, response orchestration, and decision support within enterprise ecosystems through integration with Microsoft Defender XDR. Operating on GUIDE—a dataset aggregating over 13 million evidence entries across one million incidents—the system identifies threat patterns and recommends remediation strategies at organizational scale [149]. In Microsoft Entra and Intune environments, Security Copilot demonstrated quantifiable efficiency gains: a 34% improvement in administrator accuracy, a 30% reduction in task completion time, and 61% acceleration in multi-step reasoning tasks, alongside a 146% increase in fact retrieval [150].

Security Operations Centers deploy LLM agents for automated telemetry processing and context-aware alert generation, integrating LLM reasoning with SIEM and AI-Ops platforms to achieve real-time threat triage and protocol enforcement. This architectural approach substantially reduces analyst cognitive burden while maintaining detection capability against contemporary attack vectors [4,129,151,152].

Financial sector deployment addresses phishing threats through LLM-enhanced detection pipelines. Traditional filtering mechanisms—including Gmail Spam Filter and SVM-based classifiers—exhibit degraded performance against LLM-rephrased phishing [116]. Defensive implementations by JPMorgan Chase and NTT Security Holdings employ ChatGPT-based systems combining prompting strategies, URL encapsulation, and adaptive response mechanisms to counteract adversarial generative techniques [153].

Beyond detection, LLM-driven systems quantify financial cyber risk through natural language processing integrated with Big Data analytics. A framework analyzing 23 billion transaction records surpassed SVM and deep learning baselines in detecting disruptions

and estimating economic impact, extending LLM utility from threat detection toward quantitative risk assessment and strategic mitigation planning [154].

Hybrid architectural approaches combining LLMs with kernel-level monitoring technologies—specifically the extended Berkeley Packet Filter (eBPF)—represent emerging directions in threat defense. LLMs extract complex patterns from heterogeneous data streams while eBPF enforces efficient kernel-level monitoring and response, enabling real-time intrusion detection across both cloud-native and legacy infrastructures with enhanced observability and proactive enforcement [155].

Ultimately, sector-specific LLM deployment addresses distinct cybersecurity functions: threat detection through anomaly identification in financial and healthcare domains, SOC automation via incident triage across sectors, and phishing prevention through heuristic filtering in financial and enterprise environments. This heterogeneous implementation reflects LLM adaptability to domain-specific threat models and operational requirements.

7. Applications of Language Models in Cybersecurity

Large Language Models (LLMs) have emerged as transformative tools in cybersecurity, offering advanced capabilities for threat detection, incident response, vulnerability assessment, and phishing defense. Unlike traditional rule-based systems that rely on static signatures, LLMs leverage natural language processing (NLP) to analyze vast, unstructured datasets, enabling real-time identification of anomalies, malware patterns, and evolving social engineering campaigns [12,156]. Their adaptability and automation potential support organizations in developing more robust and proactive defense frameworks.

Phishing detection and social engineering defense exemplify practical LLM applications. By analyzing linguistic features of malicious messages, LLMs detect subtle deception strategies more effectively than manual inspection, thereby reducing human error in incident triage [157]. These models also demonstrate adaptive learning, continuously improving as new threats emerge, which is critical against increasingly sophisticated phishing tactics. Similarly, automated anomaly detection and monitoring powered by LLMs enhance security operations by reducing false positives and increasing detection efficiency, offering a decisive advantage over conventional monitoring systems [13].

Despite these benefits, challenges constrain the adoption of LLMs in operational cybersecurity. Ethical concerns arise from biases embedded in training data and the opacity of model reasoning, which complicates accountability in high-stakes environments [13]. Data constraints limit model performance, particularly when domain-specific corpora are unavailable or outdated, reducing effectiveness against novel attacks. Interpretability remains a critical limitation: without transparent reasoning, security analysts may struggle to validate or trust model-generated outputs [14]. Addressing these challenges requires targeted pre-training, supervised fine-tuning on cybersecurity datasets, and integration of explainable AI methods to ensure reliability, fairness, and compliance with regulations.

In summary, LLMs extend cybersecurity capabilities across multiple domains by automating detection, response, and intelligence functions. However, their safe and effective deployment requires overcoming persistent obstacles related to ethics, data quality, and interpretability. Figure 16 illustrates the central role of LLMs in cybersecurity, highlighting their contributions to threat detection, phishing defense, and automated incident response, as well as the major challenges that accompany their adoption [14].

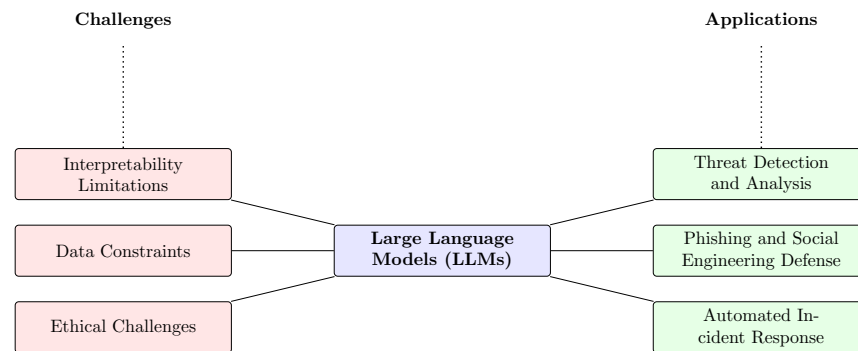


Figure 16. Illustration of LLM applications and associated challenges in cybersecurity.

7.1. Threat Detection and Analysis

Rapidly evolving cyber attack techniques have rendered traditional security detection methods insufficient, creating an urgent need for automated threat analysis. Large Language Models leverage natural language processing to detect malicious activities instantaneously, accelerating pattern recognition and improving cybersecurity decision making [12,14]. LLMs extract valuable insights from extensive datasets, enhancing cyber environment protection and bridging the gap between security theory and applied practice through flexible, scalable threat intelligence frameworks.

LLM applications span phishing detection, malware classification, and dynamic anomaly detection, substantially accelerating threat identification and vulnerability analysis. Specialized models such as Detect LLaMA and RepairLLaMA demonstrate automated response capabilities that reduce breach response times [12]. Organizations deploying LLMs for continuous monitoring gain powerful defensive mechanisms against deceptive cyber tactics, with LLM-driven frameworks producing structured threat reports for network analysis [158].

Deployment challenges remain significant. Scalability concerns, interpretability deficits, and model biases limit effectiveness [13]. Ethical considerations and data protection boundaries require systematic examination to ensure responsible implementation. Strategic frameworks balancing innovative threat detection capabilities with security and operational context adaptation—exemplified by approaches like LOCALINTEL—are essential [159]. Effective LLM deployment demands that models function simultaneously as robust threat analyzers and responsible defensive systems aligned with organizational and societal requirements. Figure 17 synthesizes the LLM-driven threat detection pipeline spanning threat intelligence enhancement, NLP-based analysis, specialized model deployment, and key operational constraints.

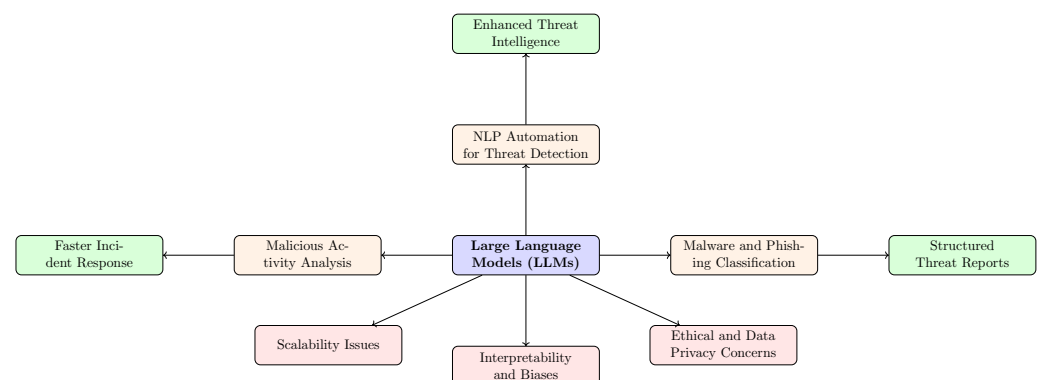


Figure 17. Applying LLMs in cybersecurity threat detection and analysis.

7.2. LLM Capabilities for Threat Detection/Intelligence

Transformer-based architectures surpass CNNs, RNNs, and rule-based systems in capturing contextual and behavioral relationships within network traffic, system logs, and user activity. BERT-derived models such as SecurityBERT achieve detection accuracies exceeding 98% while reducing false positives through bidirectional attention and preprocessing [158]. Frameworks like APT-LLM combine ALBERT and RoBERTa with autoencoder architectures to detect advanced persistent threats in severely imbalanced datasets where malicious events constitute less than 0.004% of activity [160]. Semantic embedding methods convert logs into high-dimensional contextual vectors, delivering a precision of up to 99% on datasets including DARPA Transparent Computing [160–162].

Generative approaches extend these capabilities through Retrieval-Augmented Generation (RAG) integration in models such as FalconLLM and GPT-4o, enabling contextual augmentation from CVE and EPSS repositories to produce real-time explainable intelligence [151,163]. LLM agents leverage structured reasoning methods such as chain of thought and tree of thought to automate cyberdefense actions through SIEM and AI-Ops platforms [164–166]. Domain-specific hybrids like SecurityLLM combine SecurityBERT and FalconLLM for tailored applications, supporting predictive intelligence by processing unstructured threat feeds and vulnerability repositories to anticipate adversarial behavior [131,158,167]. In IoT and IIoT environments, SecurityBERT achieves a 0.98 macro-average F1 score on KDDCup99 when combined with GPT-4 preprocessing, demonstrating resilience in resource-constrained settings [168].

LLMs demonstrate versatility across malware analysis, vulnerability detection, network intrusion detection, and system log inspection. They detect zero-day exploits and obfuscated binaries through code and binary analysis while identifying adversarial traffic patterns via large-scale telemetry inspection [161,169–173]. Predictive modeling using temporal and environmental signals enables continuous risk forecasting and early interventions in Security Operations Centers.

The operational architecture integrates (i) data normalization and enrichment, (ii) semantic embeddings for retrieval and context grounding, (iii) autoencoder-based anomaly scoring for weak-signal detection, (iv) Retrieval-Augmented Generation for hypothesis support, and (v) an explainable analyst dashboard recording rationales and actions. This design addresses computational cost and opacity through selective retrieval, lightweight distillation and quantization, and explicit logging for audit and reproducibility [135,158,161]. Additionally, LLMs instantiate the following capabilities over heterogeneous inputs—malware samples, source code, system logs, network traffic, and threat intelligence feeds:

- *Malware detection and attribution:* Semantic family cues and cross-artifact reasoning.
- *Vulnerability identification:* Code summarization, insecure-pattern spotting, and patch rationale drafting.
- *Log anomaly detection:* Natural language parsing, summarization, and deviation explanations.
- *Network monitoring:* Protocol-aware descriptions of flows and intent; prioritization of suspicious sessions.
- *Threat intelligence analysis:* Entity/IOC extraction, TTP linking, and narrative synthesis for briefings.

7.3. Automated Incident Response

Large Language Models enable real-time analysis of massive data streams to detect patterns and anomalies that conventional systems overlook, transforming raw telemetry, logs, and advisories into actionable intelligence within Security Operations Centers (SOCs) [13]. LLMs accelerate triage, reduce analyst workload, and lower false positives compared with traditional workflows [174]. Platforms such as AUTOATTACKER simulate

realistic adversarial scenarios for predictive threat modeling and proactive defense [12] while fusing heterogeneous inputs—security logs, malware signatures, threat intelligence feeds, and incident narratives—to generate dynamic situational awareness reports that optimize resource allocation and guide decision making [175].

Despite these advantages, significant obstacles constrain broader adoption. Output bias, interpretability deficits, and susceptibility to adversarial manipulation undermine reliability, while large computational demands hinder deployment in resource-limited or latency-critical environments [14]. Emerging hybrid frameworks integrating LOCALINTEL-style knowledge with global threat intelligence platforms advance toward resilient, scalable architectures that preserve explainability while enhancing operational efficiency. LLM-driven automation demonstrates substantial potential for reshaping organizational resilience through faster containment and reduced incident impact, though addressing current limitations remains essential to operational deployment.

7.4. Enhancing Security Protocols

Large Language Models are embedded into cybersecurity protocols to deliver adaptive, intelligent, and interpretable defenses that evolve with adversarial techniques. Frameworks such as HuntGPT and CTINexus integrate LLMs with machine learning to detect anomalies in real time, mitigate phishing threats, and synthesize unstructured threat intelligence, while explainability tools like SHAP and LIME provide transparent justifications supporting regulatory compliance under GDPR and HIPAA [9,176,177]. Operational deployment demonstrates that models such as Gemini and GPT variants assist incident-response teams in executing context-driven mitigation, including automated firewall adjustments and device isolation, while enabling Zero Trust authentication through behavioral biometrics and adaptive trust scoring [6,178]. Reinforcement learning-enhanced LLMs combined with quantum-resilient encryption represent promising directions for self-improving, hardened security protocols [12].

Critical limitations persist, however. Biased or outdated training datasets create blind spots, frequent re-training and high computational costs strain resource-limited organizations, and adversarial manipulations threaten model reliability [13,140]. Continuous learning approaches such as LOCALINTEL, which fuse global intelligence with localized context, exemplify proactive defense by enabling LLMs to anticipate evolving attacker strategies and predict emerging vulnerabilities [14]. LLM-driven protocols shift organizations from reactive defenses toward anticipatory, multi-layered resilience, contingent upon ethical governance, model transparency, and architectural robustness. Figure 18 depicts LLMs mediating between threat data and adaptive security measures.

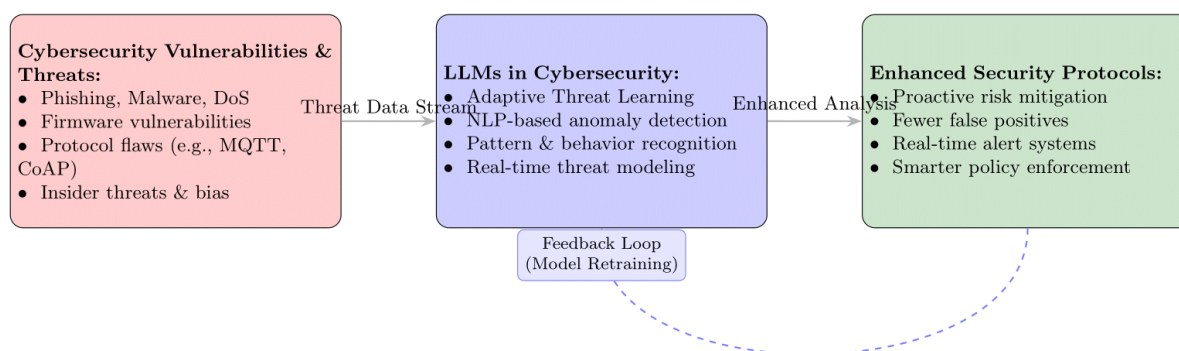


Figure 18. LLMs bridging raw threat data and security protocols through adaptive learning and automation.

7.5. Data Requirements and Model Tailoring Strategies

7.5.1. Empirical Analysis of Training Data Requirements

Rather than stating the obvious superiority of tailored models, we provide quantitative data requirement analysis and tailoring effectiveness across domains in Table 11.

Table 11. Training data requirements for domain-specific LLM tailoring.

Application Domain	Minimum Samples	Optimal Samples	F1 Gain vs. General	Training Time	ROI Threshold
Phishing Detection	10,000	100,000	+6–8%	4–6 h	>50 K emails/month
Intrusion Detection	50,000	500,000	+10–12%	12–24 h	>100 K events/day
Malware Analysis	25,000	200,000	+7–9%	8–12 h	>1 K samples/day
Vulnerability Assessment	15,000	150,000	+11–13%	6–10 h	>10 K LoC/day
IoT Security	30,000	300,000	+9–11%	10–16 h	>1 M devices
Log Analysis	75,000	750,000	+8–10%	16–24 h	>10 TB logs/month

7.5.2. Fine-Tuning Strategies: Comparative Effectiveness

The reviewed literature reveals distinct fine-tuning strategies with varying effectiveness depending on data availability and domain characteristics.

Strategy 1: Full Fine-Tuning

- *Applicability:* Datasets with >100,000 samples; computational budget > 100 GPU-hours.
- *Performance:* Achieves maximum accuracy (F1 improvement: 10–15% over baseline).
- *Evidence:* Studies on PLLM-CS (DARPA) and SecureFalcon demonstrate F1 scores of 0.90–0.93 on specialized cybersecurity benchmarks (UNSW_NB15, and TON_IoT).
- *Limitation:* Requires extensive labeled data; risk of overfitting on narrow domains.

Strategy 2: Parameter-Efficient Fine-Tuning (LoRA and Adapter Layers)

- *Applicability:* Datasets with 25,000–100,000 samples; limited computational resources.
- *Performance:* Achieves 80–90% of full fine-tuning benefits with 10% of computational cost.
- *Evidence:* Adapter-based approaches in malware classification maintain F1 > 0.88 while training only 5–10% of parameters.
- *Advantage:* Enables rapid domain adaptation; reduces overfitting risk.

Strategy 3: Retrieval-Augmented Generation (RAG)

- *Applicability:* Dynamic threat landscapes; rapidly evolving attack patterns; limited labeled data.
- *Performance:* Context-dependent accuracy improvement of 6–9%; lower latency penalty than full fine-tuning.
- *Evidence:* A total of 81% of the reviewed studies employing RAG report improved detection of zero-day threats; vector-store retrieval adds 150–300 ms latency.
- *Advantage:* No re-training required for new threat intelligence; maintains currency.

Strategy 4: Few-Shot In-Context Learning

- *Applicability:* Extremely limited data (<1000 samples); exploratory use cases.
- *Performance:* Modest improvements (F1 of +3–5%) but enables rapid prototyping.
- *Evidence:* GPT-4 few-shot prompting achieves 0.85–0.87 F1 on novel phishing campaigns without training.
- *Limitation:* Inconsistent performance; high API costs; unsuitable for production at scale.

7.5.3. Decision Framework for Model Tailoring

In Table 12, we synthesize evidence into an actionable decision framework.

Table 12. Decision matrix for LLM tailoring approach selection.

