© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718467

AI-Assisted Vulnerability Discovery and Reporting: Reliability Challenges, Security Risks, and Future Directions

SHAIK MOHAMMAD YASIN[1], KALISETTI VENKATESH[2], HEEMA CHETRI[3]
[1, 2, 3] *Department of Computer Science and Engineering, Aurora Deemed to be University, Hyderabad, India*

Abstract- ***In the contemporary world of cyber security, AI plays a significant part and is extensively recognized in finding and reporting vulnerabilities. Due to recent advancement in Machine Learning (ML), Deep Learning (DL) and Large Language Models (LLMs), various automated tools for finding vulnerabilities in software, reporting security bugs, and suggest patches for them are currently available. This work aims at identifying existing literature on AI assisted vulnerability detection and reporting system with a special focus on their security and reliability issue. Works from year 2023 to 2026 are reviewed and analyzed with categorizing them on the basis of their aim, techniques employed, merits and demerits. From our analysis, key reliability issues identified are; false positives, false negatives, hallucinations, inaccurate severity rating, and complexity in interpretations whereas the major security threats encountered were prompt injection attacks, adversarial manipulation of input data, poisoning data and leaking sensitive information. Based on the findings of this study, it is concluded that while there are numerous benefits to the AI assisted cyber security system, humans can never be out of the loop and supervision and interpretation of security flaws based on credibility cannot be disregarded in practical applications. We highlight some major challenges in current approaches along with several future research direction in order to achieve a dependable and secure AI-assisted vulnerability assessment system.***

Index Terms- ***Artificial Intelligence, Cyber Security, Vulnerability Discovery, Vulnerability Reporting, Trustworthy AI, Large Language Models.***

I. INTRODUCTION

The explosive growth of software systems, cloud platforms, and digital infrastructure has elevated cybersecurity challenges. As organizations embrace more digital technologies, software vulnerabilities and security threats continue to proliferate, and vulnerability discovery and reporting becomes crucial in identifying weaknesses before they can be exploited [1, 2].

Manual methods of vulnerability assessment, such as code reviews, penetration testing, and security audits, are efficient but often time consuming, demand substantial resources and expert knowledge. In the wake of ever-increasing volumes of software applications and security alerts, the need for automated approaches to aid cybersecurity professionals is gaining momentum [1, 2].

Recent advances in Artificial Intelligence (AI), such as machine learning, deep learning, and Large Language Models (LLMs), have led to new approaches for AI-driven vulnerability detection and reporting. AI assisted tools can be used to analyze source code for vulnerabilities, generate reports, prioritize findings, and propose fixes, increasing speed and scalability [3, 4].

Nevertheless, several concerns exist around the trustworthiness of these systems. Researchers have reported that they are prone to false positives, missed vulnerabilities, hallucinations, inconsistent risk rankings, lack of explanation, susceptibility to adversarial attacks and prompt injection, as well as privacy issues [4, 8, 13].

To that end, this paper aims to offer a structured literature review on the reliability and security issues related to AI-assisted vulnerability discovery and reporting, identify research gaps, and propose potential avenues for future investigation in trustworthy AI-enabled cybersecurity systems. Specifically, this paper aims to:

1. Survey approaches to AI assisted vulnerability discovery and reporting.
2. Assess the reliability and trustworthiness of AI based vulnerability assessments.
3. Identify key security risks of AI assisted cybersecurity systems.
4. Analyze existing research gaps in the current body of work.
5. Recommend future research directions to enhance trustworthy AI based vulnerability assessment frameworks.

This review will be guided by the following research questions:
RQ1: How reliable are AI-assisted vulnerability discovery and reporting systems at finding and documenting software vulnerabilities?
RQ2: What are the security risks and challenges of AI based vulnerability assessments and reports?
RQ3: Which factors influence the trustworthiness of AI-assisted vulnerability discovery and reporting systems?
RQ4: What research directions can improve the trustworthiness and robustness of AI-enabled security solutions

II. LITERATURE REVIEW

A. Evolution of AI-Assisted Vulnerability Discovery

Initially, finding vulnerabilities in software involved manual code reviews, penetration testing, and the application of rule-based static analysis tools. These methods, still crucial in cybersecurity practice, typically demand extensive expertise, time, and processing resources. This and the sheer volume of software vulnerabilities today led to investigations into how Artificial Intelligence (AI) could help in this discovery process [1], [2].

