Future Generation Computer Systems 175 (2026) 108107

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Cybersecurity in the age of generative AI: A systematic taxonomy of AI-powered vulnerability assessment and risk management

Seyedeh Leili Mirtaheri [a,*], Narges Movahed [b], Reza Shahbazian [c], Valerio Pascucci [d], Andrea Pugliese [a]

[a] Department of Informatics, Modeling, Electronics and System Engineering, University of Calabria, Italy
[b] Mechanical, Energy, and Management Engineering Department, University of Calabria, Rende, 87036, CS, Italy
[c] Department of Humanities, University of Palermo, Palermo, 87036, Italy
[d] Scientific Computing and Imaging Institute, and School of Computing, University of Utah, USA

ARTICLE INFO

Keywords:
Generative AI
Cybersecurity
Vulnerability assessment
Risk management
Systematic review

ABSTRACT

The article discusses the transformative impact of Generative AI (GenAI) to the field of vulnerability assessment (VA) and risk management (RM) right from the beginning of their life cycle to the end in cybersecurity (CS). Through a systematic review of over 100 publications (2021-2025), we develop a comprehensive taxonomy classifying GenAI's dual offensive and defensive applications in VA/RM. The survey spells out the dominant techniques of GenAI and also points towards challenging aspects, which include security, explainability, and trustworthiness. The resultant findings reinforce the belief that GenAI could help resolve many traditional VA/RM challenges, thus providing fertile ground for research and practice in this area.

1. Introduction

Organizations are exposed to advanced threats, the average cost estimate of which ranges from 3.9 − 8.19 million and can go up to 400 billion [1]. Regarding data-breach incidents, average estimates are about 9.5 million [2], which impose a serious challenge on credibility [3]. This calls for an urgent need for rigorous vulnerability assessment (VA) and risk management (RM) practices to identify security weaknesses and mitigate the risks. An effective VA strategy constantly analyses and pinpoints system vulnerabilities, while RM assesses their impact and likelihood of exploitation. The arrival of AI, particularly Large Language Models (LLMs), is changing VA and RM [4]. GenAI can assist with vulnerability discovery and remediation during code analysis, threat simulation, and automated actions, hence reducing response time. In RM, it will augment the risk scoring and resource prioritization and aid in threat modeling. Recent studies explored GenAI in CS covering vulnerability detection [5] and automated repairs [6]. GenAI solutions such as prompt-based scanners or threat modeling platforms help in data collection for VA detection and RM scenario assessment. However, there remain issues with data quality, adversarial use [7], model interpretability, and ethical concerns [8]. In safety-critical domains, explainability is no longer optional-it is a must. This indicates that the AI-driven insights are both trustworthy and auditable, and this assurance is very fundamental for compliance and operational justification.

We explore the evolving landscape of GenAI in CS through: **RQ1)** How GenAI is transforming the end-to-end process of VA and RM? **RQ2)** How does our taxonomy clarify GenAI's dual use in the context of VA and RM? **RQ3)** Can GenAI overcome the key prioritization and remediation challenges within VA and RM in dynamic and evolving threat landscapes? **RQ4)** What standard of explainability is needed to ensure the trustworthiness of GenAI in VA? **RQ5)** What types of cyberattacks are the focus of the GenAI researchers? Through systematically addressing these questions, we aim to clarify the state of the art, limitations, and future paths for GenAI in VA and RM.

Methodology. We systematically collected over 340 papers (2020–2025) by ACM, IEEE, ScienceDirect, and Springer on GenAI in CS. We identified relevant studies through targeted keyword combinations ("Generative AI," "Vulnerability Assessments", "Risk Management", "GANs (Generative Adversarial Networks)", "LLMs", etc.), and filtered over 100 high quality papers to develop a taxonomy of GenAI applications in CS and RM, focusing on security domain intersections (cloud, IoT) and AI techniques (Natural Language Processing (NLP), synthetic data).

Contributions. This survey: 1) analyzes GenAI's role across the VA/RM lifecycle, 2) introduces a novel taxonomy for VA/RM-specific applications and impacts, and 3) evaluates techniques and key challenges in vulnerability analysis, risk prioritization, and explainability through 100+ papers. The results provide researchers and practitioners with

* Corresponding author.
E-mail address: leili.mirtaheri@dimes.unical.it (S.L. Mirtaheri).

https://doi.org/10.1016/j.future.2025.108107
Received 27 June 2025; Received in revised form 18 August 2025; Accepted 26 August 2025
Available online 28 August 2025

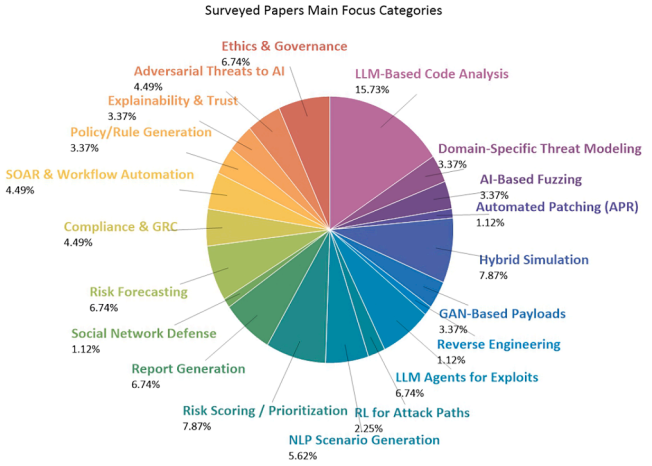


Fig. 1. Summary of main focus areas and current literature trends.

a comprehensive understanding of GenAI's current applications and future potential in enhancing VA and RM, in a stage-by-stage manner.

2. Background and related work

Research focus areas and trends in VA/RM: An overview. The current literature on GenAI in VA and RM seeks more unique paths, while many others remain gaps. Fig. 1 illustrates that the field is dominated by LLM-based code analysis (15.73 %) shows a trend toward using LLMs in finding vulnerabilities in APIs, smart contracts, and mobile apps. Interest in this space appears pronounced for hybrid simulation and risk scoring/prioritization (7.87 % each) accompanying the automation of scenario-based testing in the simulation of attacker behavior and the prioritization of vulnerabilities. Generation of reports, risk forecasting, and LLMs for exploitation modeling, which contribute equally at 6.74 %, underscore the importance of analytical rigor and automation in end-to-end VA/RM pipelines. While much progress has been made in the area, some disciplines still remain quite untrodden. Many areas, such as automatic patching, reverse engineering, and social network countermeasure (1.12 % each), remain unused and unempirical. Explainability, trust, and policy/rule generation (3.37 % each) are peripheral but necessary items for reliable AI-based VA/RM decisions. Given that today organizations want predictive and proactive tools, risk forecast and dynamic exploitation modeling with LLM agents may find some traction. Ethics, governance (6.74 %), and adversarial robustness (4.49 %) point at the need to secure GenAI models before they can be truly trusted for the purposes of VA/RM. Remediation, explainability, and formal validation remain underdeveloped, thus limiting the full application of GenAI in cyber RM.

Vulnerability assessment. VA is crucial in CS for identifying and eliminating system weaknesses [9]. Traditional levers include static analysis (scalable but prone to false positives) [10], dynamic analysis (runtime evaluation with performance overhead) [11], and hybrid methods (balanced but ineffective against zero-days (vulnerabilities exploited before patches are available)) [12,13]. Common tools like Common Vulnerability Scoring System (CVSS) [1] [14] and Common Vulnerabilities and Exposures (CVE) databases [15] certainly provide some standardization, but struggle with dynamic threats and zero-days, and often require manual confirmation. Such methods fall short against modern dynamic challenges like APTs, zero-days, and cloud/IoT attack surfaces due to reliance on static snapshots and expert judgment. Therefore, necessitating GenAI for adaptive RM (Fig. 1).

[1] https://www.first.org/cvss/v4-0/specification-document.

Prevalent cyberattack trends: An overview. Cyberattacks have greatly increased, with a direct impact on VA and RM in organizations. After COVID, software vulnerability exploitation peaked [3,16], targeting SQL injection, cross-site scripting, and AI-generated payloads on static VA and CVE/CVSS databases [2,17–20]. GenAI opened avenues for Dual-Use Weapons against VA/RM, with GANs and transformer models generating complex payloads that expose vulnerabilities [17–19], 51–62 % of which is vulnerable to API misuse, making detection challenging [21–24]. Prompt injections, data poisoning, and interruptions compromise AI-assisted VA/RM pipelines by adversaries [7,25–27], rendering the provision of risk predictions and compliance even more burdensome. Additionally, manual and responsive risk workflows can lag behind; hence, automated and adaptive VA/RM systems need to develop to facilitate real-time AI-assisted threat mitigation [28–32].

Evolution of AI in CS. AI has progressed from rule-based systems to ML and now GenAI [33]. Early systems addressed known threats, while modern ML detects subtle patterns. Deep learning (Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs)) excels in malware classification and traffic analysis [34,35], while NLP improves VA by parsing unstructured data (bug reports). GenAI synthesizes data (text, images) using models like GANs [36], transformers (GPT, BERT for log analysis) [37], and diffusion models (media synthesis) [38]. These enable both defensive applications and offensive risks.

Existing surveys. The role that GenAI plays in VA and RM has become the subject of quite a number of recent surveys, usually as part of broader CS topics with limited linking to code generation and exploit synthesis driven by GenAI [39–41]. Very few studies have been dedicated to how GenAI can be applied across the VA-RM lifecycle from the perspective of vulnerability discovery, risk forecasting, to compliance alignment. Issues such as their generality have been raised in association with model instability, privacy leakage, as well as bias [42,43], without any link to automated vulnerability analyses or predictive risk modeling. The role the XAI [44] plays in securing the GenAI-enabled VA/RM pipeline operationally or in mission-critical scenarios is equally less explored. The present study proposes: (1) an organized taxonomy describing the GenAI applications in VA and RM; (2) an analysis of GenAI across critical lifecycle stages in VA/RM; and (3) answers to five relevant research questions to direct future work toward secure, explainable, and adaptive VA/RM systems. This is a systematic survey and taxonomy-development exercise in GenAI-based tools for VA and RM. Hence, it is neither providing any alternative benchmarking scheme nor experimental validation. Primarily, it is a contribution toward establishing a framework, identifying patterns, and filling gaps in future investigations empirically focusing on a lack of standardization in datasets and evaluation criteria.

3. Taxonomy and lifecycle of GenAI in VA and RM

This section introduces an overview of the GenAI-powered VA/RM lifecycle, and a taxonomy that classifies GenAI's dual-use applications (offensive and defensive) across this lifecycle. Together, these provide a coherent structure for understanding GenAI's technical impact and practical roles in modern CS systems.

3.1. Lifecycle overview of GenAI in VA and RM

We propose a lifecycle (Fig. 2) to organize how GenAI technologies intervene across the end-to-end workflow of VA and RM. This lifecycle includes four primary stages: 1) **Discovery:** GenAI employs LLMs, GAN payloads for fuzzing, and code summarization tools for automated detection of vulnerabilities [45]. LLMs such as GPT-4 and CodeLlama surpass traditional scanning methods for smart contracts, APIs, and Android applications [46]. 2) **Assessment:** Then, GenAI aids

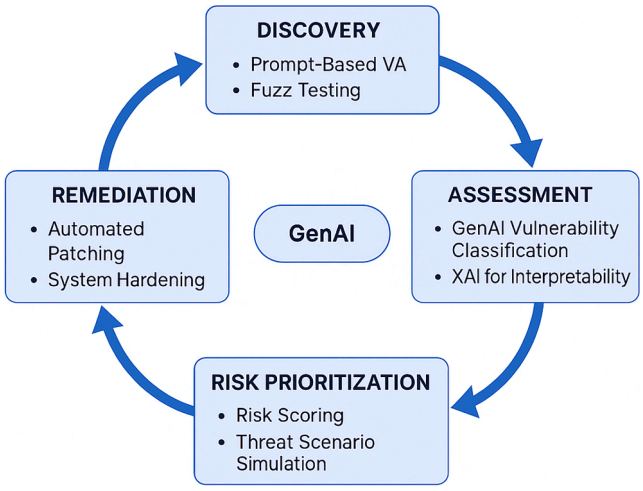


Fig. 2. GenAI-powered lifecycle for VA and RM.

Fig. 3. Taxonomies of GenAI applications in VA and RM.

in context-aware risk analysis, severity classification, and explainability [44,47]. Hybrid systems (LLMs with symbolic reasoning or formal verification) improve trust and support human validation [4,47]. 3) **Risk Prioritization:** GenAI enables dynamic risk modeling and prioritization using transformer-based CVSS scoring, RL-based threat simulations, and synthetic scenario generation (e.g., phishing or APT attack trees) [29,48–50]. 4) **Remediation:** GenAI accelerates patching, rule generation, and SOAR using systems such as DevSec-GPT (for DevSecOps) and ContrastRepair (for Automatic Repair) [51–54]. These tools generate actionable code patches, configure IDS policies, and provide human-readable mitigation plans. A feedback loop connects remediation with the discovery phase so that AI systems may learn from patched vulnerabilities and evolving attack vectors in an adaptive security operation with self-improvement. A detailed taxonomy of GenAI application for each life cycle phase is presented in Section 3.2, covering offensive and defensive use cases, models, techniques, and system-level strategies.