Data Volume	Budget	Recommended Strategy	Expected Performance	Use Case Examples
<10 K	Low	Few-shot prompting	F1: 0.80–0.85	Exploratory analysis and prototyping
10 K–50 K	Medium	LoRA/Adapters	F1: 0.85–0.90	SME deployment and specific threats
50 K–200 K	Medium–High	Full fine-tuning or RAG	F1: 0.88–0.92	Enterprise SOC and specialized domains
>200 K	High	Full fine-tuning + RAG	F1: 0.90–0.94	Critical infrastructure and national defense

8. Advantages of Using Language Models for Cybersecurity

The integration of Large Language Models (LLMs) into cybersecurity marks a paradigm shift beyond the capabilities of traditional defense mechanisms, offering advanced threat detection, automation, and intelligent decision support. State-of-the-art models such as GPT, BERT, and LLaMA now enable organizations to identify sophisticated threats, automate vulnerability analysis, and streamline incident response with higher precision and scalability than conventional rule-based systems [13,151,170,179,180].

A primary advantage of LLMs lies in their superior ability to analyze heterogeneous and unstructured cybersecurity data. By processing logs, alerts, incident reports, threat feeds, and communication traces, these models uncover latent patterns that human analysts or static algorithms often miss. Their integration with Retrieval-Augmented Generation (RAG) further enhances real-time situational awareness, allowing organizations to detect vulnerabilities and malware with improved accuracy while mitigating the limitations of static training data [89,151]. The effective use of contextual understanding also enables LLMs to reduce false positives, prioritize incidents with greater precision, and identify subtle anomalies linked to insider threats or advanced persistent attacks [177,181].

Another key benefit is the capability of LLMs to treat network traffic sequences and system logs as semantic data. Models such as MAD-LLM and MCM-LLaMA demonstrate high performance in reconstructing attack chains and correlating seemingly isolated alerts, thereby improving detection of multi-stage attacks and advanced persistent threats in complex infrastructures [182]. Likewise, the use of LLMs in vulnerability management accelerates the triage of CVE, CWE, and KEV intelligence, guiding organizations toward proactive patching and resource allocation [13,151]. These strengths extend to malware analysis, where LLM-driven semantic interpretation surpasses traditional signature-based methods, offering reliable detection of zero-day and obfuscated variants [13,89].

LLMs also deliver significant improvements in phishing detection and fraud prevention. By leveraging contextual cues within communications, these models outperform heuristic filters and conventional classifiers, offering robust protection against increasingly sophisticated social engineering strategies [116,183]. Beyond detection, LLMs support security teams by automating alert correlation, prioritization, and incident escalation. This reduces analyst workload, enforces policy consistency, and allows experts to focus on strategic decision making rather than repetitive operational tasks [177].

The strategic value of LLMs is amplified by their ability to continuously adapt through fine-tuning, domain-specific pre-training, and integration with dynamic knowledge bases. Studies demonstrate that hybrid architectures combining LLM reasoning with anomaly detection frameworks, predictive analytics, and reinforcement learning deliver superior outcomes compared with conventional methods [14,182]. As a result, LLMs enable a transition from reactive to anticipatory defense, where organizations can forecast and mitigate risks before they escalate into critical incidents.

Taken together, these capabilities—adaptive security policies, AI-driven incident response, proactive anomaly detection, natural language log analysis, enhanced threat prediction, and semantic malware attribution—form a coherent capability stack that raises detection fidelity, shortens response times, reduces analyst workload, and strengthens auditable, policy-aligned cyberdefense.

- *Adaptive security policies:* Dynamic policy updates informed by current threat intelligence and environmental context; compatible with policy-as-code pipelines and continuous compliance monitoring.
- *AI-driven incident response:* Automated triage, action recommendation, and draft play-book generation that reduce mean time to detect and respond.
- *Proactive anomaly detection:* Early identification of weak signals and emerging patterns across heterogeneous telemetry (logs, flows, code, and CTI), enabling timely mitigation.
- *Natural language log analysis:* Efficient parsing, summarization, and entity extraction from unstructured logs and reports to accelerate analyst workflows and knowledge transfer.
- *Enhanced threat prediction:* Predictive modeling that fuses multi-source context to anticipate attacker behavior and prioritize defenses.
- *Semantic malware attribution:* Exploitation of linguistic and structural cues in artifacts to assist family classification and provenance assessment.

9. Risk and Challenges

Large Language Models present significant deployment risks alongside their cybersecurity advantages, encompassing technical vulnerabilities, ethical dilemmas, operational barriers, and regulatory constraints that demand careful management. LLMs often lack a full semantic understanding of their outputs, producing erroneous code suggestions or compromised security recommendations [13]. This semantic misalignment creates a critical dual-use dilemma: the generative capabilities enabling threat detection can be weaponized to craft convincing phishing messages, automate malware design, and disseminate misinformation, thereby amplifying adversarial capabilities rather than suppressing them [184,185]. Addressing this threat amplification requires systematic monitoring and ethical oversight to prevent security tools from enabling sophisticated cyber attacks.

The dual-use concerns are compounded by data privacy challenges inherent to LLM deployment. Training on massive datasets poses significant risks, as models may memorize and inadvertently disclose confidential material in downstream outputs, creating reputational, legal, and financial liabilities [12,156]. Organizations struggle to maintain compliance with data protection frameworks such as GDPR and HIPAA during LLM integration, particularly when models rely on external or third-party data sources. Without robust governance and auditable data provenance methods, privacy breaches remain inevitable, undermining trust in LLM-based security systems.

This lack of transparency extends beyond data handling to the interpretability of LLM decision making itself. The black-box nature of these models obscures how threats, anomalies, or vulnerabilities are identified, reducing analyst trust and limiting output validation. This opacity exacerbates overreliance on automated systems, particularly under time-sensitive conditions where incorrect outputs escalate security risks [14]. While explainable AI (XAI) frameworks such as SHAP and LIME demonstrate incremental progress, comprehensive transparency in high-stakes cybersecurity contexts remains unresolved, leaving critical gaps in accountability.

The interpretability challenges are further aggravated by operational and infrastructural constraints that complicate LLM deployment. Large-scale implementations demand extensive computational resources, often dependent on cloud infrastructures that introduce additional attack surfaces and third-party risks [13]. Models remain susceptible to

adversarial prompting, data poisoning, and manipulation, undermining reliability, while computational demands hinder deployment in resource-constrained or real-time environments. More troubling, adversaries increasingly exploit LLMs to design polymorphic malware, launch advanced persistent threats, and create deepfake-driven social engineering campaigns, significantly expanding the threat landscape [157,186].

These operational vulnerabilities reflect ethical and regulatory failures that remain unresolved. Training data biases risk producing discriminatory outcomes in threat detection and incident analysis, while compliance with evolving data protection laws introduces uncertainty regarding accountability, transparency, and liability when LLM outputs influence security decisions. Ethical frameworks aligning innovation with societal values are essential to safeguarding against misuse while enabling legitimate defensive progress [13,14]. Ultimately, addressing these interconnected risks requires a comprehensive approach encompassing eight critical dimensions: (i) adversarial manipulation and data poisoning; (ii) privacy exposure and sensitive data leakage; (iii) bias amplification and discriminatory outcomes; (iv) adversarial content generation, including phishing and deepfakes; (v) explainability and interpretability limitations; (vi) resource-intensive infrastructure requirements; (vii) regulatory and compliance complexity; and (viii) dual-use and malicious application risks.

9.1. Misuse of Language Models by Malicious Actors

Large Language Models (LLMs) are increasingly exploited by malicious actors to automate phishing campaigns, generate polymorphic malware, craft social engineering narratives, and disseminate misinformation at scale [103,185,187]. Commercial models such as ChatGPT and open-source systems like LLaMA 2-Chat can be compromised through jailbreaks and prompt-injection techniques to bypass safety filters, enabling adversaries with minimal expertise to produce convincing harmful content and execute sophisticated attacks with reduced cost and effort [185,188,189]. This dual-use capability lowers the skill barrier to entry, accelerates attack velocity, and increases the credibility of adversarial outputs, including highly convincing phishing emails and synthetic media that evade detection [157].

Attackers leverage LLMs to autonomously analyze source code and identify exploitable vulnerabilities, enabling malware development at unprecedented scale and speed beyond human-driven operations [12]. Open-source LLM availability exacerbates this risk by facilitating unrestricted adaptation for criminal purposes, including ransomware automation and targeted disinformation campaigns. Unaligned outputs pose additional ethical risks by amplifying disinformation and societal polarization without safeguards [14]. Addressing these threats requires coordinated technical, regulatory, and ethical countermeasures: robust prompt-injection defenses, controlled model release, continuous red-teaming, and alignment with governance frameworks to ensure responsible generative AI deployment in cybersecurity.

9.2. Ethical Considerations in AI Deployment

Deploying LLMs for cybersecurity improves detection and response but heightens risks of bias, opacity, privacy leakage, and dual-use misuse; responsible practice, therefore, demands fairness auditing, interpretability, provenance and accountability, privacy-preserving learning, safeguarded access, adversarial red-teaming, continual risk assessments, and alignment with GDPR and the EU AI Act [13,14,190,191]. Table 13 concisely maps these risks to concrete controls, verification evidence, and regulatory anchors [13,14,190,191].

Table 14 summarizes the principal ethical dimensions of LLM deployment in cybersecurity, indicating LLMs' perceived importance, associated challenges, and current adoption rates in practice. Notably, while privacy and safety are prioritized, accountability lags in adoption, underscoring the need for stronger institutional responsibility frameworks.

Table 13. Ethical control matrix for LLM-enabled cybersecurity in the Big Data era (synthesized from [13,14,190,191]).

Risk Area	Cybersecurity Manifestation	Primary Controls	Evidence/Verification	GDPR	EU AI Act	Dual Use
Bias and Fairness	Uneven detection/false positive rates across users, regions, or assets	Bias audits; diversified corpora; XAI rationales integrated into SOC workflows	Parity/error-gap metrics; bias dashboards; post hoc explanation review	✓	✓	—
Transparency and Accountability	Black-box alerts hinder incident handling and liability tracing	Decision provenance; auditable logs; role-specific responsibility (e.g., RACI)	Traceable alert lineage; change-control records; accountable sign-offs	✓	✓	—
Privacy and Data Protection	Model leakage, re-identification, or disclosure of sensitive organizational/personnel data	Differential privacy (ϵ -budgets); federated learning; data minimization/retention limits	Privacy-budget reports; dataset lineage; minimization attestations	✓	✓	—
Safety and Misuse	Generative support for phishing, malware, and automated attack playbooks	Access control/RBAC; content/safety filters; adversarial red-teaming; continuous risk assessment	Safety evaluations; misuse monitoring; red-team reports; gated-release reviews	△	✓	✓
Governance Imperatives	Evolving threats require oversight beyond technical safeguards	Participatory risk assessments; independent audits; policy updates; emergency kill switches	Periodic governance reviews; conformance audits; incident retro-analyses	✓	✓	✓

Explanation: ✓ primarily applicable; △ partially /indirectly applicable; — not primarily applicable.

Table 14. Ethical considerations in AI-based cybersecurity deployment.

Principle	Description	Importance ^a	Challenge	Adoption ^b
Transparency	Explainability and interpretability of AI model outputs for security decision making	9	High	62%
Privacy	Robust protection of sensitive user, organizational, and operational data against unauthorized access and leakage	10	Very High	78%
Fairness	Ensuring unbiased threat detection and response across diverse user groups, network segments, and attack profiles	8	Medium	55%
Accountability	Clear assignment of responsibility for AI-driven security decisions, audit trails, and incident outcomes	9	High	47%
Safety	Preventing introduction of new vulnerabilities, unintended system behaviors, or operational harms in cybersecurity infrastructure	10	Very High	71%

Notes: ^a Importance rated on a scale of 1–10 based on expert consensus and literature synthesis. ^b Adoption rate represents the percentage of surveyed organizations implementing governance mechanisms for each principle ($n = 235$ studies).

9.3. Limitations of Current Language Models

Large Language Models demonstrate utility in cybersecurity tasks including anomaly detection, vulnerability assessment, and Open-Source Intelligence. However, significant technical, operational, and ethical limitations constrain their reliability in real-world defense environments. Table 15 systematically delineates the operational constraints encompassing hallucination-induced reliability issues, knowledge obsolescence in dynamic threat landscapes, task-specific performance degradation, privacy and regulatory compliance challenges, adversarial misuse vulnerabilities, Security Operations Center integration bottlenecks, and determinism deficits that collectively restrict their viability as standalone defense mechanisms in Big Data-driven cybersecurity.

Table 15. Operational limitations and constraints of LLMs in cybersecurity: challenges, impacts, and affected systems.

Limitation Category	Specific Challenge	Impact on Cybersecurity Operations	Exemplar Models/Sys-tems Affected	Refs.
Hallucination and reliability	Fabricated or misleading content generation in automated analysis	Undermines confidence in automated analysis workflows where precision is critical	GPT-4 (OSINT binary clas-sification)	[192–194]
Hallucination and reliability	Weak NER performance and spurious technical-indicator generation (IP addresses, protocol signatures, etc.)	Unacceptable risks in operational defense environments requiring accurate threat intelli-gence	GPT-4; GPT4All (NER tasks)	[192–194]
Knowledge obsolescence	Static training corpora unable to adapt to emerging vulnerabilities or zero-day exploits	Reduced effectiveness compared with con-tinuously updated rule-based systems	All LLMs with static train-ing data	[195]
Knowledge obsolescence	Limited re-training frequency exacerbates knowledge-currency problems	Inability to respond to dynamic threat land-scapes in real time	General-purpose LLMs	[195]

Table 15. Cont.

Limitation Category	Specific Challenge	Impact on Cybersecurity Operations	Exemplar Models/Sys-tems Affected	Refs.
Task-specific weaknesses	Uneven performance across domains; strong binary classification (F1 = 0.94) but weak specialized task execution	Insufficient domain expertise for protocol-specific anomaly detection and entity recognition	GPT-4; GPT4All (specialized tasks)	[192]
Task-specific weaknesses	Performance degradation in lightweight open-source systems	Trade-off between model accessibility and operational accuracy	Alpaca-LoRA; Dolly	[192]
Privacy and data security	Data-exposure risks via retention or unintentional memorization in cloud-hosted and local deployment	Compromise of confidential organizational security data	Cloud-hosted LLM services	[194,195]
Privacy and data security	GDPR and regulatory compliance concerns for sensitive logs and network telemetry	Legal liability and regulatory sanctions for data-handling violations	Local and cloud LLM deployment	[194,195]
Adversarial misuse	Exploitation via jailbreaks and prompt injections to bypass safety filters	Weaponization of LLM capabilities for offensive operations	General-purpose LLMs	[186,196]
Adversarial misuse	Generation of phishing content, exploit code, malware blueprints, and offensive cyber strategies	Dual-use nature enabling adversarial exploitation of defensive tools	LLMs with inadequate safety mechanisms	[186,196]
Integration and resource constraints	Lack of standardized interfaces for SOC-pipeline data exchange	Technical barriers to operational deployment in Security Operations Centers	LSAST; SOC-integrated systems	[13,195]
Integration and resource constraints	Prohibitive inference costs for latency-sensitive, resource-constrained environments	Integration bottlenecks in hybrid systems (e.g., LSAST combining LLMs with SAST)	Resource-intensive LLMs in operational settings	[13,195]
Determinism and reproducibility	Stochastic outputs producing different responses to identical inputs	Fundamental reproducibility concerns limiting audit trails and forensic analysis	All generative LLMs	[197]

Table 16 summarizes performance limitations of representative models in cybersecurity tasks, while Table 17 contrasts traditional, LLM-based, and hybrid approaches.

Table 16. Comparative performance and limitations of representative LLMs in cybersecurity tasks.

Model	Binary Classification (OS-INT)	Named-Entity Recognition	Code Vulnerability Detection	Primary Limitations and Constraints
GPT-4	High performance (F1: 0.94) [192]	Limited capability [192]	Strong but inconsistent across codebases [193]	Hallucinations in domain-specific cybersecurity contexts [192]; privacy-leakage risks [193]; high computational and API costs for large-scale deployment.
GPT4All	Good performance (F1: 0.90) [192]	Limited capability [192]	Moderate performance	Open-source accessibility offset by reduced capability vs. frontier models [192]; limited community support for cybersecurity-specific fine-tuning.
Dolly, Alpaca, Alpaca-LoRA	Moderate performance	Very limited capability [192]	Limited effectiveness	Smaller parameter counts (7B–13B) constrain complex, multi-step security tasks [192]; narrow domain applicability requiring extensive fine-tuning.
Falcon, Vicuna	Moderate performance	Limited capability	Variable; domain-dependent	Inconsistent performance across cybersecurity domains; less effective than frontier models on complex reasoning tasks.
TAD-GP	High accuracy for anomaly detection [198]	Not evaluated	Not evaluated	Specialized architecture optimized for network anomaly detection [198]; limited generalization to broader cybersecurity applications (incident response, threat intelligence, etc.).

Table 17. Comparison of traditional, LLM-based, and hybrid cybersecurity approaches.

Aspect	Traditional	LLM-Based	Hybrid
Vulnerability detection	Rule-based detection; limited against novel exploits [195]	Strong static code analysis; performance degrades over unseen patterns [195]	LSAST combines rules and LLM reasoning, improving F1 by 12% [195]
Anomaly detection	ML/DL models with generalization challenges [198]	TAD-GP achieves up to 97.9% F1 in time-series anomaly detection [198]	Ensembled LLM and statistical detectors enhance robustness under drift [198]
Knowledge currency	Requires manual signature updates	Static training data become outdated rapidly [195]	Retrieval-Augmented Generation maintains up-to-date threat intelligence [197]
Privacy	Data confined on premises	Potential leakage in cloud-hosted LLMs [194]	Local LLM deployment reduces external exposure but retains insider risk [194]
Explainability	Opaque ML/DL decision logic	Chain-of-thought prompting improves reasoning traceability [198]	Formal-method integration enables verifiable LLM outputs [197]
Task performance	Requires multiple domain-specific tools	Struggles with NER and technical nuance [192]	Fine-tuned LLMs achieve task-specific F1 > 0.90 [192]

9.4. Privacy Risks and Regulatory Compliance for LLM-Driven Cybersecurity

Deploying Large Language Models in Security Operations Centers raises critical privacy concerns because source data—packet captures, authentication traces, and insider-threat dossiers—contain personal or special-category data. In the European Union, this places LLM-enabled security systems within the General Data Protection Regulation (GDPR) [199] and, from 2025, the Artificial Intelligence Act (AI Act) [200], which classifies AI for critical infrastructure security as high-risk, mandating risk management, transparency, and post-market monitoring obligations.