The impact of Machine Learning (ML) and Deep Learning (DL) in analyzing software vulnerabilities is undeniable. AI approaches can automatically analyze source code to recognize patterns indicative of vulnerabilities and categorize these security weaknesses without relying on static, rule-based definitions. Shimmi et al. [1] noted promising results in using AI-based vulnerability detection systems that were capable of discovering flaws while lessening the manual workload. Similarly, Malkawi et al. [2] found that AI-assisted vulnerability detection could enhance the efficiency of security practices by speeding up analysis and providing intelligent decision support.

More recent research focusing on transformer architectures and Large Language Models (LLMs) has enhanced the potential of AI in discovering vulnerabilities. These models excel over previous ML-based techniques as they can process the source code contextually and even generate detailed security analyses. In various studies, LLMs have proven effective for classifying vulnerabilities, reviewing code, and providing security recommendations, thus improving the vulnerability management process [10], [11]. Furthermore, the adoption of generative AI has contributed to automatic reasoning capabilities, which were difficult to achieve with traditional security tools [3], [4].

However, researchers continue to identify limitations inherent in AI-assisted vulnerability discovery. The quality of datasets, reproducibility of results, model transparency, and evaluation consistency remain major challenges across existing research [1], [13]. These challenges question the trustworthiness and reliability of AI-generated vulnerability findings in critical security situations.

B. AI-Assisted Vulnerability Reporting

The automation of vulnerability reporting and security documentation is another area where AI is gaining traction. Reporting a vulnerability involves not only identifying it but also detailing its severity, suggesting ways to fix it, and communicating this information to relevant parties. These steps have traditionally been manual processes, often requiring significant time and skill from cybersecurity professionals.

The recent advancements in generative AI and LLMs allow for automatic generation of reports and recommendations. Literature has shown that LLMs can convert technical vulnerability findings into readable and structured reports, making communication more effective and reporting times faster [3], [4]. This ability has proved invaluable for managing numerous vulnerabilities found in large-scale software environments.

AI-assisted reporting systems have also been shown to aid in prioritizing vulnerabilities, summarizing findings, and recommending remediation steps [2], [9]. By automating the tedious aspects of documentation, AI can free up security experts to concentrate on higher-level analysis and decision-making. Organizational studies on the implementation of generative AI in cybersecurity highlight growing industry interest in its use for security operations and vulnerability management [6], [8].

Despite the benefits, reliability concerns related to AI-generated reports are significant. Reports have shown that these systems can sometimes generate factually inaccurate descriptions, inconsistent severity ratings, or unsubstantiated recommendations [4], [10]. The existence of false information, also known as hallucination, and the lack of clear explanations make verification difficult. Hence, most research advocates for human oversight in validating AI-generated reports before incorporating them into security decisions [7], [13].

C. Reliability Challenges in AI-Assisted Vulnerability Assessment

The effectiveness of AI in identifying and reporting vulnerabilities hinges on the reliability of its outputs. Despite considerable progress in using machine learning, deep learning, and LLMs for automated security analysis, certain reliability issues impede their adoption in sensitive security contexts. Among these issues are false positives, false negatives, hallucinations, inconsistent severity assessments, and the lack of explainability, all of which are widely cited as significant challenges for the trustworthy deployment of AI-assisted cybersecurity systems [1], [4], [13].

False positives are a particularly common issue, with AI systems frequently misidentifying non-existent vulnerabilities. A large volume of false positives places additional demands on security analysts to verify each one, consuming their valuable time and resources. Research indicates that the performance of AI-based vulnerability detection systems can be inconsistent due to variations in training data, software domains, and evaluation methodologies used in different studies [1], [2].