3.2. Taxonomy of GenAI applications in VA and RM

GenAI is notably impacting CS, especially in the domains of VA and RM. Offensively, it aids automated discovery of vulnerabilities [55–57] and simulated attacks to perform advanced testing of system weaknesses at scale. Defensively, it helps in vulnerability detection [58], threat modeling, and security automation [59–61] for ultimate risk reduction. Key concerns include adversarial attacks [7], LLM vulnerabilities [21,22], and trust issues countered through explainability [47]. Our taxonomy of GenAI applications in VA and RM is illustrated in Fig. 3.

4. Vulnerability discovery

Tools like GANs, variational autoencoders (VAEs) , and other GenAI learned methods are progressively used for vulnerability discovery, as well as threat analysis. While researchers have concentrated on these defensive applications, the field is also useful in code analysis, fuzz testing, and reverse engineering. Data augmentation is always enhanced using GANs, whereas vulnerability detection tasks are best served through transformer-based models.

LLM-based code analysis. Recent years have observed increasing applications of LLMs-like GPT and CodeLlama-in VA and RM for mobile applications, smart contracts, APIs, and open-source software. Kouliaridis et al. [55] assess nine LLMs in targeting Android code by means of the Mobile Top 10 of the Open Worldwide Application Security Project (OWASP) and displays variability in detection accuracies, contextual benefits of RAG, and inconsistency challenges in RM. Yildirim et al. [58] has stated that LLMs like ChatGPT-4 perform better than the traditional static analyzers in identifying OWASP API vulnerabilities with detection rates as high as 62.5 % by prompt engineering. In blockchain, Sun et al. [45] proposed the combination of GPT reasoning with static analysis to identify complex logic defects in Solidity contracts, achieving more than 90 % precision in smaller contracts and identifying faults missed by the auditors. In the same way, Boi et al. [62] suggested the use of GPT-3 to spot OWASP-classified vulnerabilities in Ethereum contracts, thus revealing risks missed by static tools. Under open-source ecosystems, Wu et al. [63] matched CVE descriptions to vulnerable functions by means of in-context learning, obtaining a localization accuracy that was over 4.25 times higher and reducing false positives with Software Composition Analysis (SCA) tools. In this regard, the mentioned works illustrate how LLMs may provide a more precise and scalable way of identifying vulnerabilities and profiling risks across current software settings.

Security of AI-generated code. LLMs are critiqued for writing vulnerable code in VA and RM. According to studies, 51–62 % of programs generated by GPT-4o-mini, Gemini Pro-1.0, and Code Llama, have reported vulnerabilities—integer overflow, buffer overflow, and memory safety issues—verified by formal methods like Efficient SMT-based Bounded Model Checker (ESBMC) [21,64]. Another dataset FormAI-v2, 331,000 compilable C programs, showed that even simple programming tasks can lead to vulnerable outputs, requiring careful risk assessment prior to deployment. An assessment of code snippets in real projects produced by GitHub Copilot also showed 24–29 % of the Python and JavaScript snippets to be insecure; across 43 categories of CWE, critical vulnerabilities such as improper randomness and cross-site scripting [22]. Misuse of security APIs emerged as a critical source of vulnerability in about 70 % of Java security API implementations generated by ChatGPT; 20 distinct patterns of misuse were identified, complicating risk mitigation [23]. Furthermore, there still exist problems of exploitation propagation due to the reliance of LLMs on training data with insecure patterns, aggravating systemic risks in software supply chains [24]. On the whole, these findings show that whereas LLMs foster speed in code generation, outputs rendered by them require VA to identify insecure code, followed by RM to address these vulnerabilities (see Section 8).

Improvement of secure code generation. Focusing on VA and risk mitigation, specialized techniques for GenAI security in code generation have been developed. From prompt engineering, [65] reasonably reviews prompting strategies (Recursive Criticism and Improvement, CWE-specific prompts, etc.) to decrease vulnerabilities in

LLM-generated code, confirming a measurable reduction in security flaws present in GPT-3.5 and GPT-4 outputs. Fine-tuning methods, per [57], propose risk mitigation through the contextualizing of LLMs with secure coding practices (SCPs), leaning toward strategies that suggest how developers may, through security-aware training, shape model outputs. Syntax-aware generation methods as those in [66] are set to leverage Abstract Syntax Trees (ASTs) and Graphs (ASGs) to generate syntactically sound code and decrease syntactic vulnerabilities; the application of this in malware detection code (trained on MITRE ATT&CK techniques [2].) showed a gain of 12.5 % over plaintext-based generation. Systems like LLMSecGuard [67] incorporate static analyzer functionality into LLMs for benchmarking and hardening generated code; this provides vulnerability detection with the possibility of post-hoc fixes. Examples of other frameworks that aim to improve security include datasets like FormAI [21,64]. However, there are hurdles in the path; adversarial prompts [68] exploit attribution-guided mutations to generate vulnerable code completions (69.6 % success rate above bases), whereas hallucinations remain an unaddressed menace in the domain of generative security workflows.

AI-driven fuzz testing. AI-assisted fuzz testing is an avenue of VA in which random inputs are replaced with well-calibrated test cases that closely resemble real-life attack scenarios. An example in this regard is the framework given in [56], where adversarial generative networks are used to generate realistic networks payloads focusing on IoT ecosystems like smart airports for improved vulnerability detection rates over traditional fuzzing techniques. It appends an LLM that assesses risk through scoring likelihood and consequence for targeted remediation. For intricate subjects like JavaScript engines, the study [69] suggests a neural strategy that uses syntax-aware generation and coverage feedback to address: (1) larger number of vocabulary with context-free grammar, (2) constraint-based mutations for satisfying semantic validity and (3) deep execution path exploration through grey-box testing. The semantic error rates in syntax-aware neural programming model (SNPM) have reduced by 24 % with the discovery of 20 new defects in JavaScript engines and 11 bugs in C compilers. Nevertheless, some obstacles remain, including the quality of training data, guarantee of semantic validity when applying lots of different test cases, and scaling of neural approaches for use with systems in enterprises. In addition, efficiency in detection must be properly balanced with cost of computation when GenAI meets traditional risk assessment; however, when carefully applied, it holds promise.

Automated software reverse engineering. GenAI with LLMs can contribute to the reverse engineering process by interpreting a large portion of the obfuscated code that helps with further malware analysis and scrutiny in cloaked-source software. From low-level data, code de-obfuscation, binary analysis, and so on, GenAI would produce high-level representations or descriptions of what the code does; the point of it is that LLMs understand code structure and semantics even in their obfuscated form.

4.1. Summary and emerging challenges

The use of Transformer-based LLMs and GANs in vulnerability discovery has increasingly become important, as they can perform particularly well in analyzing code for Android apps, smart contracts, and APIs. They are also able to supplement fuzz testing in IoT and JavaScript engines [45,55,56,58,62,69]. GenAI will be envisioned equally for reverse engineering obfuscated code and malware. A summary about such studies is shown in Table 1. Further, research on the integration of GenAI into end-to-end VA workflows has to be oriented toward using multimodal data and architectural advances to raise detection precision and scalability across different software ecosystems.

[2] a knowledge base of adversary tactics and techniques

Limitations and critical evaluation. LLMs such as GPT-4 and CodeLlama maintain more than 90 % precision for the detection of smart contract vulnerabilities while improving IoT fuzz testing performance through GANs; however, variable accuracies in detection due to contextual variability tend to increase the false positives and warrant additional large-scale manual validation [45,55,56]. So far, a noticeable percentage of the code generated by LLMs contains vulnerabilities such as integer overflows, which may potentially cause insecure software; on the other hand, classic fuzzing under GAN is hindered by low quality training data and high computation cost, which prevents it from being scalable [21,23,69]. The literature reports 62.5 % detection rates for OWASP API vulnerabilities and 4.25 times better localization in open-source software, but the real-world applicability suffers due to reliance on synthetic datasets such as FormAI-v2 (331,000 C programs) [58,63]. Hybrid human-AI validation, curated datasets, and lightweight neural models could reduce false positives and increase payload realism and scalability, yet their performance will require case studies to validate across diverse software domains [56,57,69].

5. Exploitation and penetration testing

GenAI is transforming the landscape of CS by automating and enhancing tasks central to VA and RM. The ability to generate realistic attack data, simulate exploitation scenarios, and improve detection capabilities allows GenAI to support proactive security posture evaluation and threat modeling.

GAN-based attack payload generation. GANs are increasingly being recognized as a powerful automating mechanism for creating generic synthetic payloads that could be used for all-inclusive security assessments. Improved Evolutionary GAN (IE-GAN) in [17] synthesizes complex SQL injection (SQLi) attack strings in a real-world scenario using evolutionary methods and GAN training, thus generating diverse, realistic queries that often escape conventional detection methods. Another such example would be Controller Area Network (CAN bus:a communication protocol in automotive systems) messages [70], which are used to simulate real-world exploitation scenarios and evaluate the robustness of defensive mechanisms. The Transformer-based X-squatter [20] is also exploring the detection of cross-language domain squatting, generating sound-alike candidates exposing TLS certificate and package repository vulnerabilities. An analysis of millions of certificates revealed that 15 % of sound-squatting candidates have active TLS certificates-more than twice as much as any other type of squatting-which suggests considerable platform-level weaknesses. These generative methods enable proactive RM through generation of realistic attack vectors to address significant high-impact vulnerability before exploitation. Synthetic test cases cause the enhancement of VAs, exposing inherent risks in the application logic and infrastructure that were not exposed by static analysis.

Autonomous exploitation with LLM agents. The LLaMA, GPT image and their fine-tuned versions have proven successful in automating VA tasks like the benchmarking-detecting social engineering attacks [71]. For instance, LLaMA-3-8b-instruct achieved high accuracy and F1-scores on phishing detection, and was shown to provide affordable solutions for Small and Medium-sized Enterprises (SMEs), though there are still risks because of poor dataset diversity and suboptimal parameter optimization [71]. Fine-tuning GPT has improved vulnerability detection through contextual email analysis, improving on the base models for identification of phishing and insider threats, hence reducing response time [72]. Human Feedback (RLHF)-Integrated LLMs are also shown to enhance risk communication, achieving 80 % precision in phishing semantic identification with zero false positive/negatives along with 91 % semantic similarity to expert-crafted alerts [73]. Such approach improves the clarity of warnings and offers actionable RM that connects automated detection to human perception of risk. Nevertheless, fine-

Table 1
Summary of vulnerability discovery studies.

Ref.	Domain/Focus	Method	Dataset(s)	Contribution
[55]	Android/ Vuln. Det.	LLM Eval (9), RAG	OWASP Mobile Top 10	LLM strengths/weaknesses for Android vulns; RAG impact
[58]	API/ Vuln. Det.	LLM Eval (4), Static An.	OWASP API Top 10	Prompted LLMs excel over static analyzers for API vulns
[45]	Smart Contract/ Logic Vuln. Det.	LLM (GPT) + Static An.	~400 contracts, Web3Bugs	LLM + static analysis effective for logic bugs
[62]	Smart Contract/ Vuln. Det.	LLM (GPT) - VulnHunt-GPT	SC vulns, Real contracts	GPT-based detector better for common SC vulns
[63]	Open Source Sec./ Vuln. Function ID	LLM (Chat/Mistral), ICL, SR	CVE, OSS Code	LLM approach better for finding vulnerable functions
[21]	Software/ AI Code Sec.	LLM Eval (9), Formal Verif.	FormAI-v2 (LLM-gen. C)	High vuln. rate in LLM C; validation needed
[64]	Software Sec./ AI Code Sec. Dataset	LLM (GPT), Formal Verif.	FormAI (GPT-gen. C)	Large dataset of AI CS code with vuln. labels; high rate
[22]	Software Sec./ Copilot Code Sec.	Static An., Manual Review	GitHub AI-gen. Python/JS	High vuln. rate in real-world AI code; Chat can fix
[23]	Java Sec./ API Misuse	LLM (ChatGPT), Code Review	48 Java Security API tasks	High rate of Java security API misuse by ChatGPT
[24]	Software Sec./ ChatGPT vs. SO Sec.	LLM (ChatGPT), Static An.	SO Java Q&A, ChatGPT resp.	ChatGPT less vuln., but both propagate insecure code
[65]	Software Dev./ Secure Code Gen. via Prompts	LLM Eval (GPT-3/4), Prompt Eng.	150 security-relevant prompts	Specific prompting reduces weaknesses in LLM code
[57]	Software Dev./ Secure Code Gen.	LLM Context.	LLM/Copilot code (qual.)	Strategies for more secure LLM code generation
[66]	Code Gen./ Malware Det. Code Gen.	RNN (AST/ASG)	Literature + MITRE ATT&CK	Syntax-based gen. better; generates error-free code
[67]	Software Dev./ Secure LLM Code Framework	LLMs + Static An.	Conceptual; Python Code	Framework combining LLMs and static analysis for secure code
[68]	Code Completion Sec./ Adv. Prompt Gen.	ADVPRO, Target LLMs (13)	CVE-Python Code prompts	Generates effective adversarial prompts for code LLMs
[56]	IoT Sec./ Vuln. Det.	GAN-based Fuzzer (DCGAN), LLM for Risk	IoT Net. Payloads	GAN for realistic IoT fuzzing; LLM prioritizes vulns
[69]	JS Engines / Compilers Bug Discovery	Syntax-aware NNLM (SNPM), Grey-box Fuzzing	JS Test Suites, CS code	Integrates grammar/coverage into NNLM fuzzer; found JS & CS bugs

tuning models remain a major challenge, just as sources of explainability and detection efficacy against changing adversarial tactics.