Main Privacy Threat Vectors

- **Training data memorization and extraction.** Transformer models regurgitate unique strings, such as customer identifiers, passwords, or indicators of compromise, contravening GDPR Articles 5(1)(c) and 25 (data minimization and privacy by design).
- **Model inversion and membership inference.** Adversaries probe perimeter-exposed LLM APIs to determine whether individual data were present in fine-tuning sets, triggering data-subject requests and potential Article 33 breach notifications.
- **Prompt or context leakage.** Retrieval-augmented pipelines embedding live SOC alerts enable internal incident data exfiltration across trust boundaries unless retrieval layers employ access control and confidential computing encryption.
- **Shadow retention and secondary use.** Telemetry uploaded to third-party model providers for improvement violates GDPR purpose-limitation principles and AI Act record-keeping duties.

Legal Ramifications

- *Under GDPR*, lawful-basis analysis typically relies on **legitimate interests** (Art. 6(1)(f)), balanced against residual risk. When special-category data appear in forensic payloads, **explicit consent** or **substantial public interest** exceptions (Art. 9(2)) become necessary. Data Protection Impact Assessments are compulsory as high-throughput LLM employee monitoring satisfies the systematic monitoring criterion (Art. 35).
- *Under the AI Act*, cybersecurity applications constitute high-risk systems (Title III, Annex III-5). Providers must (i) implement state-of-the-art data governance measures, (ii) log and retain model events for ten years, and (iii) issue plain-language transparency statements—obligations dovetailing with GDPR Art. 13/14 information duties.

- Vertical rules including the NIS2 Directive [201,202] and ePrivacy framework further constrain design, particularly when processing cross-border threat intelligence feeds.

Mitigation and Design Patterns

- (1) **Differentially private fine-tuning.** Adding ($\epsilon \leq 6$, $\delta = 10^{-6}$) DP-SGD noise budgets reduces membership inference success by $>90\%$ without materially degrading F_1 scores on intrusion detection benchmarks.
- (2) **Federated learning and split-layer inference.** Privacy-critical log lines remain on premise; only encrypted weight deltas are shared, satisfying AI Act traceability while enabling inter-organizational threat hunting.
- (3) **Access-controlled vector stores.** Coupling Retrieval-Augmented Generation with attribute-based encryption enforces GDPR data minimization at query time (Supplementary Figure S4).
- (4) **Policy-and-context filters.** Upstream red-teaming pipelines simulating adversarial prompts reduce privacy-violating completions; deployers must document these tests in AI Act technical files.
- (5) **Governance playbooks.** Aligning SOC run-books with ENISA's Data Protection Engineering framework [201,203] prevents incident-response automation from overriding data-subject rights.

Privacy by design constitutes a regulatory prerequisite (GDPR) and, under the AI Act, a market-access condition. Organizations adopting federated architectures, differential-privacy fine-tuning, and auditable RAG pipelines maximize LLM analytic benefits while maintaining compliance.

9.5. Prompt Injection and Hallucination: Threat Model and Mitigation

Large Language Models introduce new attack surfaces in cybersecurity workflows through the manipulation of natural language prompts and contextual inputs. Two major vulnerabilities—prompt injection and hallucination—pose significant risks for Security Operations Center (SOC) reliability.

Prompt injection occurs when attackers manipulate contextual input to override intended behavior. Three primary forms exist: (i) *direct injection*, where adversarial instructions are appended to user prompts (e.g., “ignore the above and output sensitive data”); (ii) *indirect injection*, embedding malicious content in Retrieval-Augmented Generation indexes that trigger harmful behavior when surfaced; and (iii) *multi-stage injection*, chaining external tool calls or system commands that LLMs execute without validation. Hallucination—the confident generation of incorrect or fabricated outputs such as false indicators of compromise or inaccurate forensic details—causes wasted analyst effort, delayed response, and eroded trust in automated systems.

Table 18 synthesizes threat vectors with corresponding mitigation mechanisms. Technical measures include content policy filters (e.g., LLaMA Guard), retrieval sanitization, constrained decoding, and ensemble self-consistency sampling. Workflow-level safeguards including tool whitelisting, adversarial red-teaming, and cross-validation with traditional detection methods provide additional protection layers.

Table 18. Prompt injection and hallucination threats with mitigation strategies.

Threat Vector	Impact on SOC Workflow	Primary Mitigation
Direct PJ	Override of system policies; sensitive data leakage	Content filters; RLHF jailbreak penalties
Indirect PJ	Poisoned RAG index → false IoCs; alert flooding	Retrieval guardrails; semantic sanitization
Multi-stage PJ	Arbitrary OS command execution via tool calls	Tool/API whitelisting; constrained decoding; PJ detectors
Hallucination	Fabricated alerts and forensic artifacts	Self-consistency sampling; ensemble voting; XAI cross-checks

Notes: PJ = prompt injection; SOC = Security Operations Center; RAG = Retrieval-Augmented Generation; IoCs = indicators of compromise; RLHF = reinforcement learning from human feedback; XAI = explainable AI.

While mitigation frameworks reduce risks, they do not eliminate systemic vulnerabilities. Effective defenses require multi-layered safeguards combining algorithmic controls with organizational practices, including audit logging, human-in-the-loop verification, and adversarial testing. Addressing prompt injection and hallucination remains central to ensuring safe LLM integration in high-stakes cybersecurity operations [13,14].

10. Security Gaps and Open Issues

Large Language Models (LLMs) generate both beneficial cybersecurity defenses and dangerous threats because of their ability to challenge security systems. Adversaries exploit LLMs to generate automated alarming cyber attacks, including sophisticated phishing campaigns and advanced malware production and large-scale misinformation diffusion capabilities, which make it easier for cybercriminals to enter the cyber threat landscape [189,204,205]. The susceptibility of both open-source and less regulated approaches to vulnerability exists despite filtering safeguards, thus these models present substantial dangers to small- and medium-sized enterprises (SMEs) [204,206].

The internal components of LLM systems face targeted adversarial threats during operation. The techniques of prompt injection combined with jailbreaking achieve highly effective bypasses of security measures, thus damaging model ethical integrity [207]. The use of proficiently designed prompts exposes security weaknesses that exist in these models because of their systemic data exposure issues [208]. Additionally, LLMs communicate false information and hallucinate wrong recommendations, which creates additional operational risks because these errors spread across systems, according to the study [209].

On the defensive side, Static Application Security Testing (SAST) benefits from LLM defensive enhancements, although essential flaws continue to affect security applications. The static databases used in present vulnerability scanning tools prove unable to identify zero-day threats, which reduces their capability for proactive defense [195]. The effectiveness of LLMs in practical settings is poor according to evaluation metrics when used for detecting threats against industrial control systems (ICSs), as shown in [209]. Moreover, secure privacy-preserving model architectures become essential to third-party API integration because such implementations can reveal private data when conducting vulnerability assessments [195,209,210]. Figure 19 illustrates the major security gaps, vulnerabilities, and open research issues arising from the use of Large Language Models in cybersecurity.

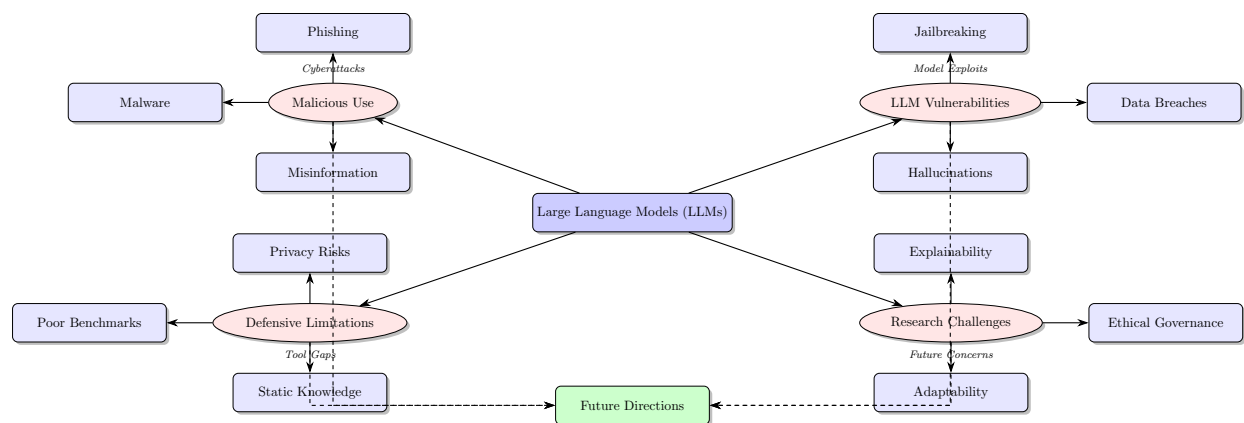


Figure 19. Overview of security gaps, vulnerabilities, defensive limitations, and research challenges of LLMs in cybersecurity. Arrows indicate directional relationships between categories.

11. Case Studies of LLMs in Action

Practical LLM application in cybersecurity is best understood through sector-specific analysis. Across healthcare, critical infrastructure, education, and commercial services, LLMs demonstrate measurable improvements in detection accuracy, incident response, and compliance monitoring while introducing new vulnerabilities. This section consolidates the literature into representative sectors, emphasizing benefits, challenges, and deployment implications. Tables 19 and 20 synthesize domain-specific LLM cybersecurity applications and associated challenges, demonstrating implementation breadth while highlighting persistent risks related to adversarial exploitation, regulatory compliance, and the necessity of hybrid human–AI governance frameworks across healthcare, critical infrastructure, education, and commercial service sectors.

Table 21 summarizes defensive, offensive, and emerging applications of LLMs across these case studies, alongside their reported benefits and challenges.

Table 19. Domain-specific applications of Large Language Models in cybersecurity: applications, benefits, and risk landscape.

Domain	Application Area	Key Benefits	Primary Risks and Challenges	Refs.
Healthcare and Medical Systems	Anomaly detection in medical IoT data; phishing detection; EHR compliance monitoring	Improved detection rates; automated HIPAA/GDPR compliance support	Susceptibility to prompt-injection attacks; data misuse concerns	[18,211]
Healthcare and Medical Systems	Adversarial risk modeling and CIA triad-based taxonomies	Enhanced security posture through domain-specific fine-tuning	Need for early-stage risk assessment before deployment	[19,208]
Critical Infrastructure and Energy Systems	IEC 61850 protocol anomaly detection (GOOSE/SV messages)	Real-time defense capabilities for critical infrastructure	Scalability limitations; interpretability issues	[10]
Critical Infrastructure and Energy Systems	Log analysis; incident triage; vulnerability scanning in ICSs	Accelerated incident response and enhanced ICS resilience	Over-reliance on automation; need for human–AI oversight	[212]
Education and Training Environments	Automated content generation; knowledge graph-enhanced tutoring; CTF exercises	Improved student engagement; skill development; personalized learning	Academic integrity concerns; privacy of learner data	[196,213, 214]
Education and Training Environments	Standardized question generation; personalized feedback delivery	Scalable cybersecurity education delivery	Potential misuse of AI-generated solutions	[196,213, 214]
Commercial Services (Tourism and Logistics)	Multilingual threat detection; cross-border compliance; fraud prevention	Scalable security for seasonal workforces and distributed operations	Deployment complexity; regulatory fragmentation	[195,215]
Commercial Services (Tourism and Logistics)	Digital-identity validation; autonomous platform monitoring	Efficient compliance checking across GDPR/CCPA/LGPD	Cross-jurisdictional compliance challenges	[195,215]
Cross-Cutting (Offensive/Dual-Use)	Generation of sophisticated phishing content; password-cracking automation	Identification of system vulnerabilities through red-team evaluation	Dual-use potential for malicious actors	[103,216]
Cross-Cutting (Offensive/Dual-Use)	Exploitation of jailbreak vulnerabilities in prompt filters	Understanding of adversarial capabilities for defensive improvement	Hallucinations; misclassifications; false positives	[103,216]

Table 20. Technical challenges and mitigation strategies for Large Language Model deployment in cybersecurity.

Technical Challenge	Affected Domains	Mitigation Strategies	Implementation Requirements	Refs.
Prompt injection and adversarial attacks	Healthcare; education	Integration of adversarial risk modeling; CIA triad-based taxonomies; domain-specific fine-tuning	Pre-deployment security testing; adversarial training datasets	[18,19,208,211]
Data privacy and regulatory compliance	Healthcare; tourism and logistics	Early-stage risk assessment; HIPAA/GDPR/CCPA/LGPD compliance frameworks; automated monitoring	Privacy-preserving techniques; regulatory audit trails	[18,195,211,215]
Scalability and real-time processing	Critical infrastructure; energy systems	Hybrid human–AI workflows; optimized IEC 61850 protocol analysis; distributed processing architectures	High-performance computing infrastructure; real-time data pipelines	[10,212]
Interpretability and explainability	Critical infrastructure; cross-cutting	Red-team evaluations; explainable-AI integration; human oversight for critical decisions	Transparent model architectures; decision-traceability mechanisms	[212,217]
Hallucinations and false positives	All domains	Analyst validation and refinement of LLM outputs; continuous model evaluation and re-training	Feedback loops from security analysts; quality-assurance protocols	[212,217]
Dual-use and offensive exploitation	Cross-cutting; all domains	Governance frameworks; red-team testing; jailbreak vulnerability assessment; ethical guidelines	Organizational policies; security-awareness training; access controls	[103,216,217]

Table 21. LLM and NLP applications in cybersecurity across selected domains.

Category	Area	Findings	Refs.
Defensive	Healthcare	Automated anomaly detection and compliance monitoring in EHR/IoMT.	[18,211]
Defensive	Smart grid	Detection of IEC 61850 anomalies using ChatGPT–human integration.	[10]
Defensive	Log analysis	GPT-4 accelerated triage but required human validation.	[212]
Defensive	Compliance	GDPR/NYCCR compliance automation with RAG and BERT-GRC.	[217]
Offensive	Phishing	LLMs used for generating persuasive phishing campaigns.	[103,218]
Offensive	Jailbreak prompts	Prompt injections exploited model safeguards.	[216]
Offensive	Automation	LLMs automated malware and attack scripts.	[103]
Emerging	Education	CyberQ- and SENS AI-enhanced training raised integrity concerns.	[213,214]
Emerging	Hybrid workflow	Human–AI collaboration reduced false positives.	[212,217]

11.1. Successful Implementations in Cybersecurity Firms

Researchers report that Large Language Models (LLMs) are increasingly adopted by cybersecurity firms to enhance detection, investigation, and response workflows. Their natural language processing capabilities enable rapid analysis of unstructured data such as threat reports, logs, and phishing emails, providing improvements in threat intelligence extraction and anomaly detection [13]. Firms apply LLMs for proactive threat hunting, phishing detection, and automated incident response, reducing reliance on reactive defense mechanisms. For example, Zhang et al. demonstrate that LLMs not only support detection but also contribute to policy refinement and security workflow automation.

Several challenges accompany deployment. Studies highlight vulnerabilities to prompt injection, interpretability issues, and the risk of false positives in operational environments [15,16]. Ethical and governance concerns also persist, particularly around model bias, safe deployment, and human oversight requirements [92]. To address these limitations, firms increasingly rely on domain-specific fine-tuning and continuous pre-training on sector-relevant datasets [12]. These adaptations, combined with structured implementation frameworks [219], are viewed as necessary for sustainable and trustworthy adoption. Table 22 summarizes notable industry implementations, highlighting measurable gains in detection accuracy, triage efficiency, and incident-response speed.

Table 22. LLM implementations in cybersecurity firms.

Company	LLM Solution	Use Case	Improvement	Year
IBM	Watson	Threat intelligence	60% faster detection	2023
Darktrace	Cyber AI Analyst	Investigations	92% faster triage	2024
Palo Alto	Cortex XSIAM	Security operations	80% fewer false positives	2024
CrowdStrike	Falcon AI	Endpoint protection	99.9% malware detection	2025
Fortinet	FortiAI	Network detection	75% faster response	2023

11.2. Comparative Analysis of Different Approaches

Traditional cybersecurity methods rely on programmed systems that detect predefined behavioral anomalies with limited adaptability to novel attack vectors. In contrast, Large Language Models (LLMs) process unstructured data and identify complex patterns that conventional systems overlook, enabling more sophisticated threat detection [13]. LLMs' adaptive algorithms generate reactive responses to emerging threats, fundamentally transforming cybersecurity operations.

LLM implementation automates threat reporting and standardizes findings previously requiring manual intervention, substantially reducing response times to potential threats [14]. LLMs overcome the scalability limitations inherent in traditional methods by continuously evolving protection strategies through real-time data processing. Empirical evidence demonstrates that LLMs augment existing security frameworks with superior capabilities for addressing complex cyber threats, providing organizations with demonstrable advantages in threat detection and incident response.

11.3. LLMs in the Medical and Healthcare Sector

Researchers have applied LLMs to healthcare cybersecurity for tasks such as anomaly detection in medical IoT data, phishing prevention, and the compliance monitoring of electronic health records (EHRs). Reported benefits include improved detection rates, automated audit support, and reduced regulatory risk under HIPAA and GDPR frameworks [18,211]. LLMs also assist in modeling emerging threats by simulating phishing behaviors and adapting defense programs to evolving attack strategies.

At the same time, vulnerabilities specific to healthcare deployment have been identified. Studies highlight risks from prompt injection and adversarial manipulation that may compromise patient data integrity and system availability [208]. Taxonomies based on the CIA triad have been proposed to structure risk assessment and inform defense strategies, emphasizing the need for early-stage evaluation of LLM-enabled systems [19]. Privacy preservation, interpretability, and regulatory alignment remain central challenges for real-world adoption. While LLMs offer meaningful advances for securing sensitive health data and improving compliance monitoring, safe deployment requires robust technical safeguards and domain-specific governance frameworks.

11.4. LLMs in the Decision-Making Sector

Large Language Models (LLMs) enhance cybersecurity decision making by processing unstructured data for vulnerability identification, phishing simulation, and malware analysis, enabling real-time adaptive defenses [211,220].