False negatives are the inverse problem, where genuine vulnerabilities are missed by AI systems. This can lead to serious security risks for organizations and erode trust in automated security tools. Studies have suggested that AI models may struggle to detect more complex vulnerabilities, particularly if the training data lacks diversity or does not adequately represent real-world attack scenarios [6], [11].

Hallucination, particularly prevalent in generative AI and LLM-based security systems, is an emergent and significant reliability concern. This refers to the generation of plausible but factually inaccurate information by AI models, which in the context of vulnerability assessment can translate to non-existent flaws, erroneous descriptions, or misleading remediation advice. Researchers have identified hallucination as a major barrier to the reliability of AI-generated security outputs [4], [10].

Beyond accuracy issues, the reliability of AI-assisted vulnerability reporting is compromised by inconsistent severity classifications. Identical vulnerabilities may be assigned different severity levels by different AI models, stemming from differences in training data, model architecture, and contextual interpretation. This lack of uniformity can affect prioritization strategies and resource allocation decisions [2], [13].

The black-box nature of many advanced AI models, which limits explainability and transparency, poses further reliability concerns. Security analysts need to understand how a vulnerability assessment was reached to justify security decisions and validate AI findings. The research consensus emphasizes the necessity of enhanced transparency mechanisms to build trust and enable responsible use of AI-assisted vulnerability assessment tools [5], [13].

In conclusion, the reviewed literature suggests that despite the potential benefits of AI in vulnerability discovery and reporting, several reliability challenges remain significant obstacles to its widespread implementation. Overcoming issues related to false positives, false negatives, hallucination, inconsistent severity classification, and lack of explainability will

be essential for developing trustworthy AI-assisted cybersecurity solutions in the future [1], [4], [13].

D. Security risks of AI-assisted vulnerability discovery and reporting.

While there are great advantages of using AI for cybersecurity tasks, a new range of security risks might appear, making it potentially compromising the entire integrity and reliability of vulnerability assessment. According to recent findings, researchers started to recognize that the AI system itself might become an attack surface and thus may present security risks for any organization adopting AI-enabled cybersecurity tools [3], [4], [8].

One of the most popular attacks discussed in this context are prompt injection attacks where in Large Language Models based AI, crafted prompts might inject a different, non-intended logic which forces the AI to deviate from its intended objectives and thus to generate misleading or malicious outputs. In the context of vulnerability reporting, the prompt injection could potentially influence on security recommendations, change vulnerability descriptions and result in inaccurate reports. With an increase use of LLMs for various cybersecurity purposes, the resilience of these models to prompt manipulation becomes an even greater concern [4], [8].

The risks of data leakage and privacy exposure are also widely discussed issues of AI-assisted vulnerability discovery, since the AI usually requires access to confidential source code, security logs, configuration data and infrastructure information. Failure to protect sensitive data can lead to unintentional revealing of confidential assets through the AI system or through third party service providers. Privacy-preserving AI architectures and strong data governance mechanism should be carefully taken into consideration while designing AI-based cybersecurity system [3], [9].

Adversarial attacks present one of the major threats for AI-assisted security systems. They can be defined as malicious inputs that are crafted to trick machine learning models and thus may compromise their functioning by lowering their detection capabilities or misinterpreting the findings. Research has demonstrated that a lot of AI models remain vulnerable to adversarial perturbations, indicating that robust defense strategies and constant monitoring are essential [4], [12].

Model poisoning attacks which influence the training data and thus make a trained AI model less reliable are another type of threat against AI systems. Poisoned training data might introduce specific flaws into an AI model and reduce its ability to detect anomalies or may cause it to function as intended by an attacker. Since a lot of AI-assisted vulnerability detection systems are based on training on massive data, data integrity becomes crucial for any trustworthy application [8], [13].