RL for attack path generation. RL techniques are gaining traction in modeling dynamic and strategic threat behaviors. By simulating adaptive attack strategies and defense mechanisms, RL frameworks contribute to more resilient and responsive CS systems. Specifically, offline RL is applied to design and refine defensive playbooks that counteract Advanced Persistent Threats (APTs) [74]. These models can continuously improve the quality of threat responses by learning from historical incidents and simulated adversarial behavior. While still in early stages of deployment, RL-based systems represent a forward-looking approach to attack path generation, vulnerability prioritization, and adaptive response planning. They enable the construction of AI-guided decision support systems that dynamically assess and mitigate risks in real-time.

AI-generated exploits vs. traditional methods. GenAI improves VA work processes automation through the exploitation of attack vectors for uncovering system weaknesses for the purposes of proactive RM [17,18,75]. Hilario et al. [75] show that LLMs such as ChatGPT improve penetration testing by adaptive learning in customized scenarios. Likewise, [17] applied IE-GAN for generating diverse SQL injection payloads for model diversity in training datasets, as well as detection frameworks for injection-based vulnerabilities. In the work by Lu et al. [18], GANs have been used to model network attacks in improving vulnerability exposure and mitigation planning under adversarial conditions. Yet, there still remain limitations. The security-focused LLMs, such as the RAG-enhanced fine-tuned Mistral 7B, could autonomously escalate privileges in a Linux environment, yet most times they generate inconsistent or non-executable commands, giving evidence to the need for high-grade training data [19]. Huang et al. [51] present a two-phase framework that integrates discovery of vulnerabilities with their automated remediation using LLMs showing improved coverage and cost-efficient remediation; yet its effectiveness is heavily dependent upon the quality, structure, and external source of knowledge of input data, thereby embedding challenges into a reliable and reproducible AI-driven paradigm for risk-mitigating actions.

5.1. Summary and emerging challenges

GenAI techniques have also found their way into exploitation and penetration testing to improve VA and RM, particularly models like GANs, LLMs, and RL. Such techniques push this domain further with GAN systems creating plausible SQL injections and CAN-bus attacks, platform-level threats analyzed by transformer-based models such as X-squatter, and automated vulnerability protection through LLM-based au-

Table 2
Summary of exploitation and penetration testing studies.

Ref.	Domain/Focus	Method	Dataset(s)	Contribution
[17]	Web Sec./SQL Inj. Gen.	IE-GAN	Custom SQL, sqli-lab	IE-GAN generates realistic SQLi, enhances detection
[70]	Auto. Sec./Stealthy CAN Attack	LSTM-GAN	Public CAN logs	GANs generate malicious IVN traffic bypassing IDS
[20]	Domain Squat./Sound-squat. Gen.	Transformer (X-squatter)	Brand names, TLS certs, PyPI	AI generates multilingual sound-squatting, reveals flaws
[71]	Phishing Email Det.	LLMs (12 Benchmarked)	Human/AI phishing emails	Open-source LLMs (Llama-3) excel for SME phishing det.
[72]	Inc. Resp. Time/Early Phishing Det.	Fine-tuned GPT	Implicit Email dataset	Fine-tuned GPTs outperform base models for early phishing det.
[73]	Phishing Awareness/Personalized Warnings	GPT-3.5, RLHF	Custom phishing emails, User studies	LLM + RLHF generates context-specific phishing warnings
[74]	Cyber Def./Adaptive Def. Playbooks	Offline RL	CyberVAN sim. (APT data)	Offline RL trains agents for adaptive defense against APTs
[75]	Pentesting	LLM (GPT 3.5)	VulnHub VM	GenAI offers useful suggestions across pentesting stages
[18]	Web Sec./SQL Inj. Gen.	DCGAN + Mutation	SQLi samples, sqli-lab	GANs generate diverse, usable SQLi for WAF testing
[19]	Pentesting/ Auto. Linux Priv. Esc.	Mistral 7B + RAG	Linux privesc articles, Debian VMs	RAG-augmented LLM performed autonomous privesc
[51]	Auto. Pentesting & Remediation	GPT-4/3.5 + RAG	CS books (KB), Metasploitable2	Integrated LLM framework automates pentesting/remediation

tonomous agents [17,20,70,72]. Attack-defense scenarios are modeled using RL techniques for real-time APT mitigation [74]. The summary of the studies can be found in Table 2. For the future, some of the issues needing research are adaptive LLMs and RL systems being trained by actual events to improve realism in exploitation and achieve expert-AI interactive frameworks for responsible deployment in VA and RM.

Limitations and critical evaluation. While GenAI can create realistic attacks and recognizes 15 % of sound-squatting TLS certificates, GANs-generated payloads lack production-level validation, which risks unrealistic scenarios. Additionally, commands produced for tasks such as Linux privilege escalation are inconsistent through LLM-based agents [17,19,20,70]. The fine-tuned LLMs can suffer from poor performance for phishing detection, because of limited dataset diversity, and the dual-use ethical issues raise the possibility of misuse [71,75]. High detection accuracy with good phishing (80 % precision) and better APT defense playbooks are reported in the literature, but again, controlled datasets like SQL-lab limit the real-world use of their results [17,73,74]. Payload realism could be ensured while improving detection robustness, as well as mitigating ethical discomfort; these included production environment validation, diversified datasets, access controls, and the use of red-team exercises. Now, all of these describe inconsistent outputs, which show a strong need for further empirical testing [17,71,75].

6. Threat modeling and attack simulation

Proactive threat modeling and attack simulation enable organizations to predict possible hazards, grasp attack paths, assess defenses, and train staff. GenAI introduces dynamic, realistic, and automated techniques for simulating and modeling threats.

NLP for attack scenarios. Recent advancements in GenAI trigger automated VA and adaptive RM, using sophisticated techniques for scenario generation. Ensemble methods [76] show how multiple generative models (ChatGPT, Llama) autonomously compose personalized anti-phishing training scenarios using a scenario vector database. They are evaluated through automated metrics (BLEU:Bilingual Evaluation Understudy, ROUGE:Recall-Oriented Understudy for Gisting Evaluation) and human-assessed metrics (feasibility, personalization). For dynamic threat environment RAG systems like [77] use structured prompt engineering to generate adaptive deception ploys, achieving 93 % engagement and 96 % accuracy against evolving malware threats, and enhancing risk mitigation. While, the work [78] harnesses RAG-enhanced LLMs to simulate new attack scenarios for push-testing security postures beyond traditional training boundaries. Ontology frameworks [79] conduct structured VAs by applying named entity recognition and graph comparison to threat intelligence, automatically generating exercise near-real-world scenarios Tactics, Techniques, and Procedures (TTPs). These methods enable a holistic risk assessment by increasing coverage through diverse synthetic attack variants, adapting to new patterns of threat emergence, and ensuring scenario relevance through structured knowledge grounding; however, challenges exist in addressing hallucinations and ensuring the generated content alignment with the current attack surface.

AI-driven cyberattack graphs. GenAI and NLP are reshaping VA by automating the conversion of attack trees into comprehensive attack-defense trees [80]. The authors used LLMs to suggest countermeasures for identified threats, thus addressing some important gaps in manual mitigation planning. Integrated defenses generated by LLMs with secondary mitigation data yield results comparable to associate-level security engineers while avoiding redundant mitigations. New metrics evaluate the semantic correctness and completeness of these countermeasures, thereby supporting their early-stage evaluation in defense strategy. This approach reduces expensive reengineering by allowing identification of the most cost-effective mitigations up front. When evolutionary models based on CAPEC and CVE datasets are integrated, in conjunction with static graph analysis for validation, this generated

framework enables robust automated assessment and enhancement of threat coverage-smoothing the process of RM along the system lifecycle.

Hybrid simulation environments. Different environments have established effectiveness ratings of GAN in VA. For instance, Dynamic Adaptive Threat Simulation GAN (DATS-GAN) [81] counteracts DDoS and spoofing attacks in IoT/WSN environments, thus improving detection and reducing response times by 30–40 %. Conditional GANs (CGANs) [82] generate deceptive network topologies to prevent reconnaissance attacks while maintaining 99.2 % availability for Industrial Control Systems (ICS). In addition, privacy-preserving synthetic data generation is important for risk-aware testing; transformer-based models [83] hold accuracy levels up to 92 % for the ICMP protocols, while Conditional VAE (CVAE)-based methods [84,85] combine differential privacy with generative modeling to produce anonymized datasets for intrusion detection, with 95 % detection effectiveness preserved. Simulations in domains make visible other looming vulnerabilities among others: IoT time-series generators [86] develop synthetic sensor data capable of exposing weaknesses in smart environments, whereas naval radar GANs [87] are used to create adversarial Inverse Synthetic Aperture Radar (ISAR) images to spoof maritime navigation, having greater than 0.85 Structural Similarity Index (SSIM) scores. Therefore, these approaches provide full-cycle VA by (1) simulating advanced attack vectors, (2) ensuring privacy in testing, and (3) exposing environment-specific weaknesses, even though challenges persist in keeping semantic consistency across complex protocols and preventing technology misuse.

Real-world case studies. NLP-driven threat assessments in healthcare infrastructure (implantable devices, biobanks) validate GenAI's capability to uncover domain-specific vulnerabilities [88]. LLM-generated threat scenarios in [79] and [78] enhance VA realism by enabling red and blue teams to identify, prioritize, and respond to high-risk attack vectors under expert-validated conditions. ChatGPT's analysis of real honeypot logs confirms its efficacy in automating log-based vulnerability detection [89]. Also, CyberWheel's use of GenAI to train autonomous defense agents contributes to adaptive RM by enabling real-time identification and mitigation of evolving vulnerabilities in complex environments [90]. Studies on social network defense (CIC-DDoS2019 datasets: a DDoS attack dataset) [91] and naval radar spoofing using NATO data [87] further illustrate how GenAI facilitates realistic, context-aware VAs.

6.1. Summary and emerging challenges

GenAI is being applied for threat modeling and attack scenario simulation to help support VA and RM, with particular attention to NLP and GANs, such as GAN-based simulation models for IoT and ICS, and LLM generating attack-defense graphs from CAPEC and CVE datasets [76,80–82]. The case studies in healthcare, defense, and social networks testify their usefulness in red-blue team collaboration [78,79,87–91]. Table 3 summarizes these studies. Further paths may include improving multimodal data usage and analyst-software interaction to enhance scenario fidelity and proactive defense planning.

Limitations and critical evaluation. While such GenAI techniques can deliver reductions of about 30 to 40 % in DDoS response times, hallucinated scenarios do not produce authentic instances, and GANs offer semantically unreal protocols for IoT/ICS-inhibiting their training effectiveness [76,81]. The ontology frameworks are also subject to scaling issues due to their computational demands, which are often combined with redundant mitigations in LLM-generated attack-defense trees and sacrifices in accuracy in privacy-preserving models for anonymity [79,80,84]. Moreover, despite the high effectiveness, simulated datasets such as CIC-DDoS2019 limit real-world applicability [76,82,84]. Human-in-the-loop validation, protocol-specific training, efficient algorithms, and differential privacy could improve realism, scalability, and accuracy in scenarios, though more longitudinal studies will be required to confirm performance in practice [76,81,84].

7. Risk evaluation and decision support

Effective vulnerability management depends highly on accurate risk assessment, prioritization of threats, and informed decision-making. GenAI and ML technologies have greatly impacted the organization, automation, and enhancement of effectiveness in these areas. Here, we analyze how GenAI, especially LLMs and NLP methods, can be used in risk scoring, vulnerability-related report generation, forecasting emerging exposures from historical vulnerability patterns, and providing support in decision-making.

Automated risk scoring and prioritization. Automated VA and RM have been improved in efficiency and accuracy by GenAI. Fine-tuning BERT-based models on NVD (National Vulnerability Database) data seems to return an accuracy of greater than 90 % in severity classification, thus accelerating analysis and prioritization [92]. Hybrid models that combine the generative power of GPT with the contextual understanding of BERT also improved the classification of vulnerability descriptions and their corresponding scoring systems [48]. The Universal Sentence Encoder (USE) models predict CVSS scores with an accuracy in the range of 72–77 %, outperforming human experts in speed and consistency [60]. CNNs coupled with Product Hygiene Index (PHI) metrics deliver an F1-score of 0.82 in predicting exploitability for high-severity vulnerabilities using product-specific vulnerability histories [49].