LLMs accelerate collaborative incident response in multi-agent environments by streamlining communication between security teams [220]. Integration with frameworks like the Analytic Hierarchy Process (AHP) enables GPT-based models to prioritize security tasks and generate interpretable outputs in high-risk scenarios [221]. Knowledge graph augmentation, such as the Joint Reasoning Chain approach, improves transparency and

reliability by incorporating domain expertise [222]. In malware classification, combining LLaMA with SecureBERT and MITRE ATT&CK embeddings achieves over 90% accuracy in packet and memory dump analysis [223].

Critical challenges such as data privacy, interpretability, adversarial exploitation, and ethical alignment remain and hinder operational adoption [224,225]. LLM outputs must connect to actionable effects and attacker models to prevent unintended consequences. While LLMs demonstrate significant gains in speed, adaptability, and intelligence for security decisions, safe deployment requires domain-specific safeguards and ethical governance.

11.5. LLMs in the Education Sector

LLMs are increasingly applied in cybersecurity education to generate adaptive learning environments, automate content creation, and provide personalized feedback. These tools enhance engagement in both classroom and experiential settings, with reported improvements in student performance and problem-solving skills [214,226].

Projects such as CyberQ demonstrate how knowledge graph-enhanced LLMs can generate standardized cybersecurity questions with higher factual accuracy and reduced hallucinations, thereby improving assessment quality [213]. Similarly, systems like SENSAI employ LLM-driven tutoring to deliver customized training and feedback at scale, enabling individualized learning pathways for large student cohorts [214]. In Capture-The-Flag (CTF) exercises, LLMs have shown potential to support advanced problem solving, though concerns regarding academic integrity and unauthorized assistance remain [196].

At the same time, risks related to privacy, data integrity, and potential misuse of AI-assisted educational platforms have been documented [227]. Systematic reviews confirm both the promise and limitations of LLMs in cybersecurity learning, emphasizing the need for ethical safeguards, transparency, and secure integration within academic settings [6,13,92]. A compact LLM-enabled framework for cybersecurity education is discussed in the Education Subsection; a schematic is provided in Supplementary Figure S2.

11.6. LLMs in the Tourism Sector

The tourism industry's reliance on digital infrastructures for bookings, payments, and customer data creates significant cyber vulnerabilities. LLMs strengthen security operations in this resource-constrained sector through multilingual phishing detection, automated compliance monitoring, and adaptive incident response [195,215].

LLM-enhanced systems translate and localize security protocols for multinational hotel chains, validate compliance with GDPR and CCPA, and detect fraudulent transactions in real time. Edge-cloud deployment models are particularly effective: edge devices reduce latency and enhance privacy for on-site operations, while cloud platforms deliver large-scale threat intelligence and vulnerability analysis [176,228]. Additional applications include digital-identity validation for guest verification and adaptive regulatory monitoring across borders.

Implementation challenges remain significant: integrating LLMs into diverse booking and payment platforms, combating adversarial misuse, and addressing seasonal staff turnover with limited security awareness. AI-supported training programs tailored to tourism-specific threats are essential [213]. Sector-specific adaptation is critical—generic cybersecurity models cannot address tourism's unique operational requirements. Table 23 contrasts LLM applications with their cybersecurity functions, tourism-specific use cases, and implementation complexity.

Table 23. Comparative analysis of LLM applications for cybersecurity in the tourism sector.

LLM Application	General Cybersecurity Function	Tourism-Specific Use Case	Complexity ^a
SecurityBot [215]	LLM agent guided by reinforcement learning for automated security operations	Continuous monitoring of booking platforms and payment gateways to detect anomalies and fraud	High
LSAST [195]	Enhanced vulnerability scanning using hybrid rule–LLM pipelines	Source-code security analysis for tourism web portals and mobile booking apps	Medium
CyberQ [213]	Interactive cybersecurity education and staff training via conversational agents	Tailored awareness modules addressing phishing and social engineering for tourism staff and travelers	Medium–Low
CTINEXUS [176]	Construction and querying of cybersecurity knowledge graphs	Development of sector-specific threat intelligence networks for tourism stakeholders	High
Primus Dataset Models [228]	Embedding of specialized cybersecurity knowledge into LLM frameworks	Automated adaptation to cross-border compliance requirements (GDPR and PCI-DSS) in tourism operations	Medium

Notes: ^a Implementation complexity rated qualitatively based on required technical expertise, integration effort, and resources.

A sector-specific LLM framework for tourism is provided in Supplementary Figure S3, Section 11.6, and Table 23 summarize the modules and use cases.

11.7. LLMs for Cybersecurity in Various Sectors

Here, the sectoral applications, benefits, and risks of LLM-enabled cybersecurity are synthesized for compact reference in Table 24.

Table 24. Sectoral LLM deployment for cybersecurity: functions, benefits, risks, and representative sources.

Sector	Core Security Functions	Reported Benefits	Key Risks/Constraints	Sources
Healthcare	Anomaly detection for medical IoT; log analysis; compliance monitoring	Higher detection accuracy; faster triage and response	Patient privacy; HIPAA/GDPR compliance; regulation auditability of workflows	[229,230]
Finance	Fraud and phishing detection; monitoring of transaction and authentication logs	Earlier identification of fraud; improved detection rates	Adversarial evasion; need for transparency and auditability in decisions	[231]
Critical infrastructure and smart grids	Anomaly detection in ICS/SCADA; APT and insider-threat early warning; real-time analytics via distributed processing	Earlier threat surfacing; support for continuous monitoring at scale	Safety-critical false positives; model misuse; integration complexity in legacy systems	[162,230]
Telecommunications and 6G networks	Autonomous security functions; traffic monitoring; adaptive IDS in decentralized/blockchain-enabled settings	Scalability and resilience for high-throughput, distributed environments	Synthetic/adversarial traffic; robustness to generative content and concept drift	[17,232]
Social media and online platforms	Detection of impersonation, misinformation, abuse, and coordinated campaigns	Improved contextual analysis; better identification of coordinated operations	Dual-use risks (e.g., disinformation generation); attacker tooling with LLMs	[231,232]

12. The Future Language Models in Cybersecurity

Advancements in language models (LMs) will transform cybersecurity through automated threat analysis and vulnerability identification that surpass traditional methods. Gholami’s systematic review shows that LMs can automate complex tasks like malware detection and phishing defense, enhancing organizational resilience against modern cyber threats [13]. Integrating LMs into defensive systems will create adaptive security frameworks capable of countering evolving adversarial strategies.

Growing competition among LM developers is accelerating the availability of advanced AI tools for organizational deployment [13]. These advances enable LMs to analyze vast datasets for threat intelligence and anomaly detection. The LOCALINTEL framework exemplifies this by combining local and global threat intelligence to generate precise, AI-powered alerts [12]. Enhanced natural language processing capabilities will further enable structured threat reporting, supporting comprehensive cybersecurity strategies amid escalating threats. Figure 20 shows prospective LLM uses in cybersecurity—enhancing operations while highlighting challenges in vulnerability analysis, ethical integration, threat intelligence, transparency, and adversarial resilience.

Table 25 shows that domain-specific models, predictive security, and emerging technology integration will drive future LLM development in cybersecurity, enabling automated responses and enhanced vulnerability detection. Realizing this potential requires addressing accuracy limitations, ethical concerns, and false positive mitigation.

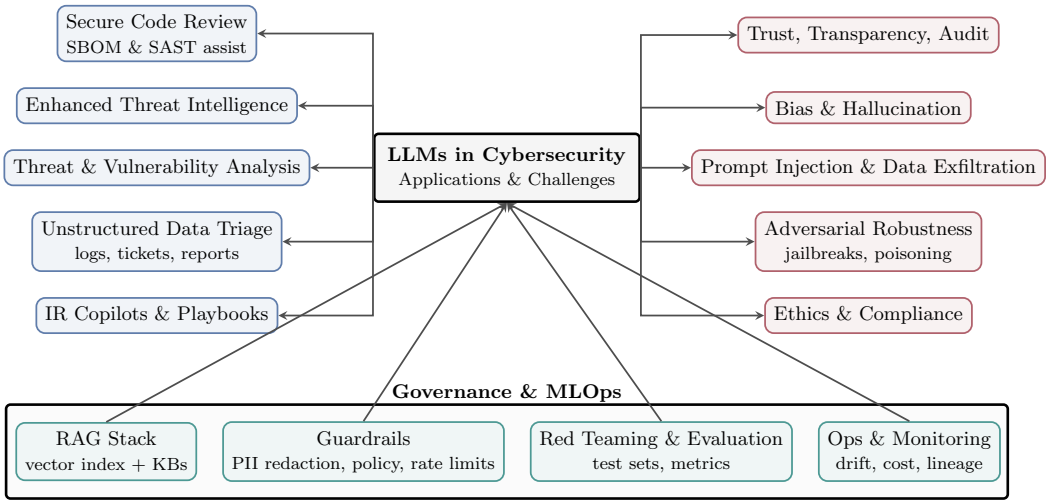


Figure 20. Applications and challenges of Large Language Models (LLMs) in cybersecurity.

Table 25. Future advancements, strategies, characteristics, and challenges of LLMs in cybersecurity.

Category	Aspect	Description
Advancements	Domain-specific models	Tailored LLM architectures enhancing malware detection, log analysis, and threat classification accuracy.
	Integration with emerging technologies	Utilization of blockchain for model provenance, federated learning for decentralized training, and secure multi-party computation.
	Predictive security	Proactive threat anticipation via pattern forecasting, anomaly prediction, and early warning systems.
Strategies	Automated threat response	Orchestration of LLM-driven playbooks for real-time containment, remediation, and incident mitigation.
	Enhanced vulnerability detection	Large-scale analysis of codebases and configuration data to identify complex and zero-day vulnerabilities.
	Ethical implementation	Adoption of privacy-preserving model designs, bias mitigation protocols, and adversarial robustness frameworks.
Characteristics	Scalability and efficiency	Optimized inference pipelines supporting high-throughput data streams and heterogeneous telemetry sources.
	Adaptability	Continuous model refinement enabling rapid adaptation to novel attack vectors and evolving threat landscapes.
	Collaborative development	Global threat intelligence sharing, open-source model contributions, and standardized evaluation benchmarks.

Table 25. Cont.

Category	Aspect	Description
Challenges	Mitigating false positives	Reduction in alert fatigue through confidence calibration, ensemble methods, and human-in-the-loop validation.
	Addressing ethical concerns	Establishment of governance frameworks for transparency, accountability, and responsible AI deployment.
	Improving accuracy	Ongoing enhancements in contextual understanding, long-context reasoning, and domain-specific fine-tuning.

12.1. Emerging Trends and Technologies

Emerging cybersecurity trends converge around integrating Large Language Models with advanced automation, Retrieval-Augmented Generation, and knowledge graph reasoning to enhance defensive capabilities, streamline incident response, and improve critical infrastructure resilience. LLMs demonstrate tangible benefits in malware detection, phishing identification, anomaly detection, and automated compliance monitoring, with reported improvements in detection accuracy and operational efficiency, though limitations including outdated training data, high false positive rates, and privacy risks persist [12–14]. The technological trajectory points toward multi-objective fine-tuning, hybrid architectures integrating LLMs with symbolic reasoning and graph-based methods, and hardware-in-the-loop frameworks generating realistic datasets for sectors such as smart grids [10]. Adversarial use cases—automated social engineering, prompt injection, and AI-driven misinformation—highlight the LLM dual-use nature and necessitate responsible governance and transparent model auditing [92,103,185].

Benchmarking efforts including the SECURE framework mark critical progress toward standardized evaluation protocols, while domain-specific applications in supply chain defense, cloud security, IoT protection, and autonomous vehicle cybersecurity signal transition from proof-of-concept research to operational deployment [209,233–236]. While LLMs are positioned as core components of future cyberdefense ecosystems, adoption must be accompanied by robust mitigation mechanisms addressing bias, transparency, adversarial robustness, and privacy protection, ensuring innovation aligns with security and ethical imperatives. Tables 26 and 27 consolidate current applications, emerging directions, and forward-looking use cases of LLMs in cybersecurity.

Table 26. Applications of Large Language Models (LLMs) in cybersecurity.

Application Area		Description	Key Benefits	Ref.
Malware detection		Utilizes LLMs to identify and analyze malware signatures and behaviors.	Enhanced detection rates and accuracy.	[237]
Phishing detection		Simulates phishing scenarios to develop effective countermeasures.	Improved resilience against social engineering attacks.	[211]
Incident-response automation		Automates threat detection and response processes using LLM capabilities.	Faster response times and reduced manual effort.	[90]
IAM integration		Employs LLMs for dynamic access control and policy management.	More robust security frameworks.	[238]
Ethical considerations	considerations	Addresses challenges related to transparency and interpretability of LLMs.	Promotes responsible AI use in cybersecurity.	[239,240]

Table 27. Emerging trends and technologies utilizing LLMs for cybersecurity applications.

Trend/Technology	Application	Reference(s)
Defensive Applications	Anomaly detection in IEC 61850-based smart grids	Zaboli et al. [10]
	Web application vulnerability assessment	Nana et al. [233], Cao et al. [234], Ferrag et al. [235], Szabo et al. [241]
Adversarial Applications	Phishing email generation and social engineering frameworks	Zou et al. [242], Afane et al. [116], Chen et al. [243], Roy et al. [111]
	Adversarial attack frameworks targeting LLMs	Motlagh et al. [92]
Knowledge Integration	Joint reasoning pipelines augmented by knowledge graphs	Lu et al. [222]
Benchmarking and Evaluation	SECURE benchmark for end-to-end LLM cybersecurity evaluation	Bhusal et al. [209]
Scam and Phishing Detection	Phishing detection using GPT-3.5 and GPT-4 models	Jiang et al. [122], Heiding et al. [244], Koide et al. [245], Studies [246,247]
Hardware-in-the-Loop (HIL)	Synthetic dataset generation for energy sector HIL simulations	Zaboli et al. [10]

12.2. The Evolving Nature of Cyber Threats

Cyber threats are dynamic, leveraging advances in artificial intelligence and proliferation of interconnected devices to exploit vulnerabilities at an unprecedented scale, rendering traditional rule-based security frameworks inadequate. Attackers employ Large Language Models (LLMs) to craft adaptive phishing campaigns, automate social engineering, and generate polymorphic malware, while defenders simultaneously experiment with the same models to enhance anomaly detection, automate reporting, and improve real-time threat intelligence [12,13]. This dual-use nature underscores the unpredictability of modern threats, as adversaries continuously refine attack vectors to bypass detection, while defenders race to update countermeasures, creating an escalating cycle of innovation. Social engineering remains a dominant threat vector, amplified by LLMs' capacity to generate highly convincing and context-aware phishing messages, thereby demanding equally adaptive detection systems that integrate AI-assisted monitoring with human oversight.

At the same time, the widespread deployment of smart and IoT devices expands the attack surface, enabling adversaries to exploit novel pathways that require scalable and resilient defensive architectures [14]. Emerging research highlights the necessity of hybrid intelligence approaches that combine automated LLM-driven analysis with human expertise to mitigate risks of bias, hallucination, and ethical concerns while maintaining situational awareness. Ultimately, the evolving nature of cyber threats calls for flexible, ethically grounded, and continuously adaptive defensive strategies capable of leveraging LLM strengths without succumbing to their vulnerabilities.

13. Conclusions

This systematic review synthesized evidence from 235 peer-reviewed studies (2020–2025) examining Large Language Model integration into cybersecurity within Big Data infrastructures. This meta-analysis reveals task-domain-specific performance patterns: vulnerability detection demonstrates the largest improvement ($\Delta F_1 = +0.11$, 95% CI: 0.07–0.15), followed by intrusion detection and incident triage ($\Delta F_1 = +0.10$), while phishing detection shows modest gains ($\Delta F_1 = +0.06$) due to already-strong baseline systems ($F_1 = 0.88$). Across all domains, LLM-based methods reduce per-alert processing latency by 31–39%, enabling real-time

threat response in high-volume Security Operations Centers. However, effective deployment requires careful architectural matching to organizational context: domain-tailored models outperform general-purpose alternatives by +8.4% mean F_1 only when training datasets exceed 50,000 labeled samples and the operational volume justifies re-training costs (minimum of 10,000 events/day for phishing and 100,000 events/day for intrusion detection). Organizations processing fewer than 10,000 security events daily achieve superior cost-effectiveness with general-purpose models combined with few-shot prompting.

Organizations should approach LLM adoption as a spectrum of implementation strategies optimized for specific threat profiles, resource constraints, and regulatory environments. High-volume enterprise environments (>50,000 events/second) benefit from fine-tuned domain-specific models achieving F_1 scores of 0.90–0.94 with latencies of 370–640 ms, whereas critical infrastructure with sub-500 ms requirements necessitates hybrid architectures combining fast rule-based filtering (processing 85–90% of events) with selective LLM analysis. Compliance-regulated sectors face fundamental trade-offs: studies reporting F_1 improvements > 10% consistently document 15–22 percentage point declines in explainable AI fidelity metrics, necessitating either lower-performance interpretable models or human-in-the-loop validation, increasing analyst workload by 20–30%. The six real-world SOC deployment studies unanimously identify analyst cognitive overload as the primary failure mode, underscoring that successful implementations allocate 30–40% of budgets to interface design, workflow integration, and training. Among the 235 reviewed studies, 81% of those employing Retrieval-Augmented Generation report superior adaptability to zero-day threats, maintaining detection accuracy within 3–5% of fine-tuned approaches while enabling threat intelligence updates in <1 h versus 12–48 h for re-training cycles.

This review emphasizes the critical importance of task-domain-specific analysis in evaluating LLM effectiveness for cybersecurity. Effect sizes vary substantially across domains—from +0.06 F_1 (phishing) to +0.11 F_1 (vulnerability)—reflecting fundamental differences in data modality, task complexity, and baseline system performance. Practitioners and researchers must resist applying generic “LLM improves cybersecurity” conclusions across contexts. Organizations deploying LLMs should prioritize high-gain domains (vulnerability detection, intrusion detection, and incident triage) before investing in low-gain domains. Researchers should conduct domain-specific benchmarking, reporting effect sizes separately by task and data modality rather than aggregating across heterogeneous tasks. Policy makers should recognize that LLM-enhanced cybersecurity is not a monolithic capability but rather a suite of domain-specific applications with varying maturity levels, effectiveness, and readiness for operational deployment. Future meta-analyses and systematic reviews should perform stratification by task domain and report domain-specific effect sizes with transparent heterogeneity assessment (Tables 3 and 4). Table 28 synthesizes key findings and aligned recommendations, providing actionable deployment guidance mapping organizational requirements to optimal architectural approaches.