Finally, the experts warn against an over-reliance on the outputs generated by an AI system, since relying solely on AI findings without verification could increase risks during operation. In most studies, human monitoring and validation is recommended for a trustworthy AI-driven security solution [7], [13].

III. METHODOLOGY

The approach chosen for this research study on the trustworthiness of AI-assisted vulnerability discovery and reporting system is the Structured Literature Review. It was employed to systematically, transparently, and reproducibly investigate the existing body of research to identify the challenges of reliability, security risks, research gaps and future scope within AI-enabled cybersecurity.

A. Literature sources

Various papers are found and collected from prominent scientific databases and digital libraries, which are IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect and Google Scholar. These sources are considered due to its relevance to quality peer-reviewed publications within cybersecurity, software engineering and artificial intelligence fields.

B. Search strategy

The literature survey for relevant papers was executed by combining keywords relating to artificial intelligence, cybersecurity, vulnerability assessment

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718467

and automatic reporting. Sample search queries conducted are:

1. AI-assisted vulnerability discovery
2. Software vulnerability detection using machine learning
3. Large language models for vulnerability assessment
4. AI-generated security reports
5. Trustworthy AI in cybersecurity
6. Automated vulnerability reporting

The investigation for papers published in the year ranging from 2023 to 2026 because of new advancement in generative AI and Large Language Model in cybersecurity.

Fig 1: Research process flowchart

C. Inclusion criteria

The research paper included if they satisfied the conditions given below:

1. Published in a peer-reviewed journal or conference paper.
2. Related to the topic of AI-assisted vulnerability discovery, vulnerability detection or vulnerability reporting.
3. Focused on or concerned with reliability, trustworthiness, explainability or security risks.
4. Written in English.
5. Contained appropriate technical or empirical evidence.

D. Exclusion criteria

The papers will be excluded if they:

1. Are duplicate papers
2. Are general AI-related topics and not concerned with the cybersecurity field.
3. Lacks any technical contribution or evaluation results.
4. Is an editorial, opinion piece, blog or any non-peer-reviewed article
5. Are not related to vulnerability reporting activities.

E. Data extraction and analysis

For each selected paper, the following data were extracted and analyzed: research objectives, AI technique applied, methodology of evaluation, challenges in reliability, security risks and future direction of the research. Based on this data, four major themes were created:

1. AI-assisted vulnerability discovery
2. AI-assisted vulnerability reporting
3. Challenges of reliability and trustworthiness
4. Security risks and limitations

The studies were then analyzed in comparison to find common research gaps, emerging research directions and persistent limitations.

IV. RESULTS AND ANALYSIS

A. Comparative Analysis of Existing Studies

The reviewed literature demonstrates that Artificial Intelligence has significantly improved the efficiency of vulnerability discovery and reporting processes. Various approaches, including Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs), have been applied to identify software vulnerabilities, automate security assessments, and generate vulnerability reports. However, the performance and reliability of these systems vary considerably depending on the underlying model architecture, training data quality, and evaluation methodology [1], [2].

Table I. Comparative Summary of Reviewed Studies

Ref	Focus Area	Strengths	Limitations
[1]	AI-Based Vulnerability Detection	Improved detection efficiency	Dataset quality issues
[2]	Vulnerability Detection & Patch Management	Faster remediation support	Limited real-world validation
[3]	Generative AI in Cybersecurity	Enhanced automation	Reliability concerns
[4]	LLM Applications in Security	Context-aware analysis	Hallucinations
[5]	Generative AI Review	Broad coverage of AI techniques	Evaluation inconsistency
[10]	LLM-Based Vulnerability Detection	Strong code understanding	False positives
[11]	Transformer-Based Detection	Improved classification capability	Training complexity
[13]	AI-Based Cybersecurity	Comprehensive security analysis	Explainability limitations

The comparative analysis indicates that AI systems consistently improve automation and scalability within cybersecurity workflows. Nevertheless, concerns regarding reliability, transparency, and robustness continue to affect their practical deployment in security-critical environments [1], [13].