Real-time dynamic threat conditions are modeled with CORAL (a specific risk assessment tool) which provides an attack graph which, thus, facilitates continuous risk assessment as the attack surface changes [50]. GAN-enhanced Neuro-Fuzzy Systems (a hybrid AI approach combining neural networks and fuzzy logic) sharpen risk profiling through pattern recognition in IEEE IoT datasets [93]. AgraBOT is RAG-based for automating vendor risk analysis and achieves 0.85 F1-score on SOC 2 reports, which saves evaluation time from days to minutes [28]. Taxonomies of GenAI risks therefore provide structured frameworks for identifying and mitigating root causes [29]. These innovations will cause a paradigm shift in vulnerability management. They will (1) automate labor-intensive scoring tasks, (2) combine context and historical data into assessments, and (3) respond in real time to emerging threats, although the interpretation and recognition of novel patterns of vulnerability still present challenges.

AI-powered report generation and summarization. LLMs are shown effective in dealing with unstructured CS data for VA and RM. CYGENT's [94] GPT-3.5 turbo, for example, achieves 97 % BERTscore accuracy in generating summaries of complex logs into actionable insights for the purpose of rapid threat detection. CANAL [95] establishes that fine-tuned BERT models produce a better performance than larger LLMs in real-time threat alerting from news feeds-namely a cost-effective option to track emerging vulnerabilities. FAIL [96] uses LLMs, for retrospective analysis, to cluster incidents and extract facts from software failure reports with an F1 score of 90 %, identifying over 2400 unique failures from large corpuses. The LLMs with knowledge graphs are integrated-through the techniques such as Quantized Low-Rank Adaptation (QLoRA) fine-tuning on the models like Llama 2 and Mistral 7B [97]. Visualization tools based models like GPT-4 and GPT-4o show a noticeable correspondence to human decisions in producing risk dashboards for smart cities and industrial systems [98]. In all, they provide applications that advance the RM paradigm through automating raw data transformation, allowing real-time threat awareness, and amplifying risk communication-though there are still difficulties mainly regarding domain-specific terminology and scale consistency.

Table 3
Summary of threat modeling and attack simulation studies.

Ref.	Domain/Focus	Method	Dataset(s)	Contribution
[76]	Auto. Scenario Gen.	Ensemble of LLMs	Scenario vector DB	Ens. LLMs automates personalized anti-phishing training
[77]	Adaptive Deception	RAG systems	Evolving malware threats	Generates adaptive deception ploys
[78]	Cyber Scenario Gen.	LLMs (GPT), RAG	Org. functions, CS Handbook	LLMs/RAG generate complex CS exercise scenarios
[79]	Structured Cyber Exercise Gen.	NLP, Text Gen.	Public CS articles, MITRE ATT&CK	ML/NLP structures info for cyber exercise scenarios
[80]	Threat Modeling/ A-D Trees	LLMs, NLP	Sys. specs, Attack Trees	LLMs automate countermeasures and Attack-Defense trees
[82]	ICS Sec. / Deceptive Net. Topologies	CGAN	Simulated ICS	CGANs dynamically generate defensive topologies for ICS
[83]	Network Traffic Gen.	GPT-3, LLM Chaining	Real network traffic	LLM framework (PAC-GPT) generates network traffic
[84]	NIDS Privacy-Preserving Training Data	CVAE	CIC-IDS2017, CSE-CIC-IDS2018	CVAE generates effective, private synthetic traffic
[85]	Synthetic/ Privacy-preserving Data	Fine-tuned LLMs	Iris, Healthcare, Finance, Cyber	LLM framework with DP generates synthetic data
[87]	Naval Radar Sec./ Deceptive ISAR	GANs	Real ISAR (NATO SET-196)	GANs generate deceptive ISAR images
[88]	Healthcare Threat intel. Extraction	NLP (BERT-NER), KBs	CS news, Healthcare scenarios	NLP automates healthcare threat assessment
[89]	Honeypot Log Analysis/ Explaining logs	LLM (ChatGPT)	Real honeypot logs	ChatGPT explains attacks, maps to ATT&CK
[90]	High-fidelity Training Env. for Auto Def.	Sim./Emul. Framework	Simulated network	High-fidelity env. (Cyberwheel) for training auto defense agents
[91]	Social Net. Def./ DDoS det.	Swarm OpenAI LLM	CIC-DDoS2019	CNN-LSTM for DDoS detection in social networks

Predictive models for vulnerability-driven risk forecasting. Beyond current-state assessment, GenAI is being used to forecast future risks by modeling patterns from historical vulnerability data and known attack vectors. It has been shown by studies that LLMs, especially BERT-like models, are very loving for predicting vulnerabilities and predicting exposure in the context of fine-tuning and retrieval-augmented generation [99]. The analysis of 12,582 global patents using BERT-based topic modeling has helped in spotting emerging threats and security gaps across domains [100]. For IoT environments, privacy-preserving BERT models with embedding perturbation deliver an accuracy rate of 91.2% using the Edge-IIoTset dataset while maintaining data privacy constraints [101,102].

Beyond LLMs, GANs paired with Neuro-Fuzzy systems generate realistic attack scenarios and augment limited datasets, improving detection accuracy through fuzzy rule-based training [93]. AutoML (Automated Machine Learning) with VAEs has fully automated the entire VA pipeline-data balancing and model ensembling-having shown quite strong results on CICIDS2017 and 5G-NIDD datasets [103]. Together, these methods will enhance predictive VA by combining analytics, privacy, and automation.

Regulatory compliance and enterprise security use cases. GenAI can automate gap analyses and risk evaluation for compliance and Third-Party Risk Assessment(TPRA) activities. The CS AI-based Data-Driven Integrated Environment (CADDIE) model, the fine-tuned BERT-GRC (Governance, Risk, and Compliance) model with RAG, scans documents for misalignments between internal policies and external regulations, extracting compliance requirements from structured knowledge bases and facilitating high-confidence gap detection and proactive remediation [30]. Again, using multi-stage processing by LLM, AgraBOT analyzes reports and security questionnaires, compressing the evaluation time from days to minutes [28]. In sum, these tools convert unstructured regulatory and vendor data into actionable vulnerability intelligence, reinforcing AI-compliance and TPRA workflows.

7.1. Summary and emerging challenges

Similar to the transformative impact of GenAI in different stages of VA/RM lifecycle we discussed, GenAI is also impacting risk evaluation and decision-making processes [4,92]. Respecting this, predictive models like USE and CNNs and hybrid models of GPT-BERT could improve CVSS predictions and dynamic threat emulation [48–50,60,93]. Moreover, log summarization will be carried out in real time via LLMs while cognitive graphs will be utilized for decision-making, and privacy-aware models for IoT security [94,97,101]. Future efforts should thus be targeted at AutoML scalable implementations and risk taxonomies integrated to enrich governance and compliance in frameworks such as GDPR [28,103]. Studies are summarized in Table 4.

Limitations and critical evaluation. Although the literature states BERT-based models and AgraBOT's high accuracy for severity classification and vendor risk analysis, dependence on structured NVD data restricts their application to unstructured sources and novel vulnerabilities [28,60,92]. Template-sensitive tools like AgraBOT fail on non-standard reports and the lack of explainability in LLMs stands in the way of regulatory compliance [28,48]. Hybrid GPT-BERT models improved scoring,

Table 4
Summary of risk evaluation and decision support studies.

Ref.	Domain/Focus	Method	Dataset(s)	Contribution
[92]	Vuln. Mgmt/ Severity Classif.	BERT-based models	NVD	BERT models achieve >90 % accuracy in severity classification
[48]	Vuln. Mgmt/ Desc.	Hybrid (GPT + BERT)	NVD	Hybrid models improve vulnerability classification and scoring
[60]	Vuln. Mgmt/ Auto. CVSS Scoring	USE, GPT, SVM	CVE descriptions	AI models automate CVSS scoring with high accuracy
[49]	Vuln. Mgmt/ CVE exploit.	CNNs, Word Embed., PHI	NVD, ExploitDB, CVEDetails	CNNs automate severity prediction; exploitability score outperforms NVD
[50]	Container Sec./ Online risk assessment	Attack Graphs	Container topology, Vulns	Efficient attack graph for near real-time container risk assessment
[93]	Risk Classif.	Neuro-Fuzzy, GANs	IEEE IoT	Neuro-fuzzy improves risk classification in GAN-based IDS
[28]	Automating TPSRM	RAG (IR, LLMs, ranking)	Real TPSRM assessments	RAG (AgraBOT) speeds up TPSRM assessments with high accuracy
[94]	Log Analysis / SecOps	LLM (Fine-tuned GPT-3)	Log files, Manual summaries	Fine-tuned GPT-3 excels at human-readable log summarization
[95]	CTI/Threat Model./ Cyber activity alert.	Fine-tuned BERT (CANAL)	CS news articles	Fine-tuned BERT (CANAL) is cost-effective for cyber threat alerts
[96]	Software Failure Analysis	LLM-based pipeline	News articles (137k +)	LLM pipeline (FAIL) automates large-scale software failure analysis
[97]	CTI Extraction/KM	LLMs, KG, Fine-tuning	Unstructured CTI texts	Fine-tuned LLMs effectively extract CTI triples for KG construction
[98]	Security Viz. Eval.	LLMs (Bing Chat, ChatGPT-4/4o)	Web security dashboards	LLMs provide valuable insights for security dashboard usability
[99]	Evaluating LLM for IDS	LLMs, Prompting, RAG	Proprietary commercial data	Compares LLM techniques for IDS using real-world data
[100]	Threat Forecasting	BERTopic	Telehealth security patents	BERT topic modeling reveals telehealth security trends
[101]	IoT Threat Detection/ Privacy	BERT, Differential Privacy	Edge-IIoTset	BERT with DP achieves high accuracy for IoT threat detection
[102]	IoT Threat Det.	BERT (SecurityBERT), PPFLE	Edge-IIoTset	Lightweight BERT with PPFLE for IoT threat detection
[103]	Auto. IDS via AutoML	AutoML (incl. TVAE)	CIC-IDS-2017	AutoML framework automates IDS creation with high performance
[30]	Reg. Comp./Risk Mgmt	BERT-GRC, RAG	Internal policies, Regs. (GDPR)	RAG-enhanced BERT detects compliance gaps

and CYGENT is reported to have high accuracy in log summarization, but controlled datasets like Edge-IIoTset, hamper validation in real-time scenarios [48,94,101]. Enhancements to applicability, explainability, and robustness can be achieved by using preprocessing for unstructured data, with XAI techniques such as SHAP, adaptive templates, and validations conducted for real-time threat feeds. There also needs to be cross-domain work focusing more on generalization [28,60,92].

8. Vulnerability remediation

Vulnerability remediation is a core step following VA, aiming to reduce risk through patching, configuration, or mitigation. Traditionally, this has been a very laborious activity, but GenAI opens exciting and promising paths for automating and enhancing all aspects of remediation-from generating code patches to recommending configuration changes and optimizing workflows for security operations.

Automated patch generation. Automated Program Repair (APR) strives to generate patches for software vulnerabilities automatically to shrink the critical gap between discovery and remediation, and LLMs appear promising for partial or even complete automation of such a solution. In this respect, the ContrastRepair system [32] demonstrates this capability by using ChatGPT with contrastive test pairs to locate the root causes of flaws better. In fact, this method surpassed existing SOTA techniques on well-known standard benchmarks such as Defects4J, QuixBugs, and HumanEval-Java, illustrating how advanced prompting strategies improve LLMs with respect to code repair. These movements probably mark the beginning of future integration of GenAI into vulnerability management pipelines for better patching and remediation implementation.

AI-driven configuration and policy recommendations. For vulnerability management, security settings and compliance with benchmarks are very important, with LLMs showing some prospects as automatic rule generators, although their performance highly depends on the quality of fine-tuning. Louro et al. [104] give a systematic assessment of how LLMs can generate security rules and reports that base models falter the majority of time. This is while fine-tuning grants 89 % accuracy for simple rule translations, 61 % for cross-prompt integrations, and 79 % for the creation of more complicated rules, allowing pre-remediation hardening. Advanced LLM agents, such as those discussed in Hu et al. [105], integrate vulnerability reports and patterns of malicious traffic to develop rules that are both comprehensive and generalized,

Table 5
Summay of vulnerability remediation studies.