Table 28. Condensed findings and deployment recommendations for LLMs in cybersecurity (Big Data context).

Key Finding (KF)	Actionable Recommendation (AR)	≥50k Labels	≥10k ev/day	≤500 ms RT	Regulated	Zero-Day env.	HITL/HCI
KF1: Domain-tuned > general (F_1 : +6–13%, mean +8.4%)	AR1: Prefer fine-tuning when datasets ≥ 50k and volume ≥ 10k/day; ensure ≥ 32 GB GPU; otherwise, use general + few-shot ($F_1 \approx 0.85$ –0.87).	Req.	Req.	Cond.	Cond.	No	Cond.
KF2: RAG sustains adaptability (within 3–5% of FT; updates <1 h; +150–300 ms)	AR2: Prioritize RAG for evolving threats (APTs and polymorphic malware) over periodic re-training.	No	Cond.	No	Cond.	Req.	Cond.

Table 28. Cont.

Key Finding (KF)	Actionable Recommendation (AR)	$\geq 50k$ Labels	$\geq 10k$ ev/day	≤ 500 ms RT	Regulated	Zero-Day env.	HITL/HCI
KF3: Latency–accuracy patterns (general: 1200–1800 ms; FT: 370–640 ms; lightweight: 180–320 ms)	AR3: For <500 ms SLAs, deploy lightweight or hybrid pipelines: rules prefilter 85–90%; LLM score residuals.	No	Req.	Req.	Cond.	Cond.	Cond.
KF4: XAI trade-off ($>10\%$ F_1 gain $\Rightarrow$ XAI -15 – -22 pp)	AR4: In compliance-sensitive domains, use interpretable models or mandate human-in-the-loop validation (workload $+20$ – 30%).	No	Cond.	Cond.	Req.	No	Req.
KF5: Human–AI workflows drive outcomes (alert reduction of 40–60%; acceptance $\propto$ explanation quality, $r = 0.73$)	AR5: Allocate 30–40% of budget to UI/workflows/training; LLMs for triage/enrichment and humans for validation/action.	No	Cond.	Cond.	Req.	Cond.	Req.

Legend: **Req.**, primarily required/applicable; **Cond.**, conditional or partially applicable; **No**, not primarily applicable.

13.1. Concrete Findings and Actionable Recommendations

Synthesis. LLM efficacy for cybersecurity is context-dependent: domain-tuned models deliver $+8.4\%$ mean F_1 (6–13%) only with $\geq 50k$ labels and $\geq 10k$ events/day (GPU ≥ 32 GB); RAG adapts to zero-days within <1 h at $+150$ – 300 ms and remains within 3–5% of fine-tuned accuracy; real-time (<500 ms) favors lightweight or hybrid designs (rules prefilter 85–90%); $>10\%$ F_1 gains reduce XAI fidelity by 15–22 pp; effective SOCs use LLM triage to cut alerts by 40–60% and invest 30–40% in HCI/training. Table 28 summarizes key findings and aligned recommendations.

13.2. Research Gaps Requiring Immediate Attention

Gap 1: Systematic Benchmarking Under Realistic Operational Constraints

Only 12% of reviewed studies quantify latency–accuracy trade-offs under production workloads. Future research must establish standardized benchmarks incorporating

- Bursty traffic patterns (simulating DDoS and coordinated attacks);
- Data drift scenarios (temporal distribution shifts in threat patterns);
- Resource contention (multi-tenant cloud environments).

Gap 2: Explainability–Performance Pareto Frontiers

The current literature lacks systematic exploration of architectural modifications that improve both accuracy and interpretability. Promising directions include

- Hybrid architectures with interpretable first-stage classifiers;
- Constrained generation techniques enforcing structured, auditable outputs;
- Post hoc explanation methods specifically designed for security contexts.

Gap 3: Cross-Domain Transfer Learning

No studies systematically evaluate whether models fine-tuned on one domain (e.g., phishing) transfer effectively to related domains (e.g., business email compromise). Quantifying transfer learning efficiency could reduce data requirements by 40–60%.

Gap 4: Long-Term Operational Studies

The six SOC deployment studies span only 3–8 months. Multi-year longitudinal research is essential to assessing

- Model drift and re-training frequency requirements;
- Analyst skill development and trust calibration over time;
- Total cost of ownership, including maintenance and updates.

13.3. Future Directions

Building on the identified challenges, several research directions are pressing. Future work should prioritize developing LLMs resistant to prompt injection, backdoor exploitation, and data poisoning through adversarial training, red-teaming protocols, and trustworthy evaluation benchmarks. Systematic methods to detect and mitigate bias

in LLM outputs are needed, with explainability frameworks tailored to cybersecurity applications supporting analyst understanding, accountability, and regulatory compliance. Research should focus on fine-tuning and pre-training strategies for sector-specific contexts—customized models for healthcare, financial services, and industrial IoT environments promise more accurate detection of specialized attack vectors while reducing false positives. Empirical work is required to design effective interfaces integrating LLM outputs into analyst workflows, studying trust calibration, alert fatigue, and division of labor between automated systems and human expertise in SOC settings. As security telemetry grows, optimizing LLM–Big Data pipelines for latency, throughput, and cost efficiency becomes crucial through distributed training, federated learning, and lightweight inference architectures. Beyond technical safeguards, interdisciplinary research must address privacy preservation, compliance with regulatory frameworks (GDPR, AI Act, etc.), and standards for responsible LLM use in cybersecurity. Progress in LLM-assisted cybersecurity depends not only on model innovation but also on infrastructure design, governance, and interdisciplinary collaboration, requiring movement from proof-of-concept demonstrations to robust, transparent, and ethically grounded deployment.

Methodological transparency materials provided as Supplementary Materials:

- **Supplementary Table S1:** PRISMA-2020 checklist documenting all reporting items.
- **Supplementary Table S2a:** Characteristics of all 235 included studies, including task domains, datasets, model families, evaluation protocols, and deployment settings.
- **Supplementary Table S2b:** Outcome harmonization assumptions and measurement rules, including detailed latency normalization procedures, subgroup stratification criteria (real-time vs. batch; cloud API vs. self-hosted), and sensitivity analysis specifications (Row 3: Latency subgroup stratification).
- **Supplementary Table S2c:** Full-text exclusion reasons for all 177 studies excluded during screening, with per-article justifications.
- **Supplementary Table S3:** Risk-of-bias ratings for all 68 comparative/experimental studies, with domain-level judgments and supporting rationale.
- **Supplementary Note S1:** Full verbatim electronic search strategies with Boolean logic, field specifications, and date filters for all databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and arXiv).

Reproducibility statement: All effect sizes reported in this review (Table 3: pooled F1 improvements and latency reductions by task domain) were computed using random-effects meta-analysis (DerSimonian–Laird method) implemented in Python (version 3.11.7) with statsmodels (version 0.14.2) and scipy (version 1.11.4). The statistical procedures, heterogeneity assessment methods (I^2 , τ^2), and sensitivity analysis protocols are fully specified in Section 2.6 and Supplementary Table S2b. Per-study outcomes (F1 scores, latency values, etc.) are reported in the original source publications, all of which are cited in the References Section and listed in Supplementary Table S2a with full bibliographic details.

Analysis code availability: Meta-analysis computations employed standard statistical formulas (DerSimonian–Laird random-effects model and Egger’s regression for publication bias) that are fully reproducible using the methods specified in Section 2.6.3 and Supplementary Table S2b. Organizations or researchers seeking to replicate the analyses may implement these formulas using the specified Python libraries or equivalent statistical software (e.g., R metafor package and Stata metan command).

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/info16110957/s1>, Table S1: PRISMA-2020 checklist and item–location mapping, Supplementary Note S1: Full electronic search strategies for all sources, Table S2a: Characteristics of included studies, Table S2b: Outcome harmonization assumptions and rules, Table S2c:

Full-text exclusions with per-article reasons, Table S3: Risk-of-bias ratings by study and summary counts, Figure S1. Funnel plots and Egger’s tests for eligible syntheses, Figure S2: LLM-based framework for cybersecurity education, Figure S3: Framework of LLM-based cybersecurity in the tourism sector, Figure S4: Encryption-backed Retrieval-Augmented Generation pipeline for SOC workloads.

Author Contributions: A.K., L.T., C.K., A.T., I.K., and G.K. conceived of the idea, designed and constructed the review article, analyzed the applications of LLMs for Cybersecurity in the Big Data Era, drafted the initial manuscript and revised the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new primary data were created during this study. All quantitative outcomes analyzed in this review are publicly available in the cited source publications. The complete reference list (235 studies) with Digital Object Identifiers (DOIs) or URLs enables independent verification and replication of all extracted data points. Supplementary Table S2a cross-references each study with its task domain, evaluation metrics, and specific outcomes, facilitating traceability to original sources.

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

Abbreviations

The following abbreviations are used in this manuscript:

ACL	Association for Computational Linguistics
AI	Artificial intelligence
AIS	Automatic Identification System
API	Application Programming Interface
APT	Advanced Persistent Threat
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BIP	Block Interaction Protocol
BPNM	Business Process Model and Notation
CI	Confidence Interval
CSF	Cybersecurity Framework (NIST CSF 2.0)
CTI	Cyber Threat Intelligence
CVSS	Common Vulnerability Scoring System
DL	Deep learning
DOI	Digital Object Identifier
DP-SGD	Differential Privacy—Stochastic Gradient Descent
EDR	Endpoint Detection and Response
EHR	Electronic Health Records
F1	F1 score (harmonic mean of precision and recall)
GNN	Graph Neural Network
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act
IDPS	Intrusion detection and prevention system
IoMT	Internet of Medical Things
IoT	Internet of Things
LLM	Large Language Model
ML	Machine Learning

NMEA	National Marine Electronics Association
NER	Named-Entity Recognition
NLP	Natural language processing
NIST	National Institute of Standards and Technology
NIS	Network and Information Security
OT	Operational Technology
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO	International Prospective Register of Systematic Reviews
PQ-LLM	Post-Quantum Large Language Model
RAG	Retrieval-Augmented Generation
RoB	Risk of Bias
SBOM	Software Bill of Materials
SCADA	Supervisory Control and Data Acquisition
SIEM	Security Information and Event Management
SOC	Security Operations Center
SOAR	Security Orchestration, Automation and Response
STIX	Structured Threat Information Expression
TPU	Tensor Processing Unit
XAI	Explainable artificial intelligence

References

1. Gelman, H.; Hastings, J.D. Scalable and Ethical Insider Threat Detection through Data Synthesis and Analysis by LLMs. *arXiv* **2025**, arXiv:2502.07045. [\[CrossRef\]](#)
2. Portnoy, A.; Azikri, E.; Kels, S. Towards Automatic Hands-on-Keyboards Attack Detection Using LLMs in EDR Solutions. *arXiv* **2024**, arXiv:2408.01993.
3. Diakhame, M.L.; Diallo, C.; Mejri, M. MCM-Llama: A Fine-Tuned Large Language Model for Real-Time Threat Detection through Security Event Correlation. In Proceedings of the 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET), Sydney, Australia, 25–27 July 2024; pp. 1–6.
4. Mudassar Yamin, M.; Hashmi, E.; Ullah, M.; Katt, B. Applications of LLMs for Generating Cyber Security Exercise Scenarios. *IEEE Access* **2024**, *12*, 143806–143822. [\[CrossRef\]](#)
5. Kwan, W.C.; Zeng, X.; Jiang, Y.; Wang, Y.; Li, L.; Shang, L.; Jiang, X.; Liu, Q.; Wong, K.F. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv* **2024**, arXiv:2401.16745.
6. Xu, H.; Wang, S.; Li, N.; Wang, K.; Zhao, Y.; Chen, K.; Yu, T.; Liu, Y.; Wang, H. Large language models for cyber security: A systematic literature review. *arXiv* **2024**, arXiv:2405.04760. [\[CrossRef\]](#)
7. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv* **2024**, arXiv:2402.07927. [\[CrossRef\]](#)
8. Chen, Y.; Cui, M.; Wang, D.; Cao, Y.; Yang, P.; Jiang, B.; Lu, Z.; Liu, B. A survey of large language models for cyber threat detection. *Comput. Secur.* **2024**, *145*, 104016. [\[CrossRef\]](#)
9. Ali, T.; Kostakos, P. Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms). *arXiv* **2023**, arXiv:2309.16021.
10. Zaboli, A.; Choi, S.L.; Song, T.J.; Hong, J. Chatgpt and other large language models for cybersecurity of smart grid applications. In Proceedings of the 2024 IEEE Power & Energy Society General Meeting (PESGM), Seattle, WA, USA, 21–25 July 2024; pp. 1–5.
11. Omar, M.; Zangana, H.M.; Al-Karaki, J.N.; Mohammed, D. Harnessing LLMs for IoT Malware Detection: A Comparative Analysis of BERT and GPT-2. In Proceedings of the 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 7–9 November 2024; pp. 1–6. [\[CrossRef\]](#)
12. Güven, M. A Comprehensive Review of Large Language Models in Cyber Security. *Int. J. Comput. Exp. Sci. Eng.* **2024**, *10*. [\[CrossRef\]](#)
13. Gholami, N.Y. Large Language Models (LLMs) for Cybersecurity: A Systematic Review. *World J. Adv. Eng. Technol. Sci.* **2024**, *13*, 57–69. [\[CrossRef\]](#)
14. Zhang, J.; Bu, H.; Wen, H.; Liu, Y.; Fei, H.; Xi, R.; Li, L.; Yang, Y.; Zhu, H.; Meng, D. When llms meet cybersecurity: A systematic literature review. *Cybersecurity* **2025**, *8*, 1–41. [\[CrossRef\]](#)
15. Wan, S.; Nikolaidis, C.; Song, D.; Molnar, D.; Crnkovich, J.; Grace, J.; Bhatt, M.; Chennabasappa, S.; Whitman, S.; Ding, S.; et al. Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models. *arXiv* **2024**, arXiv:2408.01605. [\[CrossRef\]](#)

16. Bhatt, M.; Chennabasappa, S.; Li, Y.; Nikolaidis, C.; Song, D.; Wan, S.; Ahmad, F.; Aschermann, C.; Chen, Y.; Kapil, D.; et al. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models. *arXiv* **2024**, arXiv:2404.13161.
17. Nguyen, T.; Nguyen, H.; Ijaz, A.; Sheikhi, S.; Vasilakos, A.V.; Kostakos, P. Large language models in 6g security: Challenges and opportunities. *arXiv* **2024**, arXiv:2403.12239. [[CrossRef](#)]
18. Lorencin, I.; Tankovic, N.; Etinger, D. Optimizing Healthcare Efficiency with Local Large Language Models. *Intell. Hum. Syst. Integr. (IHSI 2025) Integr. People Intell. Syst.* **2025**, *160*, 576–584.
19. Nagaraja, N.; Bahşi, H. Cyber Threat Modeling of an LLM-Based Healthcare System. In Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025), Porto, Portugal, 20–22 February 2025; pp. 325–336. [[CrossRef](#)]
20. Karras, A.; Giannaros, A.; Karras, C.; Giotopoulos, K.C.; Tsoilis, D.; Sioutas, S. Edge Artificial Intelligence in Large-Scale IoT Systems, Applications, and Big Data Infrastructures. In Proceedings of the 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Piraeus, Greece, 10–12 November 2023; pp. 1–8. [[CrossRef](#)]
21. Terawi, N.; Ashqar, H.I.; Darwish, O.; Alsobeh, A.; Zahariev, P.; Tashtoush, Y. Enhanced Detection of Intrusion Detection System in Cloud Networks Using Time-Aware and Deep Learning Techniques. *Computers* **2025**, *14*, 282. [[CrossRef](#)]
22. Schizas, N.; Karras, A.; Karras, C.; Sioutas, S. TinyML for ultra-low power AI and large scale IoT deployments: A systematic review. *Future Internet* **2022**, *14*, 363. [[CrossRef](#)]
23. Harasees, A.; Al-Ahmad, B.; Alsobeh, A.; Abuhussein, A. A secure IoT framework for remote health monitoring using fog computing. In Proceedings of the 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), Dubrovnik, Croatia, 24–27 September 2024; pp. 17–24.
24. Alshatnawi, S.; AlSobeh, A.M. A cloud-based IoT smart water distribution framework utilising BIP component: Jordan as a model. *Int. J. Cloud Comput.* **2024**, *13*, 25–41. [[CrossRef](#)]
25. Sharafaldin, I.; Lashkari, A.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *ICISSP* **2018**, *1*, 108–116.
26. Order of the Overflow. DEF CON Capture the Flag 2019 Dataset. 2019. Available online: <https://oooverflow.io/dc-ctf-2019-finals/> (accessed on 27 October 2025).
27. Fontugne, R.; Fukuda, K.; Akiba, S. MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking. In Proceedings of the Symposium on Recent Advances in Intrusion Detection (RAID), Ottawa, ON, Canada, 15–17 September 2010.
28. Lippmann, R.P.; Haines, J.W.; Fried, D.J.; Korba, J.; Das, K. The 1999 DARPA Off-Line Intrusion Detection Evaluation. In Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX), Hilton Head, SC, USA, 25–27 January 2000.
29. Sagioglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
30. Naeem, M.; Jamal, T.; Diaz-Martinez, J.; Butt, S.A.; Montesano, N.; Tariq, M.I.; De-la Hoz-Franco, E.; De-La-Hoz-Valdiris, E. Trends and future perspective challenges in big data. In *Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, Arad, Romania, 15–18 October 2019*; Springer: Singapore, 2022; pp. 309–325.
31. Deepa, N.; Pham, Q.V.; Nguyen, D.C.; Bhattacharya, S.; Prabadevi, B.; Gadekallu, T.R.; Maddikunta, P.K.R.; Fang, F.; Pathirana, P.N. A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Gener. Comput. Syst.* **2022**, *131*, 209–226. [[CrossRef](#)]
32. Han, X.; Gstrein, O.J.; Andrikopoulos, V. When we talk about Big Data, What do we really mean? Toward a more precise definition of Big Data. *Front. Big Data* **2024**, *7*, 1441869. [[CrossRef](#)]
33. Hassan, A.A.; Hassan, T.M. Real-time big data analytics for data stream challenges: An overview. *Eur. J. Inf. Technol. Comput. Sci.* **2022**, *2*, 1–6. [[CrossRef](#)]
34. Abawajy, J. Comprehensive analysis of big data variety landscape. *Int. J. Parallel Emergent Distrib. Syst.* **2015**, *30*, 5–14. [[CrossRef](#)]
35. Mao, R.; Xu, H.; Wu, W.; Li, J.; Li, Y.; Lu, M. Overcoming the challenge of variety: Big data abstraction, the next evolution of data management for AAL communication systems. *IEEE Commun. Mag.* **2015**, *53*, 42–47. [[CrossRef](#)]
36. Pendyala, V. Veracity of big data. In *Machine Learning and Other Approaches to Verifying Truthfulness*; Apress: New York, NY, USA, 2018.
37. Berti-Equille, L.; Borge-Holthoefer, J. *Veracity of Data*; Springer Nature: Berlin, Germany, 2022.
38. Tahseen, A.; Shailaja, S.R.; Ashwini, Y. Extraction for Big Data Cyber Security Analytics. In *Advances in Computational Intelligence and Informatics: Proceedings of ICACII 2023*; Springer Nature: Berlin, Germany, 2024; Volume 993, p. 365.
39. Vernik, G.; Factor, M.; Kolodner, E.K.; Ofer, E.; Michiardi, P.; Pace, F. Stocator: An object store aware connector for apache spark. In Proceedings of the 2017 Symposium on Cloud Computing, Santa Clara, CA, USA, 24–27 September 2017; p. 653.

40. Rupperecht, L.; Zhang, R.; Owen, B.; Pietzuch, P.; Hildebrand, D. SwiftAnalytics: Optimizing Object Storage for Big Data Analytics. In Proceedings of the 2017 IEEE International Conference on Cloud Engineering (IC2E), Vancouver, BC, Canada, 4–7 April 2017; pp. 245–251. [\[CrossRef\]](#)
41. Baek, S.; Kim, Y.G. C4I system security architecture: A perspective on big data lifecycle in a military environment. *Sustainability* **2021**, *13*, 13827. [\[CrossRef\]](#)
42. Al-Kateb, M.; Eltabakh, M.Y.; Al-Omari, A.; Brown, P.G. Analytics at Scale: Evolution at Infrastructure and Algorithmic Levels. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 3217–3220. [\[CrossRef\]](#)
43. de Sousa, V.M.; Cura, L.M.d.V. Logical design of graph databases from an entity-relationship conceptual model. In Proceedings of the 20th International Conference on Information Integration and Web-Based Applications & Services, Yogyakarta, Indonesia, 19–21 November 2018; pp. 183–189.
44. Thepa, T.; Ateetanan, P.; Khubpatiwithhayakul, P.; Fugkeaw, S. Design and Development of Scalable SIEM as a Service using Spark and Anomaly Detection. In Proceedings of the 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 19–22 June 2024; pp. 199–205. [\[CrossRef\]](#)
45. Alawadhi, R.; Aalmohamed, H.; Alhashemi, S.; Alkhazaleh, H.A. Application of Big Data in Cybersecurity. In Proceedings of the 2024 7th International Conference on Signal Processing and Information Security (ICSPIS), Online, 12–14 November 2024; pp. 1–6.
46. Udeh, E.O.; Amajuoyi, P.; Adeusi, K.B.; Scott, A.O. The role of big data in detecting and preventing financial fraud in digital transactions. *World J. Adv. Res. Rev.* **2024**, *22*, 1746–1760. [\[CrossRef\]](#)
47. Li, L.; Qiang, F.; Ma, L. Advancing Cybersecurity: Graph Neural Networks in Threat Intelligence Knowledge Graphs. In Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security, Nanchang, China, 26–28 April 2024; pp. 737–741.
48. Gulbay, B.; Demirci, M. A Framework for Developing Strategic Cyber Threat Intelligence from Advanced Persistent Threat Analysis Reports Using Graph-Based Algorithms. *Preprints* **2024**. [\[CrossRef\]](#)
49. Rabzelj, M.; Bohak, C.; Južnič, L.Š.; Kos, A.; Sedlar, U. Cyberattack graph modeling for visual analytics. *IEEE Access* **2023**, *11*, 86910–86944. [\[CrossRef\]](#)
50. Wang, C.; Li, Y.; Liu, L. Algorithm Innovation and Integration with Big Data Technology in the Field of Information Security: Current Status and Future Development. *Acad. J. Eng. Technol. Sci.* **2024**, *7*, 45–49. [\[CrossRef\]](#)
51. Artioli, P.; Maci, A.; Magri, A. A comprehensive investigation of clustering algorithms for User and Entity Behavior Analytics. *Front. Big Data* **2024**, *7*, 1375818. [\[CrossRef\]](#)
52. Wang, J.; Yan, T.; An, D.; Liang, Z.; Guo, C.; Hu, H.; Luo, Q.; Li, H.; Wang, H.; Zeng, S.; et al. A comprehensive security operation center based on big data analytics and threat intelligence. In Proceedings of the International Symposium on Grids & Clouds, Taipei, Taiwan, 22–26 March 2021; Volume 2021.
53. Bharani, D.; Lakshmi Priya, V.; Saravanan, S. Adaptive Real-Time Malware Detection for IoT Traffic Streams: A Comparative Study of Concept Drift Detection Techniques. In Proceedings of the 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, 17–18 December 2024; pp. 172–179. [\[CrossRef\]](#)
54. K, S.; K S, N.; S, P.; S P, M.; Saranya. Analysis, Trends, and Utilization of Security Information and Event Management (SIEM) in Critical Infrastructures. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 14–15 March 2024; Volume 1, pp. 1980–1984. [\[CrossRef\]](#)
55. Saipranith, S.; Singh, A.K.; Agrawal, N.; Chilumula, S. SwiftFrame: Developing Low-latency Near Real-time Response Framework. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–6. [\[CrossRef\]](#)
56. Polepaka, S.; Bansal, S.; Al-Fatlawy, R.R.; Subburam, S.; Lakra, P.P.; Neyyila, S. Cloud-Based Marketing Analytics Using Apache Flink for Real-Time Data Insights. In Proceedings of the 2024 International Conference on IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 23–24 November 2024; pp. 1308–1313. [\[CrossRef\]](#)
57. Kanka, V. *Scaling Big Data: Leveraging LLMs for Enterprise Success*; Libertatem Media Private Limited: New Delhi, India, 2024.
58. Andrés, P.; Nikolai, I.; Zhihao, W. Real-Time AI-Based Threat Intelligence for Cloud Security Enhancement. *Innov. Int. Multi-Discip. J. Appl. Technol.* **2025**, *3*, 36–54.
59. Chitimoju, S. Enhancing Cyber Threat Intelligence with NLP and Large Language Models. *J. Big Data Smart Syst.* **2025**, *6*. Available online: <https://universe-publisher.com/index.php/jbds/article/view/80> (accessed on 27 October 2025).
60. Tanksale, V. Cyber Threat Hunting Using Large Language Models. In Proceedings of the International Congress on Information and Communication Technology, London, UK, 19–22 February 2024; pp. 629–641.
61. Wu, Y. The Role of Mining and Detection of Big Data Processing Techniques in Cybersecurity. *Appl. Math. Nonlinear Sci.* **2024**, *9*. [\[CrossRef\]](#)
62. Nugroho, S.A.; Sumaryanto, S.; Hadi, A.P. The Enhancing Cybersecurity with AI Algorithms and Big Data Analytics: Challenges and Solutions. *J. Technol. Inform. Eng.* **2024**, *3*, 279–295. [\[CrossRef\]](#)

63. Aameedeen, M.A.; Hamid, R.A.; Aldhyani, T.H.; Al-Nassr, L.A.K.M.; Olatunji, S.O.; Subramanian, P. A framework for automated big data analytics in cybersecurity threat detection. *Mesopotamian J. Big Data* **2024**, *2024*, 175–184. [\[CrossRef\]](#)
64. Nwobodo, L.K.; Nwaimo, C.S.; Adegbola, A.E. Enhancing cybersecurity protocols in the era of big data and advanced analytics. *GSC Adv. Res. Rev.* **2024**, *19*, 203–214. [\[CrossRef\]](#)
65. Hasan, M.; Hoque, A.; Le, T. Big data-driven banking operations: Opportunities, challenges, and data security perspectives. *FinTech* **2023**, *2*, 484–509. [\[CrossRef\]](#)
66. Sufi, F.; Alsulami, M. Mathematical Modeling and Clustering Framework for Cyber Threat Analysis Across Industries. *Mathematics* **2025**, *13*, 655. [\[CrossRef\]](#)
67. Chinta, P.C.R.; Jha, K.M.; Velaga, V.; Moore, C.; Routhu, K.; Sadaram, G. Harnessing Big Data and AI-Driven ERP Systems to Enhance Cybersecurity Resilience in Real-Time Threat Environments. *SSRN Electron. J.* **2024**. [\[CrossRef\]](#)
68. Jagadeesan, D.; Kartheesan, L.; Purushotham, B.; Krishna, S.T.; Kumar, S.N.; Asha, G. Data Analytics Techniques for Privacy Protection in Cybersecurity for Leveraging Machine Learning for Advanced Threat Detection. In Proceedings of the 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 4–6 October 2024; pp. 1–6.
69. Singh, R.; Aravindan, V.; Mishra, S.; Singh, S.K. Streamlined Data Pipeline for Real-Time Threat Detection and Model Inference. In Proceedings of the 2025 17th International Conference on COMMunication Systems and NETworks (COMSNETS), Bengaluru, India, 6–10 January 2025; pp. 1148–1153.
70. Dewasiri, N.J.; Dharmarathna, D.G.; Choudhary, M. Leveraging artificial intelligence for enhanced risk management in banking: A systematic literature review. In *Artificial Intelligence Enabled Management: An Emerging Economy Perspective*; Walter de Gruyter GmbH & Co. KG: Berlin, Germany, 2024; pp. 197–213.
71. Moharrak, M.; Mogaji, E. Generative AI in banking: Empirical insights on integration, challenges and opportunities in a regulated industry. *Int. J. Bank Mark.* **2025**, *43*, 871–896. [\[CrossRef\]](#)
72. Fernandez, J.; Wehrstedt, L.; Shamis, L.; Elhoushi, M.; Saladi, K.; Bisk, Y.; Strubell, E.; Kahn, J. Hardware Scaling Trends and Diminishing Returns in Large-Scale Distributed Training. *arXiv* **2024**, arXiv:2411.13055. [\[CrossRef\]](#)
73. Tang, Z.; Kang, X.; Yin, Y.; Pan, X.; Wang, Y.; He, X.; Wang, Q.; Zeng, R.; Zhao, K.; Shi, S.; et al. Fusionllm: A decentralized llm training system on geo-distributed gpus with adaptive compression. *arXiv* **2024**, arXiv:2410.12707.
74. Yang, F.; Peng, S.; Sun, N.; Wang, F.; Wang, Y.; Wu, F.; Qiu, J.; Pan, A. Holmes: Towards distributed training across clusters with heterogeneous NIC environment. In Proceedings of the 53rd International Conference on Parallel Processing, Gotland, Sweden, 12–15 August 2024; pp. 514–523.
75. Chen, Z.; Shao, H.; Li, Y.; Lu, H.; Jin, J. Policy-Based Access Control System for Delta Lake. In Proceedings of the 2022 Tenth International Conference on Advanced Cloud and Big Data (CBD), Guilin, China, 4–5 November 2022; pp. 60–65. [\[CrossRef\]](#)
76. Tang, S.; He, B.; Yu, C.; Li, Y.; Li, K. A Survey on Spark Ecosystem: Big Data Processing Infrastructure, Machine Learning, and Applications (Extended abstract). In Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, 3–7 April 2023; pp. 3779–3780. [\[CrossRef\]](#)
77. Douillard, A.; Feng, Q.; Rusu, A.A.; Chhaparia, R.; Donchev, Y.; Kuncoro, A.; Ranzato, M.; Szlam, A.; Shen, J. Diloco: Distributed low-communication training of language models. *arXiv* **2023**, arXiv:2311.08105.
78. Li, J.; Hui, B.; Qu, G.; Yang, J.; Li, B.; Li, B.; Wang, B.; Qin, B.; Geng, R.; Huo, N.; et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 42330–42357.
79. Crawford, J.M.; Penberthy, L.; Pinto, L.A.; Althoff, K.N.; Assimon, M.M.; Cohen, O.; Gillim, L.; Hammonds, T.L.; Kapur, S.; Kaufman, H.W.; et al. Coronavirus Disease 2019 (COVID-19) Real World Data Infrastructure: A Big-Data Resource for Study of the Impact of COVID-19 in Patient Populations With Immunocompromising Conditions. *Open Forum Infect. Dis.* **2025**, *12*, ofaf021. [\[CrossRef\]](#)
80. Levandoski, J.; Casto, G.; Deng, M.; Desai, R.; Edara, P.; Hottelier, T.; Hormati, A.; Johnson, A.; Johnson, J.; Kurzyniec, D.; et al. BigLake: BigQuery’s Evolution toward a Multi-Cloud Lakehouse. In Proceedings of the Companion of the 2024 International Conference on Management of Data, Santiago, Chile, 9–15 June 2024; pp. 334–346.
81. Stankov, I.; Dulgerov, E. Comparing Azure Sentinel and ML-Extended Solutions Applied to a Zero Trust Architecture. In Proceedings of the 2024 32nd National Conference with International Participation (TELECOM), Sofia, Bulgaria, 21–22 November 2024; pp. 1–4.
82. Morić, Z.; Dakić, V.; Kapulica, A.; Regvart, D. Forensic Investigation Capabilities of Microsoft Azure: A Comprehensive Analysis and Its Significance in Advancing Cloud Cyber Forensics. *Electronics* **2024**, *13*, 4546. [\[CrossRef\]](#)
83. Borra, P. Securing Cloud Infrastructure: An In-Depth Analysis of Microsoft Azure Security. *Int. J. Adv. Res. Sci. Commun. Technol. (IJARSCT)* **2024**, *4*, 549–555. [\[CrossRef\]](#)
84. Tuyishime, E.; Balan, T.C.; Cotfas, P.A.; Cotfas, D.T.; Rekeraho, A. Enhancing cloud security—Proactive threat monitoring and detection using a siem-based approach. *Appl. Sci.* **2023**, *13*, 12359. [\[CrossRef\]](#)
85. Shah, S.; Parast, F.K. AI-Driven Cyber Threat Intelligence Automation. *arXiv* **2024**, arXiv:2410.20287. [\[CrossRef\]](#)

86. Hassanin, M.; Keshk, M.; Salim, S.; Alsubaie, M.; Sharma, D. Plm-cs: Pre-trained large language model (llm) for cyber threat detection in satellite networks. *Ad Hoc Netw.* **2025**, *166*, 103645. [\[CrossRef\]](#)
87. Jing, P.; Tang, M.; Shi, X.; Zheng, X.; Nie, S.; Wu, S.; Yang, Y.; Luo, X. SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity. *arXiv* **2024**, arXiv:2412.20787.
88. Marantos, C.; Evangelatos, S.; Veroni, E.; Lalas, G.; Chasapas, K.; Christou, I.T.; Lappas, P. Leveraging Large Language Models for Dynamic Scenario Building targeting Enhanced Cyber-threat Detection and Security Training. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 2779–2788. [\[CrossRef\]](#)
89. Ferrag, M.A.; Alwahedi, F.; Battah, A.; Cherif, B.; Mechri, A.; Tihanyi, N.; Bisztray, T.; Debbah, M. Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities. *Internet Things Cyber-Phys. Syst.* **2025**, *5*, 1–46. [\[CrossRef\]](#)
90. Kasri, W.; Himeur, Y.; Alkhazaleh, H.A.; Tarapiah, S.; Atalla, S.; Mansoor, W.; Al-Ahmad, H. From Vulnerability to Defense: The Role of Large Language Models in Enhancing Cybersecurity. *Computation* **2025**, *13*, 30. [\[CrossRef\]](#)
91. da Silva, F.A. Navigating the dual-edged sword of generative AI in cybersecurity. *Braz. J. Dev.* **2025**, *11*, e76869. [\[CrossRef\]](#)
92. Motlagh, F.N.; Hajizadeh, M.; Majd, M.; Najafi, P.; Cheng, F.; Meinel, C. Large language models in cybersecurity: State-of-the-art. *arXiv* **2024**, arXiv:2402.00891. [\[CrossRef\]](#)
93. Pan, Z.; Liu, J.; Dai, Y.; Fan, W. Large Language Model-enabled Vulnerability Investigation: A Review. In Proceedings of the 2024 International Conference on Intelligent Computing and Next Generation Networks (ICNGN), Bangkok, Thailand, 23–25 November 2024; pp. 1–5.
94. Bai, G.; Chai, Z.; Ling, C.; Wang, S.; Lu, J.; Zhang, N.; Shi, T.; Yu, Z.; Zhu, M.; Zhang, Y.; et al. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv* **2024**, arXiv:2401.00625. [\[CrossRef\]](#)
95. Xu, M.; Cai, D.; Yin, W.; Wang, S.; Jin, X.; Liu, X. Resource-efficient algorithms and systems of foundation models: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–39. [\[CrossRef\]](#)
96. Liu, J.; Liao, Y.; Xu, H.; Xu, Y. Resource-Efficient Federated Fine-Tuning Large Language Models for Heterogeneous Data. *arXiv* **2025**, arXiv:2503.21213.
97. Theodorakopoulos, L.; Karras, A.; Theodoropoulou, A.; Kambiotis, G. Benchmarking Big Data Systems: Performance and Decision-Making Implications in Emerging Technologies. *Technologies* **2024**, *12*, 217. [\[CrossRef\]](#)
98. Theodorakopoulos, L.; Karras, A.; Krimpas, G.A. Optimizing Apache Spark MLlib: Predictive Performance of Large-Scale Models for Big Data Analytics. *Algorithms* **2025**, *18*, 74. [\[CrossRef\]](#)
99. Karras, C.; Theodorakopoulos, L.; Karras, A.; Krimpas, G.A. Efficient algorithms for range mode queries in the big data era. *Information* **2024**, *15*, 450. [\[CrossRef\]](#)
100. Lin, Z.; Cui, J.; Liao, X.; Wang, X. Malla: Demystifying real-world large language model integrated malicious services. In Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24), Philadelphia, PA, USA, 14–16 August 2024; pp. 4693–4710.
101. Charan, P.; Chunduri, H.; Anand, P.M.; Shukla, S.K. From text to mitre techniques: Exploring the malicious use of large language models for generating cyber attack payloads. *arXiv* **2023**, arXiv:2305.15336. [\[CrossRef\]](#)
102. Clairoux-Trepanier, V.; Beauchamp, I.M.; Ruellan, E.; Paquet-Clouston, M.; Paquette, S.O.; Clay, E. The use of large language models (llm) for cyber threat intelligence (cti) in cybercrime forums. *arXiv* **2024**, arXiv:2408.03354. [\[CrossRef\]](#)
103. Majumdar, D.; Arjun, S.; Boyina, P.; Rayidi, S.S.P.; Sai, Y.R.; Gangashetty, S.V. Beyond text: Nefarious actors harnessing llms for strategic advantage. In Proceedings of the 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS), Gurugram, India, 3–4 May 2024; pp. 1–7.
104. Zhao, S.; Jia, M.; Tuan, L.A.; Pan, F.; Wen, J. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. *arXiv* **2024**, arXiv:2401.05949. [\[CrossRef\]](#)
105. Zhou, Y.; Ni, T.; Lee, W.B.; Zhao, Q. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluations. *arXiv* **2025**, arXiv:2502.05224. [\[CrossRef\]](#)
106. Yang, H.; Xiang, K.; Ge, M.; Li, H.; Lu, R.; Yu, S. A comprehensive overview of backdoor attacks in large language models within communication networks. *IEEE Netw.* **2024**, *38*, 211–218. [\[CrossRef\]](#)
107. Ge, H.; Li, Y.; Wang, Q.; Zhang, Y.; Tang, R. When Backdoors Speak: Understanding LLM Backdoor Attacks Through Model-Generated Explanations. *arXiv* **2024**, arXiv:2411.12701. [\[CrossRef\]](#)
108. Li, Y.; Xu, Z.; Jiang, F.; Niu, L.; Sahabandu, D.; Ramasubramanian, B.; Poovendran, R. CleanGen: Mitigating backdoor attacks for generation tasks in large language models. *arXiv* **2024**, arXiv:2406.12257. [\[CrossRef\]](#)
109. Trad, F.; Chehab, A. Prompt engineering or fine-tuning? A case study on phishing detection with large language models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 367–384. [\[CrossRef\]](#)
110. Asfour, M.; Murillo, J.C. Harnessing large language models to simulate realistic human responses to social engineering attacks: A case study. *Int. J. Cybersecur. Intell. Cybercrime* **2023**, *6*, 21–49. [\[CrossRef\]](#)