Ref.	Domain/Focus	Method	Dataset(s)	Contribution
[32]	Auto. Prog. Repair/ Patch Gen.	LLM (ChatGPT), Contrastive tests	Defects4J, QuixBugs, HumanEval-Java	ContrastRepair improves LLM-based APR
[104]	Auto. Rule Gen.	Fine-tuning LLMs (Chat Bots)	Eval. of 3 fine-tuning approaches	Fine-tuning improves LLM ability to generate correct rules
[105]	Auto. Rule Gen.	LLM-based agent	Exploit-DB, CSIC 2010, Public IDS Rules	LLM agent effectively generates & generalizes IDS rules
[31]	GRC policy Gen.	GPT-4 for policy gen.	Trend Micro, SANS, NIST, NCSC	GPT-4 can outperform human policies for ransomware mitigation
[106]	Secure code (RBAC)	LLMs (ChatGPT, Bard, CoPilot), Formal Specs (JML)	JML, OpenJML RAC/ESC	Human-AI pipeline builds secure software adhering to RBAC
[61]	Auto. cloud SecOps	GANs, Transformers	Cloud Activity Logs, Threat Intel. Feeds	GenAI for faster/better cloud threat handling
[59]	DevSecOps/ Auto. Threat Discovery	LLM for design threat discovery	Use case (retail); User stories	Proposed LLM & Chaos Eng. for early threat/flaw detection
[53]	Cloud-Native Sec./ Verify Pipelines	LLM (Llama2) for PBOM (Pipeline Bill of Materials), Blockchain/NFTs	Pull request details, SBOM info.	DevSec-GPT for verifiable container security
[52]	Vuln. Remed./ Optimize Remed.	LLM (GPT-4) support	Field study	LLM collab. reduced remediation time & boosted engagement

bridging the gap between AI-driven detection and the interpretable systems. In GRC applications, GPT-4, given guided prompts and human oversight, generates ransomware mitigation policies that are greater in both completeness and effectiveness than those created by humans, thereby addressing configuration vulnerabilities that are most exploited during data exfiltration [31]. These tools take preventive aspects of vulnerability management further by turning complex defense knowledge into standards-compliant, implementable policies.

AI-augmented security orchestration and response (SOAR). Security Orchestration and Automated Response (SOAR) platforms increasingly incorporate GenAI to streamline vulnerability-focused workflows. LLMs support VA and RM by enabling intelligent log and alert triage, assisting in correlating vulnerabilities with asset criticality and exposure, and recommending context-aware remediation actions. In cloud environments, GenAI models can analyze configuration drift and log anomalies to prioritize unpatched vulnerabilities and misconfigurations [61]. Tools like DevSec-GPT extend these capabilities by generating Pipeline Bills of Materials to identify and remediate vulnerable dependencies [53,59]. LLMs can also explain complex vulnerabilities and associated remediation options in a human-interpretable manner [52], contributing to transparency and trust in automated response systems. However, explainability and decision traceability remain key challenges, especially in high-stakes environments requiring human oversight and risk accountability.

AI-driven vs. human-driven remediation. GenAI is accelerating vulnerability management with automated policy development, rule formulation, and code repairs. A study shows that GPT-4 formulates extensive ransomware mitigation policies beyond human-written ones, but still requires expert supervision [31]. Regarding the rule generation, the utilization of LLM agents that assimilate information from vulnerability reports and security rules can accelerate the generation of generalized Snort (Network Intrusion Detection and Prevention System) rules, enhancing the threat detection abilities beyond manual techniques [105]. In terms of code repair, ContrastRepair augments LLMs with contrastive test pairs so that ChatGPT can fix 143 out of 337 bugs in Defects4J, thus achieving state-of-the-art performance [32]. To enhance reliability, coupling LLMs with formal methods such as Java Modeling Language (JML) specifications and automated verification tools contributes to secure generation and validation of Role-based Access Control (RBAC) policies [106]. All these combined advances make the otherwise automated, AI-assisted vulnerability remediation systematic and reliability concerns are addressed.

8.1. Summary and emerging challenges

GenAI automates vulnerability remediation using LLMs for patch generation, configuration synthesis, and governance policies. For instance, tools such as DevSec-GPT automated cloud dependency analysis, and SOAR integration improved alert triage [31,32,52,59,104]. In this regard, Table 5 gives a brief summary of the studies presenting applications of GenAI in mitigating vulnerabilities. Future work needs to look at the integration of LLMs into the DevSecOps pipeline and a combination of GenAI with formal verification in order to improve technical fidelity and transparency in these remediation workflows.

Limitations and critical evaluation. Although models like ContrastRepair and fine-tuned LLMs showed effectiveness, continuous integration and continuous delivery (CI/CD) integration is missing from the system, limiting its scalability, and LLMs are reported to have failed 89 % of rule-generation tasks without fine-tuning [32,104]. Moreover, human oversight is required to audit policies generated by GPT-4. On the other hand, SOAR-integrated LLMs provide no explanation, and tools like DevSec-GPT are dependent on the quality of the input data fed to them [31,52,59]. Though outperforming human-generated ransomware policies and speeding up remediation by LLMs, other existing benchmarks like Defects4J fail to provide evidence of any validation at production scale [31,52]. CI/CD integration can advance automated fine-tuning, human-AI collaboration, and formal verification like JML for better scalability, reliability, and transparency, though evidence from case studies at a production scale will be required to ascertain this [32,104,106].

9. Comparative evaluation of GenAI tools

Table 6 presents a comparative evaluation of several GenAI tools across various security tasks and VA/RM stages. Despite the table showing GenAI's potential for transformation in different parts of the VA/RM

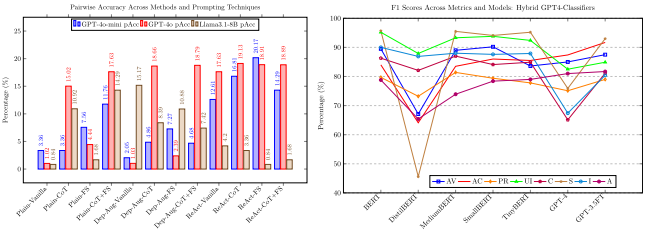
Table 6
Comparative evaluation of GenAI tools in different VA/RM stages.

Ref.	Stage	GenAI	Domain	Performance	Limitations	XAI/Trust	Concern
[45]	Disc.	LLM	Contracts	High Precision	No large code test	No explainability	Hallucination; logic gaps
[62]	Disc.	LLM (GPT3)	Contracts	Improced OWASP Detection	Not real-time	No transparency	Prompt injection; LLM bias
[63]	Disc.	LLM; RAG	OSS Code	Better Localization	FPs in new domains	Weak human validation	Low adversarial robustness
[57]	Fuzz.	GAN; LLM	IoT	Improved IoT Fuzzing	Weak semantic checks	Not interpretable	Unrealistic test cases
[69]	Fuzz.	NNLM	JS Engines	Improved Semantic Error	Needs target tuning	Grammar-aware traceability	Overfitting; generalization issues
[17]	Exploit	GAN	SQLi-Lab	Realistic SQLi Inputs	Not tested on prod	Opaque GAN behavior	Adversarial misuse
[20]	Exploit	Trans.	TLS, PyPI	Detects 15 % active squats (2×)	No prod eval.	No interpretability	Abuse if misused
[28]	Risk Eval.	RAG; LLM	SOC2, Vendor	F1 = 0.85; Time Efficient	Template-sensitive	Explainability not evaluated	Lower hallucination due to context
[94]	Log Sum.	LLM (GPT3.5)	Sec Logs	97 % BERTScore	No robustness tests	Black-box summaries	Misinterpretation risk
[32]	Remed.	LLM (GPT)	Defects4J, QuixBugs	Fixed 143/337 Bugs	No CI/CD integration	Contrastive tests aid audit	Patch security unverified
[53]	Remed./ SOAR	LLM; SBOM; PBOM	Cloud Pipelines	Verifiable Patch Automation	Lacks benchmarks	Blockchain-backed logs	Incomplete compliance check

pipeline such as smart contract analyses, OWASP detection, and generating effective patches, each of these works carries its own subtle limitations. The major critical limitations faced by studies here, in short, comprise one or more of the issues like depending on controlled datasets (e.g., FormAI-v2, SQL-lab)[17,63]; lack of matching production-scale validation [20,53]; non-consistent detection accuracies and false-positive incidence in new domains [63]; weak explainability among LLMs and GANs [28,45]. Hallucination and regulatory non-compliance, adversarial misuse, and overfitting are all risks underlying the critical need for rigorous validation, explainability, and real-world testing to ensure sound and trustworthy GenAI deployment [57,94].

9.1. Benchmarking GenAI tools for VA/RM

Different GenAI models are benchmarked here, particularly with respect to LLMs like GPT-4o and GPT-4o-mini, ReAct Agents, Llama, and bidirectional models (BERT variants) and their applications in VA/RM. This is for identifying vulnerabilities and severity scoring, as well as threat prioritization based on CVSS metrics and interprocedural analysis. In JITVUL, ReAct agents leverage on-demand retrievals and generally improve pairwise accuracy over plain LLMs, with ReAct + CoT on GPT4-o (resulting in a pAcc of around 19.13) and ReAct + FS on GPT-4o-mini (with a pAcc of 20.17) showing to be the strongest variants. This can highlight the benefits of interprocedural context, while pointing inconsistent reasoning across vulnerable and benign pairs [107]. On the other hand, in CVSS scoring [48], hybrid models like GPT + BERT pipelines outperform single-family baselines, resulting in multiple per-metric F1 exceeding 90 %, while the CVSS components experienced a mean F1 of around 0.868 in hybrid architecture, against 0.842 on GPT-4-only and about 0.783 on BERT-only architecture. The findings show that GPT can help in producing standardized and information-rich description, while BERT excels at stable multi-class labeling of CVSS components when fed those improved descriptions. Overall, JIT agents benefit from agentic, interprocedural retrieval, and hybrids are strongest for CVSS component inference. However, there are still open problems such as robustness and generalization issues (Fig. 4).



(a) (b)

Fig. 4. a) Performance of GenAI models on JIT vulnerability detection (from JITVUL Benchmark), b) Performance of GenAI GPT-4-enhanced models on severity assessment.

10. Open challenges and future research directions

The integration of GenAI into VA and RM presents both opportunities and significant challenges. Ethical considerations, adversarial robustness, explainability, and integration with existing vulnerability frameworks are critical research areas to ensure safe and effective adoption.

Ethical considerations. Concerns have been raised with the rapid adoption of GenAI for VA and RM about fairness, transparency, and an alignment with human values. Adversarial testing reveals weaknesses in LLM detection under perturbed inputs for issues of ethics and security [8]. Also, foundational taxonomies [46,108] have outlined 22 types of risks-from bias to malicious use-that should be considered in a security system. Recent studies [109,110] elaborately depict the gaps in ethical

assessment, reading into real-world context and involvement from stakeholders, warning against systemic risks of perception bias and market manipulation, among others. These risks are, however, taken care of by newly emerging frameworks [111,112], all calling for ongoing ethical evaluations even though they are faced with challenges such as maintaining a particular balance between explainability and privacy. The human-centric framework further stresses inputs by multi-stakeholders and post-audit tools toward maintenance of human rights by GenAI-driven security workflows. Altogether, these will point to the necessity for balanced governance towards utility and ethical integrity in AI-based VAs.

Adversarial attacks on AI models. GenAI models used for vulnerability detection and risk assessment remain vulnerable to adversarial manipulations. The authors in [25] show that Adaptive Gradient-based Word Embedding Perturbations (AG-WEP) could fool CNN-based classifiers with semantically intact spam inputs and therefore display weaknesses in text-based systems. More powerful threats such as compound data poisoning [26] achieve 90 % attack success by mixing label flipping and adversarial artifacts and, while just poisoning 0.5 % of the training data, pose rather stealthy threats to transformer models. Prompt injection vulnerabilities are just as widespread: [27] used a hybrid XLM-RoBERTa + Bi-GRU model to detect Thai-language instruction attacks with 96.52 % accuracy, while [7] showed that vision-language models in oncology could be compromised with sub-visual prompting by means of medical images. In response, [113] introduced a method for adversarial training based on XAI, combining Fast Gradient Sign Method (FGSM) perturbations with SHAP explanations to increase the robustness and accuracy of models on malware and intrusion detection tasks. All these studies indicate an immediate need for resilient training methods and input sanitization measures for AI-driven security systems.

Explainability and trust in AI-driven CS. Effective VA and RM requires that AI model decisions be interpretable and trustworthy. Example of the application of SHAP explanations on a hybrid CNN-BiLSTM intrusion detection model (95.28 % accuracy) was shown to reveal essential feature patterns in Layer-2 attacks [47]. To address the limitations of generic XAI metrics, [114] proposes coherence-based inter-model assessment for alignment of domain knowledge and explanation achieves 86 % correctness in automated rule validation via chain-of-thought prompting. Much of what is perceived to influence human trust in AI was analyzed in a study by 243 security professionals by [115]: trust in GenAI for threat intelligence relies more on perceived ability, integrity, and benevolence than the risk factors. Risk assessment is further complicated by the observation by [116] that GPT now has less information regarding model development and explainability in its documentation. Collectively these studies show while XAI is viewed to increase interpretability, effective RM requires transparent design, domain-informed validation, and continuous human-AI partnership.