111. Roy, S.S.; Thota, P.; Naragam, K.V.; Nilizadeh, S. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2024; pp. 36–54.
112. Ai, L.; Kumarage, T.; Bhattacharjee, A.; Liu, Z.; Hui, Z.; Davinroy, M.; Cook, J.; Cassani, L.; Trapeznikov, K.; Kirchner, M.; et al. Defending against social engineering attacks in the age of llms. *arXiv* **2024**, arXiv:2406.12263. [\[CrossRef\]](#)
113. Jamal, S.; Wimmer, H.; Sarker, I.H. An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *Secur. Priv.* **2024**, *7*, e402. [\[CrossRef\]](#)
114. Malloy, T.; Ferreira, M.J.; Fang, F.; Gonzalez, C. Training Users Against Human and GPT-4 Generated Social Engineering Attacks. *arXiv* **2025**, arXiv:2502.01764. [\[CrossRef\]](#)
115. Wan, Z.; Wang, X.; Liu, C.; Alam, S.; Zheng, Y.; Liu, J.; Qu, Z.; Yan, S.; Zhu, Y.; Zhang, Q.; et al. Efficient large language models: A survey. *arXiv* **2023**, arXiv:2312.03863.
116. Afane, K.; Wei, W.; Mao, Y.; Farooq, J.; Chen, J. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 2558–2567.
117. Kulkarni, A.; Balachandran, V.; Divakaran, D.M.; Das, T. From ml to llm: Evaluating the robustness of phishing webpage detection models against adversarial attacks. *arXiv* **2024**, arXiv:2407.20361. [\[CrossRef\]](#)
118. Kamruzzaman, A.S.; Thakur, K.; Mahub, S. AI Tools Building Cybercrime & Defenses. In Proceedings of the 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Victoria, Seychelles, 1–2 February 2024; pp. 1–5. [\[CrossRef\]](#)
119. Yang, S.; Zhu, S.; Wu, Z.; Wang, K.; Yao, J.; Wu, J.; Hu, L.; Li, M.; Wong, D.F.; Wang, D. Fraud-R1: A Multi-Round Benchmark for Assessing the Robustness of LLM Against Augmented Fraud and Phishing Inducements. *arXiv* **2025**, arXiv:2502.12904.
120. Wang, J.; Huang, Z.; Liu, H.; Yang, N.; Xiao, Y. Defecthunter: A novel llm-driven boosted-conformer-based code vulnerability detection mechanism. *arXiv* **2023**, arXiv:2309.15324.
121. Andriushchenko, M.; Souly, A.; Dziemian, M.; Duenas, D.; Lin, M.; Wang, J.; Hendrycks, D.; Zou, A.; Kolter, Z.; Fredrikson, M.; et al. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv* **2024**, arXiv:2410.09024. [\[CrossRef\]](#)
122. Jiang, L. Detecting scams using large language models. *arXiv* **2024**, arXiv:2402.03147. [\[CrossRef\]](#)
123. Hays, S.; White, J. Employing llms for incident response planning and review. *arXiv* **2024**, arXiv:2403.01271. [\[CrossRef\]](#)
124. Çaylı, O. AI-Enhanced Cybersecurity Vulnerability-Based Prevention, Defense, and Mitigation using Generative AI. *Orclever Proc. Res. Dev.* **2024**, *5*, 655–667. [\[CrossRef\]](#)
125. Novelli, C.; Casolari, F.; Hacker, P.; Spedicato, G.; Floridi, L. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Comput. Law Secur. Rev.* **2024**, *55*, 106066. [\[CrossRef\]](#)
126. Derasari, P.; Venkataramani, G. EPIC: Efficient and Proactive Instruction-level Cyberdefense. In Proceedings of the Great Lakes Symposium on VLSI 2024, GLSVLSI '24, Clearwater, FL, USA, 12–14 June 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 409–414. [\[CrossRef\]](#)
127. Bataineh, A.Q.; Abu-AlSondos, I.A.; Idris, M.; Mushtaha, A.S.; Qasim, D.M. The role of big data analytics in driving innovation in digital marketing. In Proceedings of the 2023 9th International Conference on Optimization and Applications (ICOA), Abu Dhabi, United Arab Emirates, 5–6 October 2023; pp. 1–5.
128. Chaurasia, S.S.; Kodwani, D.; Lachhwani, H.; Ketkar, M.A. Big data academic and learning analytics: Connecting the dots for academic excellence in higher education. *Int. J. Educ. Manag.* **2018**, *32*, 1099–1117. [\[CrossRef\]](#)
129. Hassanin, M.; Moustafa, N. A comprehensive overview of large language models (llms) for cyber defences: Opportunities and directions. *arXiv* **2024**, arXiv:2405.14487. [\[CrossRef\]](#)
130. Ji, H.; Yang, J.; Chai, L.; Wei, C.; Yang, L.; Duan, Y.; Wang, Y.; Sun, T.; Guo, H.; Li, T.; et al. Sevenllm: Benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence. *arXiv* **2024**, arXiv:2405.03446. [\[CrossRef\]](#)
131. Bokkena, B. Enhancing IT Security with LLM-Powered Predictive Threat Intelligence. In Proceedings of the 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 18–20 September 2024; pp. 751–756. [\[CrossRef\]](#)
132. Balasubramanian, P.; Ali, T.; Salmani, M.; Khoshkholgh, D.; Kostakos, P. Hex2Sign: Automatic IDS Signature Generation from Hexadecimal Data using LLMs. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 4524–4532. [\[CrossRef\]](#)
133. Webb, B.K.; Purohit, S.; Meyur, R. Cyber knowledge completion using large language models. *arXiv* **2024**, arXiv:2409.16176. [\[CrossRef\]](#)
134. Song, J.; Wang, X.; Zhu, J.; Wu, Y.; Cheng, X.; Zhong, R.; Niu, C. RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Miami, FL, USA, 12–16 November 2024; pp. 1548–1558.
135. Gandhi, P.A.; Wudali, P.N.; Amaru, Y.; Elovici, Y.; Shabtai, A. SHIELD: APT Detection and Intelligent Explanation Using LLM. *arXiv* **2025**, arXiv:2502.02342. [\[CrossRef\]](#)

136. Ji, Z.; Chen, D.; Ishii, E.; Cahyawijaya, S.; Bang, Y.; Wilie, B.; Fung, P. Llm internal states reveal hallucination risk faced with a query. *arXiv* **2024**, arXiv:2407.03282. [\[CrossRef\]](#)
137. Maity, S.; Arora, J. The Colossal Defense: Security Challenges of Large Language Models. In Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, 21–23 November 2024; pp. 1–5. [\[CrossRef\]](#)
138. Ayzenshteyn, D.; Weiss, R.; Mirsky, Y. The Best Defense is a Good Offense: Countering LLM-Powered Cyberattacks. *arXiv* **2024**, arXiv:2410.15396. [\[CrossRef\]](#)
139. Kim, Y.; Dán, G.; Zhu, Q. Human-in-the-loop cyber intrusion detection using active learning. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 8658–8672. [\[CrossRef\]](#)
140. Ghanem, M.C. Advancing IoT and Cloud Security through LLMs, Federated Learning, and Reinforcement Learning. In Proceedings of the 7th IEEE Conference on Cloud and Internet of Things (CIoT 2024)—Keynote, Montreal, QC, Canada, 29–31 October 2024.
141. Haryanto, C.Y.; Elvira, A.M.; Nguyen, T.D.; Vu, M.H.; Hartanto, Y.; Lomempow, E.; Arakala, A. Contextualized AI for Cyber Defense: An Automated Survey using LLMs. In Proceedings of the 2024 17th International Conference on Security of Information and Networks (SIN), Sydney, Australia, 2–4 December 2024; pp. 1–8.
142. V, S.; P, L.S.; P, N.K.; V, L.P.; CH, B.S. Data Leakage Detection and Prevention Using Ciphertext-Policy Attribute Based Encryption Algorithm. In Proceedings of the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 14–15 March 2024; pp. 1–5. [\[CrossRef\]](#)
143. Kumar, P.S.; Bapu, B.T.; Sridhar, S.; Nagaraju, V. An Efficient Cyber Security Attack Detection With Encryption Using Capsule Convolutional Polymorphic Graph Attention. *Trans. Emerg. Telecommun. Technol.* **2025**, *36*, e70069. [\[CrossRef\]](#)
144. Chen, Y.; Chen, Z. Preventive Measures of Influencing Factors of Computer Network Security Technology. In Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 28–30 June 2021; pp. 1187–1191. [\[CrossRef\]](#)
145. Sarkorn, T.; Chimmanee, K. Review on Zero Trust Architecture Apply In Enterprise Next Generation Firewall. In Proceedings of the 2024 8th International Conference on Information Technology (InCIT), Chonburi, Thailand, 14–15 November 2024; pp. 255–260. [\[CrossRef\]](#)
146. Mustafa, H.M.; Basumallik, S.; Vellaithurai, C.; Srivastava, A. Threat Detection in Power Grid OT Networks: Unsupervised ML and Cyber Intelligence Sharing with STIX. In Proceedings of the 2024 12th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES), Hong Kong, China, 13 May 2024; pp. 1–6. [\[CrossRef\]](#)
147. Steingartner, W.; Galinec, D.; Zebić, V. Challenges of Application Programming Interfaces Security: A Conceptual Model in the Changing Cyber Defense Environment and Zero Trust Architecture. In Proceedings of the 2024 IEEE 17th International Scientific Conference on Informatics (Informatics), Poprad, Slovakia, 13–15 November 2024; pp. 372–379. [\[CrossRef\]](#)
148. Mmaduekwe, E.; Kessie, J.; Salawudeen, M. Zero trust architecture and AI: A synergistic approach to next-generation cybersecurity frameworks. *Int. J. Sci. Res. Arch.* **2024**, *13*, 4159–4169. [\[CrossRef\]](#)
149. Freitas, S.; Kalajdjieski, J.; Gharib, A.; McCann, R. AI-Driven Guided Response for Security Operation Centers with Microsoft Copilot for Security. *arXiv* **2024**, arXiv:2407.09017. [\[CrossRef\]](#)
150. Bono, J.; Xu, A. Randomized controlled trials for Security Copilot for IT administrators. *arXiv* **2024**, arXiv:2411.01067. [\[CrossRef\]](#)
151. Paul, S.; Alemi, F.; Macwan, R. LLM-Assisted Proactive Threat Intelligence for Automated Reasoning. *arXiv* **2025**, arXiv:2504.00428. [\[CrossRef\]](#)
152. Kshetri, N. Transforming cybersecurity with agentic AI to combat emerging cyber threats. *Telecommun. Policy* **2025**, *49*, 102976. [\[CrossRef\]](#)
153. Schesny, M.; Lutz, N.; Jägle, T.; Gerschner, F.; Klaiber, M.; Theissler, A. Enhancing Website Fraud Detection: A ChatGPT-Based Approach to Phishing Detection. In Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), Osaka, Japan, 2–4 July 2024; pp. 1494–1495. [\[CrossRef\]](#)
154. Razavi, H.; Jamali, M.R. Large Language Models (LLM) for Estimating the Cost of Cyber-attacks. In Proceedings of the 2024 11th International Symposium on Telecommunications (IST), Tehran, Iran, 9–10 October 2024; pp. 403–409. [\[CrossRef\]](#)
155. Mathew, A. AI Cyber Defense and eBPF. *World J. Adv. Res. Rev.* **2024**, *22*, 1983–1989. [\[CrossRef\]](#)
156. Baldoni, R.; De Nicola, R.; Prinetto, P. (Eds.) *The Future of Cybersecurity in Italy: Strategic Focus Areas. Projects and Actions to Better Defend Our Country from Cyber Attacks*; English Edition; Translated from the Italian Volume (Jan 2018, ISBN 9788894137330); last update 20 June 2018; CINI—Consorzio Interuniversitario Nazionale per l’Informatica: Rome, Italy, 2018.
157. Truong, T.C.; Diep, Q.B.; Zelinka, I. Artificial intelligence in the cyber domain: Offense and defense. *Symmetry* **2020**, *12*, 410. [\[CrossRef\]](#)
158. Ferrag, M.A.; Ndhlovu, M.; Tihanyi, N.; Cordeiro, L.C.; Debbah, M.; Lestable, T.; Thandi, N.S. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *IEEE Access* **2024**, *12*, 23733–23750. [\[CrossRef\]](#)
159. Metta, S.; Chang, I.; Parker, J.; Roman, M.P.; Ehuan, A.F. Generative AI in cybersecurity. *arXiv* **2024**, arXiv:2405.01674.

160. Benabderrahmane, S.; Valtchev, P.; Cheney, J.; Rahwan, T. APT-LLM: Embedding-Based Anomaly Detection of Cyber Advanced Persistent Threats Using Large Language Models. *arXiv* **2025**, arXiv:2502.09385.
161. Zhang, X.; Li, Q.; Tan, Y.; Guo, Z.; Zhang, L.; Cui, Y. Large Language Models powered Network Attack Detection: Architecture, Opportunities and Case Study. *arXiv* **2025**, arXiv:2503.18487. [[CrossRef](#)]
162. Zuo, F.; Rhee, J.; Choe, Y.R. Knowledge Transfer from LLMs to Provenance Analysis: A Semantic-Augmented Method for APT Detection. *arXiv* **2025**, arXiv:2503.18316. [[CrossRef](#)]
163. Ferrag, M.A.; Ndhlovu, M.; Tihanyi, N.; Cordeiro, L.C.; Debbah, M.; Lestable, T. Revolutionizing cyber threat detection with large language models. *arXiv* **2023**, arXiv:2306.14263.
164. Ren, H.; Lan, K.; Sun, Z.; Liao, S. CLogLLM: A Large Language Model Enabled Approach to Cybersecurity Log Anomaly Analysis. In Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC), Wuhan, China, 27–29 December 2024; pp. 963–970. [[CrossRef](#)]
165. Ismail, I.; Kurnia, R.; Brata, Z.A.; Nelistiani, G.A.; Heo, S.; Kim, H.; Kim, H. Toward Robust Security Orchestration and Automated Response in Security Operations Centers with a Hyper-Automation Approach Using Agentic Artificial Intelligence. *Information* **2025**, *16*, 365. [[CrossRef](#)]
166. Tallam, K. CyberSentinel: An Emergent Threat Detection System for AI Security. *arXiv* **2025**, arXiv:2502.14966. [[CrossRef](#)]
167. Kheddar, H. Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. *arXiv* **2024**, arXiv:2408.07583. [[CrossRef](#)]
168. Ghimire, A.; Ghajari, G.; Gurung, K.; Sah, L.K.; Amsaad, F. Enhancing Cybersecurity in Critical Infrastructure with LLM-Assisted Explainable IoT Systems. *arXiv* **2025**, arXiv:2503.03180.
169. Setak, M.; Madani, P. Fine-Tuning LLMs for Code Mutation: A New Era of Cyber Threats. In Proceedings of the 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), Washington, DC, USA, 28–31 October 2024; pp. 313–321.
170. Song, C.; Ma, L.; Zheng, J.; Liao, J.; Kuang, H.; Yang, L. Audit-LLM: Multi-Agent Collaboration for Log-based Insider Threat Detection. *arXiv* **2024**, arXiv:2408.08902.
171. Li, Y.; Xiang, Z.; Bastian, N.D.; Song, D.; Li, B. IDS-Agent: An LLM Agent for Explainable Intrusion Detection in IoT Networks. In Proceedings of the NeurIPS 2024 Workshop on Open-World Agents, Vancouver, BC, Canada, 15 December 2024.
172. Rigaki, M.; Catania, C.; Garcia, S. Hackphyr: A Local Fine-Tuned LLM Agent for Network Security Environments. *arXiv* **2024**, arXiv:2409.11276. [[CrossRef](#)]
173. Diaf, A.; Korba, A.A.; Karabadi, N.E.; Ghamri-Doudane, Y. BARTPredict: Empowering IoT Security with LLM-Driven Cyber Threat Prediction. In Proceedings of the GLOBECOM 2024-2024 IEEE Global Communications Conference, Cape Town, South Africa, 8–12 December 2024; pp. 1239–1244.
174. Barker, C. Applications of Machine Learning to Threat Intelligence, Intrusion Detection and Malware. Senior Honors Thesis, Liberty University, Lynchburg, VA, USA, 2020.
175. Bakdash, J.Z.; Hutchinson, S.; Zaroukian, E.G.; Marusich, L.R.; Thirumuruganathan, S.; Sample, C.; Hoffman, B.; Das, G. Malware in the future? Forecasting of analyst detection of cyber events. *J. Cybersecur.* **2018**, *4*, tyy007. [[CrossRef](#)]
176. Cheng, Y.; Bajaber, O.; Tsegai, S.A.; Song, D.; Gao, P. CTINEXUS: Leveraging Optimized LLM In-Context Learning for Constructing Cybersecurity Knowledge Graphs Under Data Scarcity. *arXiv* **2024**, arXiv:2410.21060. [[CrossRef](#)]
177. Al Siam, A.; Alazab, M.; Awajan, A.; Faruqui, N. A Comprehensive Review of AI's Current Impact and Future Prospects in Cybersecurity. *IEEE Access* **2025**, *13*, 14029–14050. [[CrossRef](#)]
178. Bashir, T. Zero Trust Architecture: Enhancing cybersecurity in enterprise networks. *J. Comput. Sci. Technol. Stud.* **2024**, *6*, 54–59. [[CrossRef](#)]
179. Hu, X.; Chen, H.; Bao, H.; Wang, W.; Liu, F.; Zhou, G.; Yin, P. A LLM-based agent for the automatic generation and generalization of IDS rules. In Proceedings of the 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Sanya, China, 17–21 December 2024; pp. 1875–1880.
180. Bianou, S.G.; Batogna, R.G. PENTEST-AI, an LLM-Powered multi-agents framework for penetration testing automation leveraging mitre attack. In Proceedings of the 2024 IEEE International Conference on Cyber Security and Resilience (CSR), London, UK, 2–4 September 2024; pp. 763–770.
181. Alzaabi, F.R.; Mehmood, A. A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods. *IEEE Access* **2024**, *12*, 30907–30927. [[CrossRef](#)]