Integration with existing security frameworks. Integrating GenAI into vulnerability management does put forward challenges regarding how well traditional security frameworks adapted to AI-specific risks. The author of [117] points to how current paradigms like CVE struggle with flaws in ML systems, particularly when academic findings fail to make bridges to operational security teams. Such is also the case with supply chain security[118]; advocates a three-tier regulatory model for Large GenAI Models (LGAIMs), distributing charges across the AI value chain and making clear, particular recommendations for high-impact use cases. Similarly, there is [119], which calls for lifecycle audits and operational mandates that translate abstract AI policies into actionable security practices, borrowing from software engineering to fill governance gaps. Such studies emphasize the necessity of updating security processes and regulations related to generative models from a technical point of view. There are much more future works required in this domain for setting the standard to manage future AI security vulnerability issues effectively.

Reducing AI model bias in VA. Mitigating biases in GenAI models used for vulnerability and risk evaluation is critical to avoid skewed assessments that could impact decision quality. Continuous Ethical Assessments (CEAs) and human-centered engineering approaches provide pathways to reduce bias and improve fairness in automated vulnerability management [112,120]. In this respect, a CEA framework is proposed by the study [112]. This refers to the automating bias testing together with human audits throughout the AI lifecycle applying Security-integrated DevOps (SecDevOps) principles applied in bias detection and mitigation processes upfront before deployment in the VA systems. At the same time, [120] proposes a human-centered framework incorporating multi-stakeholder round input and post-audit tools to examine fairness and alignment in vulnerability scoring driven by LLMs. In unison, these approaches highlight the need for continuous validation and governance mechanisms specially designed for VA and RM workflows.

Adaptation in enterprise settings. The embedding of the GenAI-driven VA/RM tools of operations in the enterprise framework would require integration with the existing security framework, as well as aligning with stakeholder requirements. A phased-in approach works best; enterprises might want to kick off with proof-of-concept projects such as using ContrastRepair for automated patch generation [32] or using DevSecGPT for automating dependency analysis in cloud-native DevSecOps pipelines [53]. Links with Security Information and Event Management (SIEM) and SOAR systems are important, where LLM can improve log triage such as CYGENT in summarizing logs [94] and prioritizing vulnerabilities by asset criticality [61]. Compatibility is ensured by aligning with enterprise frameworks like NIST CSF or MITRE ATTCK [50], allowing for integration with existing workflows without much disruption or reorienting to meet regulatory compliance in real-time threat response like General Data Protection Regulation (GDPR) and New York Codes, Rules, and Regulations, part 500 (NYCRR 500) [30].

Targeted adoption strategies for organizations must be designed around barriers including cultural resistance, skill gap, and financial apprehension. Human-in-the-loop frameworks can be implemented to allow analysts to approve of AI-assisted findings so that transparency can be guaranteed in cases like those of LLM-generated remediation plans [52]. Training programs integrating interactive dashboards will help fill the skill gaps for analysts and compliance teams to evaluate GenAI outputs responsibly [98]. For cost-sensitive organizations, such fine-tuned models as CANAL might provide economical alerting options for SMEs [95], while larger companies could tap into AI services off-cloud. A structured adoption framework-from a needs-assessment stage to pilot testing, integration roadmaps, training stakeholders, and using performance metrics like MTTR reduction-ensures smooth deployment. For instance, in healthcare, companies have deployed NLP-driven threat assessments to improve time for vulnerability detection [88]. Social media companies use swarm-based LLMs for DDoS detection [91]. Such mechanisms will allow the enterprises to harness the transformation potential of GenAI in their regions while remaining conscious of operational and regulatory constraints.

10.1. Summary

The true promise of GenAI in VA and RM is challenged by ethical issues such as bias propagation, invasion of privacy, and misalignment with human values, thus demanding robust governance frameworks that may ensure fairness and transparency [8,46,108–112,120]. GenAI models for VA and risk scoring are themselves vulnerable to threats of evasion, data poisoning, and prompt injection that compromise assessment reliability, thus making adversarial robustness critical in this respect [7,25–27]. The future research should therefore put more emphasis on the development of resilient training paradigms, input sanitization methods, and architecture-level robustness to protect AI-powered VA systems against malicious attacks [113].

Table 7
Discussion over specific research challenges.

Challenge	Description	Potential Solution
Ethical Concerns	Concerns regarding fairness, transparency, bias propagation, privacy concerns, and alignment with human values in automated decision-making by GenAI models for VA and RM.	Ongoing ethical evaluations, human-centric frameworks incorporating multi-stakeholder input, post-audit tools, balanced governance frameworks, and CEA frameworks with automated bias testing and human audits.
Adversarial Attacks on GenAI Models	GenAI models are vulnerable to adversarial manipulations such as evasion via adversarial examples, data poisoning, and prompt injection, compromising model reliability and decision integrity.	Resilient training methods, input sanitization measures, and architecturally resilient designs to protect AI assessment systems.
Explainability and Trust	The need for model decisions in VA and RM to be interpretable and trustworthy. Limitations of generic XAI metrics and the importance of domain expertise in interpreting results.	XAI techniques (like SHAP, EBM) for interpretability, transparent design, domain-informed validation, continuous human-AI partnership, and human-AI coordination frameworks for critical evaluation and accountability.
Integration with Existing Security Frameworks	Challenges in adapting traditional security frameworks (like CVEs) to AI-specific risks, supply chain security risks for GenAI models, and the need for new standards and auditing processes.	Updating security processes and regulations for generative models, standardized protections, balanced regulatory frameworks, lifecycle audits, and operational mandates that translate abstract GenAI policies into actionable security practices.
Reducing Model Bias in VA	Mitigating biases in GenAI models used for vulnerability and risk evaluation to avoid skewed assessments that could impact decision quality and fairness.	CEA, human-centered engineering approaches, CEA framework with automated bias testing and human audits, and multi-stakeholder input for fairness and alignment in vulnerability scoring.

Explainability and trust are the other fundamental aspects. While methods such as SHAP or EBM can be helpful in understanding the black-box of AI models, they require domain knowledge for a reliable risk contextualization, particularly in security tasks [47,114]. Human-AI coordination frameworks are also necessary for better transparency and accountability [115,116]. As it is normally challenging to align with security standards such as CVEs, especially when it comes to threat scenarios in the supply chain, this demands developing AI-specific auditing frameworks [117–119]. Also, bias mitigation in AI algorithms is essential to prevent skewed assessments, calling for continuous ethical assessments and human-centered designs so that fairness and reliability in GenAI-driven VA and RM, in particular, are assured [112,120]. Table 7 summarizes the discussed challenges existing in the current adaptation of GenAI techniques for VA and RM, with their potential solutions.

11. Answering research questions

Here the focus shifts again back to the research questions that guided this review toward answering them with evidence from the reviewed literature, clearly at the interface of GenAI and CS, with clear emphasis on VA and RM.

(RQ1) How GenAI is transforming the end-to-end process of VA and RM?. Gen-AI is transforming VAs and RMs in an automated solution lifecycle covering processes from vulnerability identification to remediation. The LLMs like the GPT4 or CodeLlama examine source code and binary and configuration scripts with precision in invoking flaws [53,58]. The fuzzers based on GANs generate semantically rich payloads, defeating the traditional random fuzzers for the discovery of deeply hidden/complex vulnerabilities in IoT, CAN bus, and ICS [81]. In RM, the RAG-augmented LLMs help build dynamic attack models and provide near-real-time risk scoring by generating evolving attack graphs and prioritizing threats based on asset criticality [50,61]. Neuro-Fuzzy hybrid models improve risk scoring through the integration of historical data and probabilistic reasoning [99]. GenAI tools like ContrastRepair shorten patching latency, while SOAR-integrated LLMs recommend remediation based on real-time contextualization of the threats-a change that has flipped the entire VA/RM process from a statically reactive workflow to that of continuous predictive cycle [32,94].

RQ2) How does our taxonomy clarify GenAI's dual use in the context of VA and RM? . Taxonomy shows GenAI's dual-use ability, respecting offensive and defensive applications of LLMs, GANs, VAEs, and RL models in VA/RM. For example, an offensive descriptor would be automated finding of vulnerabilities, generation of exploits, and evasive tactics (prompt injection) by GenAI [17,20]. The same family of models, when subverted, provides the basis for advanced defensive mechanisms such as fuzz testing with GANs, triage of vulnerabilities with LLMs, and the creation of attack-defense trees [50,81]. These dynamics will help researchers and practitioners explore the benefit-risk trade-offs as they deploy GenAI in high-stakes security contexts. Scenarios of dual risk also emerge from the taxonomy: the instruments intended to be used proactively for RM can also be misused in nefarious ways-for example, an LLM-based patch suggestion might be hijacked for malicious use [32].

RQ3) Can GenAI overcome the key prioritization and remediation challenges within VA and RM in dynamic and evolving threat landscapes?. Yes, GenAI can uniquely address the real-time nature of today's threat landscapes. BERT-CNN hybrids and GAN-based risk engines process real-time CVE feeds and generate context-aware risk scores that consider exploitability, asset exposure, and environmental context [47,95]. LLMs with attack graphs allow adaptive threat modeling that can evolve with TTPs [50]. The remediation is becoming very fast as well. The tools like DevSec-GPT automate the patching process, dependency tracing, and configuration hardening [53]. GenAI could, for example, translate complex vulnerabilities into actionable security rules or policies compliant with GRC frameworks [30]. The disadvantages are hallucination and overfitting; thus, there is a constant need for retraining, fine-tuning to the domain, and human validation to ensure trustworthiness [52,114].

RQ4) What standard of explainability is needed to ensure the trustworthiness of GenAI in VA?. Explainability in GenAI assisted VA is more than just technical transparency-it should contain traceability that matches the regulations and possess business relevance [47]. Feature attribution methods like SHAP and EBMs explain the used input elements (derived from CVSS vectors and code segments) to conclude the judgment. The narrative generation tools link technical outputs with the clear risk-oriented language framing them in a business conference [98]. The credibility of the AI system's outputs has to be auditable. This entails logging the earlier version of the model, a list of prompt structures, and the rationale for the output [114]. The gold standard is a post-hoc XAI tracing model audits combined with interactive dashboards that enable real-time RM analysts to interrogate and validate AI outputs.

RQ5) What types of cyberattacks are the focus of the GenAI researchers?. GenAI research in VA and RM takes on advancing cyberattacks to expose vulnerabilities in traditional automated methods for vulnerability discovery, risk scoring, and mitigation. Recent models such as LLMs,

GANs, and RL simulate high-impact cyberattacks to strengthen CS resilience. For instance, LLMs and GANs can simulate SQL injection and API abuse to generate adaptive payloads and improve vulnerability detection and risk scoring based on exploitability [17,20]. GANs are used to simulate spoofed sensor data to test IoT/ICS vulnerabilities, thereby helping in the prioritization of patches and segmentation of networks [81]. LLMs generate phishing content to examine human vulnerabilities and improve risk assessments and training [71]. In addition, multi-stage advanced persistent threats are modeled using LLMs and RL to dynamically update the attack-defense tree based on current threat intelligence [74]. GenAI also leverages compliance documents and honeypot data to inform on supply chain risk and emerging TTPs [30,89]. Besides simulation, GenAI can improve the precision of VA so that vulnerabilities are discovered, prioritized, and dynamically remediated effectively.

12. Conclusion and future directions

Conclusion. With LLMs, GANs, and RL, Generative AI (GenAI) transforms vulnerability assessment (VA) and risk management (RM) by automating and enriching the security lifecycle. Complex vulnerabilities in software, IoT, and cloud systems are detected with improved accuracy by means of semantic code analysis, realistic exploit generation, and adversarial attack simulation of these models. Techniques such as real-time CVE parsing and threat graphs enable fine-grained risk prioritization, while GenAI-driven remediation using automated code repair, rule generation, and SOAR integration reduces Mean Time to Response (MTTR). This survey demonstrates GenAI's dual-use nature, where tools that contribute to enhanced VA/RM, can craft phishing lures or give rise to evasive APT strategies, thus demanding solid explainability, adversarial robustness, and ethical safeguards. Under this thrust, the taxonomy and systematic review delineate the application of GenAI across the themes of vulnerability discovery, threat modeling, risk evaluation, and remediation while pointing out the gaps within model interpretability, human-in-the-loop workflows, and regulatory alignment.

Future direction. GenAI's prospective research for VA/RM should focus on four broad areas. The first is that continuous learning should incorporate live threat intelligence (for example, zero-day exploits, supply-chain indicators) without destabilizing the model. Second, explainability should become more evolved beyond post-hoc visualizations, such as interpretable by design architectures, causal inference, and auditability systems for analyst validation and regulatory compliance. Third, hybrid human-AI frameworks that allow for real-time analyst feedback and risk threshold calibration, especially in high-stakes environments, would be an important innovation. GenAI's ethics-related issues would require federated learning, differential privacy, and model watermarking in order to preserve data confidentiality and system integrity. Standardized evaluation frameworks for benchmarking GenAI tools for trustworthiness, strength, explainability, and operational readiness are also required with metrics for offensive realism (payload diversity, evasiveness) and defensive accuracy (patch precision, false-positive rate). Directions along these lines will protect the reliability and ethical introduction of GenAI into next-generation CS operations.

CRediT authorship contribution statement

Seyedeh Leili Mirtaheri: Writing – original draft, Resources, Conceptualization; **Narges Movahed:** Writing – original draft, Visualization, Validation, Investigation; **Reza Shahbazian:** Writing – review & editing, Writing – original draft, Visualization, Resources; **Valerio Pascucci:** Writing – review & editing, Resources; **Andrea Pugliese:** Writing – review & editing, Supervision, Funding acquisition.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. One of the authors has received financial support from PNRR MUR project FAIR (PE-0000013). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU. One of the authors has received financial support from PNRR MUR project FAIR (PE-0000013).

References

[1] B. Guembe, A. Azeta, S. Misra, V.C. Osamor, L. Fernandez-Sanz, V. Pospelova, The emerging threat of AI-driven cyber attacks: a review, Appl. Artif. Intell. 36 (1) (2022) 2037254.
[2] S. Kumar, M. Mahajan, S. Batra, A recent study of machine learning based techniques for the detection of cyber-attacks on web applications, in: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), 6, IEEE, 2023, pp. 153–158.
[3] F. Quader, V.P. Janeja, Insights into organizational security readiness: lessons learned from cyber-attack case studies, J. Cybersecurity Priv. 1 (4) (2021) 638–659.
[4] S.L. Mirtaheri, A. Pugliese, Leveraging generative AI to enhance automated vulnerability scoring, in: 2024 IEEE Conference on Dependable, Autonomic and Secure Computing (DASC), IEEE, 2024, pp. 57–64.
[5] D. Noever, Can large language models find and fix vulnerable software?, arXiv preprint arXiv:2308.10345 (2023).
[6] H. Pearce, B. Tan, B. Ahmad, R. Karri, B. Dolan-Gavitt, Examining zero-shot vulnerability repair with large language models, in: 2023 IEEE Symposium on Security and Privacy (SP), IEEE, 2023, pp. 2339–2356.
[7] J. Clusmann, D. Ferber, I.C. Wiest, C.V. Schneider, T.J. Brinker, S. Foersch, D. Truhn, J.N. Kather, Prompt injection attacks on vision language models in oncology, Nat. Commun. 16 (1) (2025) 1239.
[8] K. Liu, Y. Li, L. Cao, D. Tu, Z. Fang, Y. Zhang, Research of multidimensional adversarial examples in LLMs for recognizing ethics and security issues, in: International Conference on Computer Science and Education, Springer, 2023, pp. 286–302.
[9] F. Heiding, S. Katsikeas, R. Lagerström, Research communities in cyber security vulnerability assessments: a comprehensive literature review, Computer Sci. Rev. 48 (2023) 100551.
[10] S. Lipp, S. Banescu, A. Pretschner, An empirical study on the effectiveness of static C code analyzers for vulnerability detection, in: Proceedings of the 31st ACM SIGSOFT international symposium on software testing and analysis, 2022, pp. 544–555.
[11] G. Bella, P. Biondi, S. Bognanni, S. Esposito, Petiot: penetration testing the internet of things, Internet of Things 22 (2023) 100707.
[12] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, Y. Zhong, Vuldeepecker: A deep learning-based system for vulnerability detection, arXiv preprint arXiv:1801.01681 (2018).
[13] P. Parkar, A. Bilimoria, A survey on cyber security IDS using ML methods, in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 352–360.
[14] P. Mell, K. Scarfone, S. Romanosky, Common vulnerability scoring system, IEEE Sec. Priv 4 (6) (2006) 85–89.
[15] MITRE, Common Vulnerabilities and Exposures, . [n. d.]. Retrieved from https://cve.mitre.org/.
[16] M. Alawida, A.E. Omolara, O.I. Abiodun, M. Al-Rajab, A deeper look into cybersecurity issues in the wake of covid-19: a survey, J. King Saud Univ.-Comput. Inf. Sci. 34 (10) (2022) 8176–8206.
[17] M. Xu, B. Xie, F. Cui, C. Jin, Y. Wang, SQL injection attack sample generation based on IE-GAN, in: 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE, 2023, pp. 1014–1021.
[18] D. Lu, J. Fei, L. Liu, Z. Li, A GAN-based method for generating SQL injection attack samples, in: 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 10, IEEE, 2022, pp. 1827–1833.
[19] J. Gregory, Q. Liao, Autonomous cyberattack with security-augmented generative artificial intelligence, in: 2024 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2024, pp. 270–275.
[20] R.V. Valentim, I. Drago, M. Mellia, F. Cerutti, X-squatter: AI multilingual generation of cross-language sound-squatting, ACM Trans. Priv. Sec. 27 (3) (2024) 1–27.

[21] N. Tihanyi, T. Bisztray, M.A. Ferrag, R. Jain, L.C. Cordeiro, How secure is AI-generated code: a large-scale comparison of large language models, Emp. Software Eng. 30 (2) (2025) 1–42.

[22] Y. Fu, P. Liang, A. Tahir, Z. Li, M. Shahin, J. Yu, J. Chen, Security weaknesses of copilot generated code in github, arXiv preprint arXiv:2310.02059 (2023).

[23] Z. Mousavi, C. Islam, K. Moore, A. Abuadbba, M.A. Babar, An investigation into misuse of java security apis by large language models, in: Proceedings of the 19th ACM Asia Conference on Computer and Communications Security, 2024, pp. 1299–1315.

[24] S. Hamer, M. d'Amorim, L. Williams, Just another copy and paste? Comparing the security vulnerabilities of ChatGPT generated code and StackOverflow answers, in: 2024 IEEE Security and Privacy Workshops (SPW), IEEE, 2024, pp. 87–94.

[25] J. Gregory, Q. Liao, Adversarial spam generation using adaptive gradient-based word embedding perturbations, in: 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), IEEE, 2023, pp. 1–5.

[26] E. Begoli, M. Mahbub, L. Passarella, S. Srinivasan, A compound data poisoning technique with significant adversarial effects on transformer-based sentiment classification tasks, ACM J. Data Inf. Qual. 16 (4) (2024) 1–15.

[27] V. Vajrobol, B.B. Gupta, A. Gaurav, Thai-language chatbot security: detecting instruction attacks with XLM-RoBERTa and Bi-GRU, Comput. Electr. Eng. 116 (2024) 109186.

[28] M. Toslali, E. Snible, J. Chen, A. Cha, S. Singh, M. Kalantar, S. Parthasarathy, AgraBOT: Accelerating third-party security risk management in enterprise setting through generative AI, in: Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, 2024, pp. 74–79.

[29] H. Tanaka, M. Ide, J. Yajima, S. Onodera, K. Munakata, N. Yoshioka, Taxonomy of generative AI applications for risk assessment, in: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, 2024, pp. 288–289.

[30] B. Lodge, RAGe against the machine with BERT for proactive cybersecurity posture, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 3579–3588.

[31] T. McIntosh, T. Liu, T. Susnjak, H. Alavizadeh, A. Ng, R. Nowrozy, P. Watters, Harnessing GPT-4 for generation of cybersecurity GRC policies: a focus on ransomware attack mitigation, Comput. Sec. 134 (2023) 103424.

[32] J. Kong, M. Cheng, X. Xie, S. Liu, X. Du, Q. Guo, Contrastrepair: enhancing conversation-based automated program repair via contrastive test case pairs, arXiv preprint arXiv:2403.01971 (2024).

[33] S. Ankalaki, A.A. Rajesh, M. Pallavi, G.S. Hukkeri, T. Jan, G.R. Naik, Cyber attack prediction: from traditional machine learning to generative artificial intelligence, IEEE Access 13 (2025) 44662–44706.

[34] A.H. Salem, S.M. Azzam, O.E. Emam, A.A. Abohany, Advancing cybersecurity: a comprehensive review of AI-driven detection techniques, J Big Data 11 (1) (2024) 105.

[35] J. Yang, X. Chen, S. Chen, X. Jiang, X. Tan, Conditional variational auto-encoder and extreme value theory aided two-stage learning approach for intelligent fine-grained known/unknown intrusion detection, IEEE Trans. Inf. Forensics Secur. 16 (2021) 3538–3553.

[36] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014) 2672–2680.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008.

[38] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[39] A. Ding, G. Li, X. Yi, X. Lin, J. Li, C. Zhang, Generative artificial intelligence for software security analysis: fundamentals, applications, and challenges, IEEE Software 41 (6) (2024) 46–54.

[40] R. Kaur, T. Klobučar, D. Gabrijelčič, Harnessing the power of language models in cybersecurity: a comprehensive review, Intern. J. Inf. Managem. Data Insights 5 (1) (2025) 100315.

[41] A. Golda, K. Mekonen, A. Pandey, A. Singh, V. Hassija, V. Chamola, B. Sikdar, Privacy and security concerns in generative AI: a comprehensive survey, IEEE Access 12 (2024) 48126–48144.

[42] S.M. Taghavi Far, F. Feyzi, Large language models for software vulnerability detection: a guide for researchers on models, methods, techniques, datasets, and metrics, Int. J. Inf. Secur. 24 (2) (2025) 78.

[43] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, Y. Zhang, A survey on large language model (llm) security and privacy: the good, the bad, and the ugly, High-Confid. Comput. 4 (2024) 100211.

[44] F. Charmet, H.C. Tanuwidjaja, S. Ayoubi, P.-F. Gimenez, Y. Han, H. Jmila, G. Blanc, T. Takahashi, Z. Zhang, Explainable artificial intelligence for cybersecurity: a literature survey, Ann. Telecommun. 77 (11) (2022) 789–812.

[45] Y. Sun, D. Wu, Y. Xue, H. Liu, H. Wang, Z. Xu, X. Xie, Y. Liu, Gptscan: detecting logic vulnerabilities in smart contracts by combining GPT with program analysis, in: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, 2024, pp. 1–13.

[46] C. Bird, E. Ungless, A. Kasirzadeh, Typology of risks of generative text-to-image models, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, pp. 396–410.

[47] I.F. Kilincer, Explainable AI supported hybrid deep learnig method for layer 2 intrusion detection, Egyptian Inf. J. 30 (2025) 100669.

[48] S.L. Mirtaheri, A. Pugliese, N. Movahed, R. Shahbazian, A comparative analysis on using GPT and BERT for automated vulnerability scoring, Intell. Syst. Appl. 26 (2025) 200515.

[49] A. Okutan, M. Mirakhorli, Predicting the severity and exploitability of vulnerability reports using convolutional neural nets, in: Proceedings of the 3rd international workshop on engineering and cybersecurity of critical systems, 2022, pp. 1–8.

[50] D. Tayouri, O.S. Cohen, I. Maimon, D. Mimran, Y. Elovici, A. Shabtai, CORAL: Container online risk assessment with logical attack graphs, Comput. Secur. 150 (2025) 104296.

[51] J. Huang, Q. Zhu, Penheal: a two-stage LLM framework for automated pentesting and optimal remediation, in: Proceedings of the Workshop on Autonomous Cybersecurity, 2023, pp. 11–22.

[52] X. Wang, Y. Tian, K. Huang, B. Liang, Practically implementing an LLM-supported collaborative vulnerability remediation process: a team-based approach, Comput. Secur. 148 (2025) 104113.

[53] E. Bandara, P. Foytik, S. Shetty, A. Hassanzadeh, Generative-AI (with custom-trained meta's Llama2 LLM), blockchain, NFT, federated learning and PBOM enabled data security architecture for metaverse on 5G/6G environment, in: 2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS), IEEE, 2024, pp. 118–124.

[54] E. Bandara, S. Shetty, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, K. De Zoysa, N.W. Keong, DevSec-GPT - generative-AI (with custom-trained meta's Llama2 LLM), blockchain, NFT and PBOM enabled cloud native container vulnerability management and pipeline verification platform, in: 2024 IEEE Cloud Summit, 2024, pp. 28–35. https://doi.org/10.1109/Cloud-Summit61220.2024.00012

[55] V. Kouliaridis, G. Karopoulos, G. Kambourakis, Assessing the effectiveness of LLMs inandroid application vulnerability analysis, in: International Conference on Attacks and Defenses for Internet-of-Things, Springer, 2024, pp. 139–154.

[56] M.T. Masud, N. Koroniotis, M. Keshk, B. Turnbull, S.K. Kermanshahi, N. Moustafa, Generative fuzzer-driven vulnerability detection in the internet of things networks, Appl. Soft Comput. 174 (2025) 112973.

[57] S. Arjun, D. Majumdar, A.Z. Nabin, A.K. Sharma, S.M. Raheman, Beyond copy-pasting-contextualizing LLMs for secure code generation, in: International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS, Springer, 2024, pp. 185–200.

[58] R. Yıldırım, K. Aydın, O. Çetin, Evaluating the impact of conventional code analysis against large language models in API vulnerability detection, in: Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference, 2024, pp. 57–64.

[59] M. Bedoya, S. Palacios, D. Díaz-López, E. Laverde, P. Nespoli, Enhancing DevSecOps practice with large language models and security chaos engineering, Int. J. Inf. Secur. 23 (2024) 3765–3788.