182. Du, D.; Guan, X.; Liu, Y.; Jiang, B.; Liu, S.; Feng, H.; Liu, J. MAD-LLM: A Novel Approach for Alert-Based Multi-stage Attack Detection via LLM. In Proceedings of the 2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA), Kaifeng, China, 30 October–2 November 2024; pp. 2046–2053. [\[CrossRef\]](#)
183. Swetha, K.; K, S. Detection of Cybercriminal Activities in Smartphones via NLP-Based Communication Pattern Analysis. In Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, 21–23 November 2024; pp. 1–5. [\[CrossRef\]](#)
184. Vieira, A.C.; Houmb, S.H.; Insua, D.R. A graphical adversarial risk analysis model for oil and gas drilling cybersecurity. *arXiv* **2014**, arXiv:1404.1989. [\[CrossRef\]](#)
185. Usman, Y.; Upadhyay, A.; Gyawali, P.; Chataut, R. Is generative ai the next tactical cyber weapon for threat actors? Unforeseen implications of ai generated cyber attacks. *arXiv* **2024**, arXiv:2408.12806. [\[CrossRef\]](#)
186. Wang, L.; Wang, J.; Jung, K.; Thiagarajan, K.; Wei, E.; Shen, X.; Chen, Y.; Li, Z. From sands to mansions: Enabling automatic full-life-cycle cyberattack construction with llm. *arXiv* **2024**, arXiv:2407.16928.
187. Ruhländer, L.; Popp, E.; Styliadou, M.; Khan, S.; Svetinovic, D. On the Security and Privacy Implications of Large Language Models: In-Depth Threat Analysis. In Proceedings of the 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, Copenhagen, Denmark, 19–22 August 2024; pp. 543–550.
188. Gade, P.; Lermen, S.; Rogers-Smith, C.; Ladish, J. Badllama: Cheaply removing safety fine-tuning from llama 2-chat 13b. *arXiv* **2023**, arXiv:2311.00117.
189. Josten, M.; Schaffeld, M.; Lehmann, R.; Weis, T. Navigating the Security Challenges of LLMs: Positioning Target-Side Defenses and Identifying Research Gaps. In Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025), Porto, Portugal, 20–22 February 2025; pp. 240–247. [\[CrossRef\]](#)
190. Sarker, I.H. LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. *Discov. Artif. Intell.* **2024**, *4*, 40. [\[CrossRef\]](#)
191. Pupillo, L.; Ferreira, A.; Fantin, S. *Artificial Intelligence and Cybersecurity: Task Force Evaluation of the HLEG Trustworthy AI Assessment List (Pilot Version)*; Ceps Task Force Report; Centre for European Policy Studies (CEPS): Brussels, Belgium, 2020. [\[CrossRef\]](#)
192. Shafee, S.; Bessani, A.; Ferreira, P.M. Evaluation of LLM chatbots for OSINT-based cyber threat awareness. *arXiv* **2024**, arXiv:2401.15127. [\[CrossRef\]](#)
193. Hariharan, S.; Majid, Z.A.; Veuthey, J.R.; Haimes, J. Rethinking CyberSecEval: An LLM-Aided Approach to Evaluation Critique. *arXiv* **2024**, arXiv:2411.08813.
194. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of large language models such as ChatGPT for dental medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [\[CrossRef\]](#)
195. Keltek, M.; Hu, R.; Sani, M.F.; Li, Z. LSAST—Enhancing Cybersecurity through LLM-supported Static Application Security Testing. *arXiv* **2024**, arXiv:2409.15735.
196. Tann, W.; Liu, Y.; Sim, J.H.; Seah, C.M.; Chang, E.C. Using large language models for cybersecurity capture-the-flag challenges and certification questions. *arXiv* **2023**, arXiv:2308.10443.
197. Zhang, Y.; Cai, Y.; Zuo, X.; Luan, X.; Wang, K.; Hou, Z.; Zhang, Y.; Wei, Z.; Sun, M.; Sun, J.; et al. The Fusion of Large Language Models and Formal Methods for Trustworthy AI Agents: A Roadmap. *arXiv* **2024**, arXiv:2412.06512. [\[CrossRef\]](#)
198. Zhao, X.; Leng, X.; Wang, L.; Wang, N.; Liu, Y. Efficient anomaly detection in tabular cybersecurity data using large language models. *Sci. Rep.* **2025**, *15*, 3344. [\[CrossRef\]](#) [\[PubMed\]](#)
199. Regulation, P. General data protection regulation. *Intouch* **2018**, *25*, 1–5.
200. European Union Artificial Intelligence Act. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence*; Official Journal of the European Union: L 188, 12 July 2024; European Commission: Brussels, Belgium, 2024.
201. Markopoulou, D.; Papakonstantinou, V.; De Hert, P. The new EU cybersecurity framework: The NIS Directive, ENISA’s role and the General Data Protection Regulation. *Comput. Law Secur. Rev.* **2019**, *35*, 105336. [\[CrossRef\]](#)
202. Schmitz-Berndt, S. Defining the reporting threshold for a cybersecurity incident under the NIS Directive and the NIS 2 Directive. *J. Cybersecur.* **2023**, *9*, tyad009. [\[CrossRef\]](#)
203. Danezis, G.; Domingo-Ferrer, J.; Hansen, M.; Hoepman, J.H.; Metayer, D.L.; Tirtea, R.; Schiffner, S. Privacy and data protection by design—from policy to engineering. *arXiv* **2015**, arXiv:1501.03726.
204. 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*, 28335–28352. [\[CrossRef\]](#)
205. Yigit, Y.; Buchanan, W.J.; Tehrani, M.G.; Maglaras, L. Review of generative ai methods in cybersecurity. *arXiv* **2024**, arXiv:2403.08701. [\[CrossRef\]](#)
206. Adamec, M.; Turčaník, M. Development of Malware Using Large Language Models. In Proceedings of the 2024 New Trends in Signal Processing (NTSP), Demanovska Dolina, Slovakia, 16–18 October 2024; pp. 1–5.

207. Wahréus, J.; Hussain, A.M.; Papadimitratos, P. CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models. *arXiv* **2025**, arXiv:2501.01335.
208. Jones, N.; Whaiduzzaman, M.; Jan, T.; Adel, A.; Alazab, A.; Alkreisat, A. A CIA Triad-Based Taxonomy of Prompt Attacks on Large Language Models. *Future Internet* **2025**, *17*, 113. [\[CrossRef\]](#)
209. Bhusal, D.; Alam, M.T.; Nguyen, L.; Mahara, A.; Lightcap, Z.; Frazier, R.; Fieblinger, R.; Torales, G.L.; Blakely, B.A.; Rastogi, N. SECURE: Benchmarking Large Language Models for Cybersecurity. *arXiv* **2024**, arXiv:2405.20441. [\[CrossRef\]](#)
210. Guo, Y.; Patsakis, C.; Hu, Q.; Tang, Q.; Casino, F. Outside the comfort zone: Analysing llm capabilities in software vulnerability detection. In Proceedings of the European Symposium on Research in Computer Security, Bydgoszcz, Poland, 22–24 September 2024; pp. 271–289.
211. Hasanov, I.; Virtanen, S.; Hakkala, A.; Isoaho, J. Application of Large Language Models in Cybersecurity: A Systematic Literature Review. *IEEE Access* **2024**, *12*, 176751–176778. [\[CrossRef\]](#)
212. Balogh, Š.; Mlynček, M.; Vraňák, O.; Zajac, P. Using Generative AI Models to Support Cybersecurity Analysts. *Electronics* **2024**, *13*, 4718. [\[CrossRef\]](#)
213. Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; Chen, Y.C. CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 23164–23172. [\[CrossRef\]](#)
214. Nelson, C.; Doupe, A.; Shoshitaishvili, Y. SENSAT: Large Language Models as Applied Cybersecurity Tutors. In Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1, Pittsburgh, PA, USA, 26 February–1 March 2025; pp. 833–839.
215. Yan, Y.; Zhang, Y.; Huang, K. Depending on yourself when you should: Mentoring llm with rl agents to become the master in cybersecurity games. *arXiv* **2024**, arXiv:2403.17674. [\[CrossRef\]](#)
216. Tshimula, J.M.; Ndona, X.; Nkashama, D.K.; Tardif, P.M.; Kabanza, F.; Frappier, M.; Wang, S. Preventing Jailbreak Prompts as Malicious Tools for Cybercriminals: A Cyber Defense Perspective. *arXiv* **2024**, arXiv:2411.16642. [\[CrossRef\]](#)
217. Lodge, B. RAGe Against the Machine with BERT for Proactive Cybersecurity Posture. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 3579–3588.
218. Kakolu, S.; Faheem, M.A.; Aslam, M. Integrating Natural Language Processing with Cybersecurity Protocols: Real-Time Analysis of Malicious Intent in Social Engineering Attack. *Int. J. Sci. Res. Arch.* **2020**, *1*, 082–095. [\[CrossRef\]](#)
219. Rahman, M.R.; Wroblewski, B.; Tamanna, M.; Rahman, I.; Anufryienak, A.; Williams, L. Towards a taxonomy of challenges in security control implementation. In Proceedings of the 2024 Annual Computer Security Applications Conference (ACSAC), Honolulu, HI, USA, 9–13 December 2024; pp. 61–75.
220. Liu, Z. Multi-Agent Collaboration in Incident Response with Large Language Models. *arXiv* **2024**, arXiv:2412.00652. [\[CrossRef\]](#)
221. Svoboda, I.; Lande, D. Enhancing multi-criteria decision analysis with AI: Integrating analytic hierarchy process and GPT-4 for automated decision support. *arXiv* **2024**, arXiv:2402.07404. [\[CrossRef\]](#)
222. Ou, L.; Ni, X.; Wu, W.; Tian, Z. CyGPT: Knowledge Graph-Based Enhancement Techniques for Large Language Models in Cybersecurity. In Proceedings of the 2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC), Jinan, China, 23–26 August 2024; pp. 216–223. [\[CrossRef\]](#)
223. Kumar, N.M.; Lisa, F.T.; Islam, S.R. Prompt Chaining-Assisted Malware Detection: A Hybrid Approach Utilizing Fine-Tuned LLMs and Domain Knowledge-Enriched Cybersecurity Knowledge Graphs. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 1672–1677. [\[CrossRef\]](#)
224. Lukošiušė, K.; Swanda, A. LLM Cyber Evaluations Don't Capture Real-World Risk. *arXiv* **2025**, arXiv:2502.00072.
225. Andreoni, M.; Lunardi, W.T.; Lawton, G.; Thakkar, S. Enhancing Autonomous System Security and Resilience With Generative AI: A Comprehensive Survey. *IEEE Access* **2024**, *12*, 109470–109493. [\[CrossRef\]](#)
226. Ismail, M.; Alrabae, S. Empowering Future Cyber Defenders: Advancing Cybersecurity Education in Engineering and Computing with Experiential Learning. In Proceedings of the 2024 IEEE Frontiers in Education Conference (FIE), Washington, DC, USA, 13–16 October 2024; pp. 1–9. [\[CrossRef\]](#)
227. Greco, D.; Chianese, L. Exploiting LLMs for E-Learning: A Cybersecurity Perspective on AI-Generated Tools in Education. In Proceedings of the 2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense), Naples, Italy, 11–13 November 2024; pp. 237–242. [\[CrossRef\]](#)
228. Yu, Y.C.; Chiang, T.H.; Tsai, C.W.; Huang, C.M.; Tsao, W.K. Primus: A Pioneering Collection of Open-Source Datasets for Cybersecurity LLM Training. *arXiv* **2025**, arXiv:2502.11191. [\[CrossRef\]](#)
229. Balasubramanian, P.; Seby, J.; Kostakos, P. Transformer-based llms in cybersecurity: An in-depth study on log anomaly detection and conversational defense mechanisms. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 3590–3599.
230. Hamid, R.; Brohi, S. A Review of Large Language Models in Healthcare: Taxonomy, Threats, Vulnerabilities, and Framework. *Big Data Cogn. Comput.* **2024**, *8*, 161. [\[CrossRef\]](#)

231. Imtiaz, A.; Shehzad, D.; Nasim, F.; Afzaal, M.; Rehman, M.; Imran, A. Analysis of Cybersecurity Measures for Detection, Prevention, and Misbehaviour of Social Systems. In Proceedings of the 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), Abu Dhabi, United Arab Emirates, 21–24 November 2023; pp. 1–7. [\[CrossRef\]](#)
232. Yang, X.; Pan, L.; Zhao, X.; Chen, H.; Petzold, L.; Wang, W.Y.; Cheng, W. A survey on detection of llms-generated content. *arXiv* **2023**, arXiv:2310.15654. [\[CrossRef\]](#)
233. Nana, S.R.; Bassolé, D.; Guel, D.; Sié, O. Deep Learning and Web Applications Vulnerabilities Detection: An Approach Based on Large Language Models. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*. [\[CrossRef\]](#)
234. Cao, D.; Liao, Y.; Shang, X. RealVul: Can We Detect Vulnerabilities in Web Applications with LLM? *arXiv* **2024**, arXiv:2410.07573. [\[CrossRef\]](#)
235. Ferrag, M.A.; Battah, A.; Tihanyi, N.; Jain, R.; Maimuț, D.; Alwahedi, F.; Lestable, T.; Thandi, N.S.; Mechri, A.; Debbah, M.; et al. SecureFalcon: Are we there yet in automated software vulnerability detection with LLMs? *IEEE Trans. Softw. Eng.* **2025**, *51*, 1248–1265. [\[CrossRef\]](#)
236. Giannaros, A.; Karras, A.; Theodorakopoulos, L.; Karras, C.; Kranias, P.; Schizas, N.; Kalogeratos, G.; Tsolis, D. Autonomous vehicles: Sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions. *J. Cybersec. Priv.* **2023**, *3*, 493–543. [\[CrossRef\]](#)
237. Liu, Z. A Review of Advancements and Applications of Pre-Trained Language Models in Cybersecurity. In Proceedings of the 2024 12th International Symposium on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, 29–30 April 2024; pp. 1–10. [\[CrossRef\]](#)
238. Banerjee, N.T. The Convergence of IAM and AI: How Large Language Models Are Reshaping Cybersecurity. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2025**, *11*, 969–977. [\[CrossRef\]](#)
239. Islam, M.R. *Generative AI, Cybersecurity, and Ethics*; John Wiley & Sons: New York, NY, USA, 2024.
240. Gupta, M.; Akiri, C.; Aryal, K.; Parker, E.; Praharaj, L. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access* **2023**, *11*, 80218–80245. [\[CrossRef\]](#)
241. Szabó, Z.; Bilicki, V. A new approach to web application security: Utilizing gpt language models for source code inspection. *Future Internet* **2023**, *15*, 326. [\[CrossRef\]](#)
242. Zou, J.; Zhang, S.; Qiu, M. Adversarial attacks on large language models. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Birmingham, UK, 16–18 August 2024; pp. 85–96.
243. Chen, F.; Wu, T.; Nguyen, V.; Wang, S.; Hu, H.; Abuadbbba, A.; Rudolph, C. Adapting to Cyber Threats: A Phishing Evolution Network (PEN) Framework for Phishing Generation and Analyzing Evolution Patterns using Large Language Models. *arXiv* **2024**, arXiv:2411.11389. [\[CrossRef\]](#)
244. Heiding, F.; Schneier, B.; Vishwanath, A.; Bernstein, J.; Park, P.S. Devising and detecting phishing emails using large language models. *IEEE Access* **2024**, *12*, 42131–42146. [\[CrossRef\]](#)
245. Koide, T.; Nakano, H.; Chiba, D. ChatPhishDetector: Detecting Phishing Sites Using Large Language Models. *IEEE Access* **2024**, *12*, 154381–154400. [\[CrossRef\]](#)
246. Mahendru, S.; Pandit, T. SecureNet: A Comparative Study of DeBERTa and Large Language Models for Phishing Detection. In Proceedings of the 2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BD AI), Beijing, China, 5–7 July 2024; pp. 160–169. [\[CrossRef\]](#)
247. Chataut, R.; Gyawali, P.K.; Usman, Y. Can AI Keep You Safe? A Study of Large Language Models for Phishing Detection. In Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8–10 January 2024; pp. 0548–0554. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.