[60] Z. Zhang, V. Kumar, B. Pfahringer, A. Bifet, Ai-enabled automated common vulnerability scoring from common vulnerabilities and exposures descriptions, Int. J. Inf. Secur. 24 (1) (2025) 1–20.

[61] A. Patel, P. Pandey, H. Ragothaman, R. Molleti, D.R. Peddinti, Generative AI for automated security operations in cloud computing, in: 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), IEEE, 2025, pp. 1–7.

[62] B. Boi, C. Esposito, S. Lee, VulnHunt-GPT: a smart contract vulnerabilities detector based on OpenAI chatGPT, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, 2024, pp. 1517–1524.

[63] Y. Wu, M. Wen, Z. Yu, X. Guo, H. Jin, Effective vulnerable function identification based on CVE description empowered by large language models, in: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, 2024, pp. 393–405.

[64] N. Tihanyi, T. Bisztray, R. Jain, M.A. Ferrag, L.C. Cordeiro, V. Mavroeidis, The formai dataset: generative AI in software security through the lens of formal verification, in: Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering, 2023, pp. 33–43.

[65] C. Tony, N.E.D. Ferreyra, M. Mutas, S. Dhiff, R. Scandariato, Prompting techniques for secure code generation: a systematic investigation, arXiv preprint arXiv:2407.07064 (2024).

[66] A.-G. Sîrbu, G. Czibula, Automatic code generation based on abstract syntax-based encoding. application on malware detection code generation based on MITRE ATT&CK techniques, Expert Syst. Appl. 264 (2025) 125821.

[67] A. Kavian, M.M. Pourhashem Kallehbasti, S. Kazemi, E. Firouzi, M. Ghafari, LLM security guard for code, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, 2024, pp. 600–603.

[68] X. Li, G. Meng, S. Liu, L. Xiang, K. Sun, K. Chen, X. Luo, Y. Liu, Attribution-guided adversarial code prompt generation for code completion models, in: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, 2024, pp. 1460–1471.

[69] H. Xu, Y. Wang, Z. Jiang, S. Fan, S. Fu, P. Xie, Fuzzing javascript engines with a syntax-aware neural program model, Comput. Secur. 144 (2024) 103947.

[70] F. Merola, Q. Ahmed, et al., GAttack: generative attack on in-vehicle network, in: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2024, pp. 2556–2561.

[71] J. Zhang, P. Wu, J. London, D. Tenney, Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: a comprehensive analysis, IEEE Access 13 (2025) 28335–28352.

[72] A.B. Beydemir, U. Sezgin, U. Doğan, B.E. Aşıklar, F.A. Yerlikaya, Ş. Bahtiyar, A dynamically selected GPT model for phishing detection, in: 2024 14th International Conference on Advanced Computer Information Technologies (ACIT), IEEE, 2024, pp. 481–484.

[73] Q.H. Nguyen, T. Wu, V. Nguyen, X. Yuan, J. Xue, C. Rudolph, Utilizing large language models with human feedback integration for generating dedicated warning

for phishing emails, in: Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems, 2024, pp. 35–46.
[74] A. Wei, D. Bierbrauer, E. Nack, J. Pavlik, N. Bastian, Offline reinforcement learning for autonomous cyber defense agents, in: 2024 Winter Simulation Conference (WSC), IEEE, 2024, pp. 1978–1989.
[75] E. Hilario, S. Azam, J. Sundaram, K. Imran Mohammed, B. Shanmugam, Generative AI for pentesting: the good, the bad, the ugly, Int. J. Inf. Secur. 23 (3) (2024) 2075–2097.
[76] J.Y. Park, T.-S. Kim, An automated scenario generation model for anti-phishing using generative AI, in: 2025 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, 2025, pp. 368–370.
[77] S. Ahmed, A.M. Rahman, M.M. Alam, M.S.I. Sajid, SPADE: enhancing adaptive cyber deception strategies with generative AI and structured prompt engineering, in: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2025, pp. 01007–01013.
[78] M.M. Yamin, E. Hashmi, M. Ullah, B. Katt, Applications of LLMs for generating cyber security exercise scenarios, IEEE Access 12 (2023) 143806–143822.
[79] A. Zacharis, C. Patsakis, AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition, Int. J. Inf. Secur. 22 (5) (2023) 1333–1354.
[80] A. Birchler De Allende, B. Sultan, L. Apvrille, From attack trees to attack-defense trees with generative AI & natural language processing, in: Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems, 2024, pp. 561–569.
[81] P.K. Rao, S. Chatterjee, P.S. Prakash, K.S. Ramana, Adaptive cyber defence: leveraging GANs for simulating and mitigating advanced network attacks in IoT environments, in: International Symposium on Applied Computing for Software and Smart Systems, Springer, 2024, pp. 309–322.
[82] X. Qin, F. Jiang, X. Qin, L. Ge, M. Lu, R. Doss, CGAN-based cyber deception framework against reconnaissance attacks in ICS, Comput. Netw. 251 (2024) 110655.
[83] D.K. Kholgh, P. Kostakos, PAC-GPT: a novel approach to generating synthetic network traffic with GPT-3, IEEE Access 11 (2023) 114936–114951.
[84] G. Aceto, F. Giampaolo, C. Guida, S. Izzo, A. Pescapè, F. Piccialli, E. Prezioso, Synthetic and privacy-preserving traffic trace generation using generative AI models for training network intrusion detection systems, J. Netw. Comput. Appl. 229 (2024) 103926.
[85] M. Goyal, Q.H. Mahmoud, An LLM-based framework for synthetic data generation, in: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2025, pp. 00340–00346.
[86] D. Gkoulis, Creating interpretable synthetic time series for enhancing the design and implementation of internet of things (IoT) solutions, Internet of Things 30 (2025) 101500.
[87] G. Meucci, B. Karahoda, A.H. Oveis, F. Mancuso, E. Jajaga, A. Cantelli-Forti, Naval cybersecurity in the age of AI: deceptive Isar images generation with GANs, in: 2023 IEEE 48th Conference on Local Computer Networks (LCN), IEEE, 2023, pp. 1–6.
[88] S. Silvestri, S. Islam, D. Amelin, G. Weiler, S. Papastergiou, M. Ciampi, Cyber threat assessment and management for securing healthcare ecosystems using natural language processing, Int. J. Inf. Secur. 23 (1) (2024) 31–50.
[89] M.B. Ozkok, B. Birinci, O. Cetin, B. Arief, J. Hernandez-Castro, Honeypot's best friend? Investigating ChatGPT's ability to evaluate honeypot logs, in: Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference, 2024, pp. 128–135.
[90] S. Oesch, A. Chaulagain, B. Weber, M. Dixson, A. Sadovnik, B. Roberson, C. Watson, P. Austria, Towards a high fidelity training environment for autonomous cyber defense agents, in: Proceedings of the 17th Cyber Security Experimentation and Test Workshop, 2024, pp. 91–99.
[91] M. Nadeem, C. Hongsong, Protecting social networks against dual-vector attacks using swarm openAI, large language models, swarm intelligence, and transformers, Expert Syst. Appl. 278 (2025) 127307.
[92] S.L. Mirtaheri, A. Pugliese, N. Movahedkor, A. Majd, Advanced automated vulnerability scoring: improving performance with a fine-tuned BERT-CNN model, in: 2024 11th International Symposium on Telecommunications (IST), IEEE, 2024, pp. 109–113.
[93] M. Torres, S.M. Errapotu, V. Gonzalez, Cyberattack risk classification for detecting intrusions through N euro-fuzzy inference based generative adversarial networks, in: 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET, IEEE, 2024, pp. 1–6.
[94] P. Balasubramanian, J. Seby, P. Kostakos, Cygent: a aybersecurity conversational agent with log summarization powered by Gpt-3, in: 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), IEEE, 2024, pp. 1–6.
[95] U. Patel, F.-C. Yeh, C. Gondhalekar, Canal-cyber activity news alerting language model: empirical approach vs. expensive LLMs, in: 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC), IEEE, 2024, pp. 1–12.
[96] D. Anandayuvaraj, M. Campbell, A. Tewari, J.C. Davis, FAIL: analyzing software failures from the news using LLMs, in: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, 2024, pp. 506–518.
[97] R. Fieblinger, M.T. Alam, N. Rastogi, Actionable cyber threat intelligence using knowledge graphs and large language models, in: 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2024, pp. 100–111.
[98] H. Zhao, B. Silverajan, Evaluating cyber security dashboards for smart cities and buildings: enhancing user modeling with LLMs, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–10.
[99] M.-T. Bui, M. Boffa, R.V. Valentim, J.M. Navarro, F. Chen, X. Bao, Z.B. Houidi, D. Rossi, A systematic comparison of large language models performance for intrusion detection, Proceedings of the ACM on Networking 2 (CoNEXT4) (2024) 1–23.
[100] U.H. Govindarajan, D.K. Singh, H.A. Gohel, Forecasting cyber security threats landscape and associated technical trends in telehealth using bidirectional encoder representations from transformers (BERT), Computers Secur. 133 (2023) 103404.
[101] T.N. Shree, K. Sundarakantham, J. Dhivya, B. Gayathri, Detecting cyber threat using Gen-AI, in: 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), IEEE, 2025, pp. 845–851.
[102] M.A. Ferrag, M. Ndhlovu, N. Tihanyi, L.C. Cordeiro, M. Debbah, T. Lestable, N.S. Thandi, Revolutionizing cyber threat detection with large language models: a privacy-preserving bert-based lightweight model for iot/iiot devices, IEEE Access 12 (2024) 23733–23750.
[103] L. Yang, A. Shami, Towards autonomous cybersecurity: an intelligent autoML framework for autonomous intrusion detection, in: Proceedings of the Workshop on Autonomous Cybersecurity, 2023, pp. 68–78.
[104] B. Louro, R. Abreu, J. Cabral Costa, João B.F. Sequeiros, Pedro R M. Inácio, Analysis of the capability and training of chat bots in the generation of rules for firewall or intrusion detection systems, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–7.
[105] X. Hu, H. Chen, H. Bao, W. Wang, F. Liu, G. Zhou, P. Yin, A LLM-based agent for the automatic generation and generalization of IDS rules, in: 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE, 2024, pp. 1875–1880.
[106] C.E. Rubio-Medrano, A. Kotak, W. Wang, K. Sohr, Pairing human and artificial intelligence: enforcing access control policies with LLMs and formal specifications, in: Proceedings of the 29th ACM Symposium on Access Control Models and Technologies, 2024, pp. 105–116.
[107] A. Yildiz, S.G. Teo, Y. Lou, Y. Feng, C. Wang, D.M. Divakaran, Benchmarking LLMs and LLM-based agents in practical vulnerability detection for code repositories, arXiv preprint arXiv:2503.03586 (2025).
[108] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, in: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022, pp. 214–229.
[109] Y. Lyu, Y. Du, The ethical evaluation of large language models and its optimization, AI and Ethics (2025) 1–14. https://link.springer.com/article/10.1007/s43681-024-00654-9
[110] Y. Zhang, Z. Zhu, Z. Xu, Study on derived risks and its governances of\" social embedding\" of artificial intelligence algorithm, in: Proceedings of the 2023 International Conference on Artificial Intelligence, Systems and Network Security, 2023, pp. 331–336.
[111] S. Berning, V. Dunning, D. Spagnuelo, T. Veugen, J. Van Der Waa, The trade-off between privacy & quality for counterfactual explanations, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–9.
[112] G. Steinke, R. LaBrie, S. Sarkar, Recommendation for continuous ethical analysis of AI algorithms, in: Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference, 2022, pp. 104–106.
[113] A.-E. Malik, G. Andresini, A. Appice, D. Malerba, An XAI-based adversarial training approach for cyber-threat detection, in: 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), IEEE, 2022, pp. 1–8.
[114] A. Alnahdi, S. Narain, Towards transparent intrusion detection: a coherence-based framework in explainable AI integrating large language models, in: 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), IEEE, 2024, pp. 87–96.
[115] V. Ramteke, D. Pramod, K.P. Patil, Is artificial intelligence trustworthy? An empirical investigation to adopt generative AI for cyber threat intelligence using valence framework, in: Congress on Smart Computing Technologies, Springer, 2023, pp. 53–65.
[116] Z. Xu, E. Mustafaraj, Tracing the evolution of information transparency for openAI's GPT models through a biographical approach, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 2024, pp. 1684–1695.
[117] J.M. Spring, A. Galyardt, A.D. Householder, N. VanHoudnos, On managing vulnerabilities in AI/ML systems, in: Proceedings of the New Security Paradigms Workshop 2020, 2020, pp. 111–126.
[118] P. Hacker, A. Engel, M. Mauer, Regulating chatGPT and other large generative AI models, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 1112–1123.
[119] L. Lucaj, P. Van Der Smagt, D. Benbouzid, Ai regulation is (not) all you need, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1267–1279.
[120] B. Fernandez Nieto, Toward a human-centered framework for trustworthy, safe and ethical generative artificial intelligence: a multi-level analysis of large language models social impact, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, 2024, pp. 505–509.