B. Reliability Analysis

One of the most crucial issues that can emerge in AI-driven vulnerability discovery and reporting is its reliability. According to various research papers, many AI systems could report vulnerabilities that do not actually exist (false positives), thereby increasing the work required by security experts to verify them [1], [10]. If a real vulnerability goes undetected (false negatives), the system's detection mechanism might put an organization at risk of being exploited by attackers [2], [11].

New reliability issues were brought by LLMs in form of "hallucinations" where they may produce wrong but plausible descriptions of a vulnerability, or provide incorrect fix suggestions, leading to low user trust and negative impact on the decision making of security personnel [4], [10]. The lack of commonality between the reported security severity of a same vulnerability between different AI tools shows the need for uniform evaluation mechanism for AI vulnerability reports [13].

Studies reviewed show that human intervention and verification are still critical in any AI-assisted vulnerability reporting systems to overcome reliability challenges and its shortcomings [7]. Thus, trustworthiness of an AI assisted vulnerability assessment tool is also dependent on its transparency, consistency and explainability of results along with its accuracy.

C. Security Risk Analysis

Introduction of AI in the cybersecurity domain also presents new attack surfaces and new operational risks. One prominent threat that has been highlighted in several studies is vulnerability to prompt injection attacks or adversarial manipulations where inputs are crafted to mislead the AI model and tamper its outputs in such a way that it does not work correctly [4], [8].

Data Leakage is another critical threat in this domain as many of the vulnerability discovery tools require extensive access to sensitive source code, configurations or other documents. Insecure architecture could lead to unauthorized data leakage and potential privacy violations [3], [9]. Attackers could also attack the models to modify training data, a form of "model poisoning" that may result in degraded performance and inability to detect malicious code [8], [13].

Research indicates the need for robust security controls, secure model deployment, continuous monitoring and also combine human expertise with automated detection in such AI systems to effectively mitigate the security risks [7], [13].

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718467

D. Identified Research Gaps

Based on the reviewed studies, there are several important research gaps which require further investigation:

1. Lack of standardized approaches or framework to measure trustworthiness of AI generated vulnerability reports.
2. Limited real-world study of AI driven vulnerability discovery tools in large scale deployment.
3. Insufficient explainability mechanisms to facilitate decision-making on security.
4. Lack of uniform benchmarks to measure the reliability of AI systems and report quality.
5. Limited research conducted on AI assisted vulnerability reporting aspect in comparison to vulnerability detection.

The existence of these gaps proves that more research needs to be carried out to develop trustworthy, explainable and secure AI assisted vulnerability reporting systems.

Table II. Research Gaps and Future Opportunities

Research Gap	Future Direction
Lack of trustworthiness evaluation frameworks	Develop standardized assessment models
Limited explainability mechanisms	Integrate Explainable AI techniques
Insufficient real-world validation	Conduct industry-scale deployment studies
Lack of reporting quality benchmarks	Establish common evaluation metrics
Limited focus on vulnerability reporting	Expand research beyond detection tasks

Figure 2: AI Vulnerability Assessment Trust Framework

V. FUTURE RESEARCH DIRECTIONS

The conclusions of this structured literature review are that AI-assisted vulnerability discovery and reporting technologies offer immense capabilities for enhancing security operations. There are however a number of factors which are posing challenges to the integration of these systems into high-stakes security environments; these include trustworthiness, reliability, security, and transparency. Many areas of future research can be identified from the research gaps illustrated in the above section.

A. XAI for Security Assessment

One of the major challenges encountered throughout the literature is a lack of transparency in the way AI models make decisions on security assessments. Research needs to focus on developing techniques for Explainable AI (XAI) which are able to transparently demonstrate why specific vulnerabilities have been flagged, what severity they hold, and which remediation actions would best address them. Greater transparency would provide greater trust in AI models and allow for easier validation. [5] [13]

B. Standardized Frameworks for Reliable Security Assessment

There is no consistency in the evaluation methodologies used to benchmark different AI-assisted vulnerability assessment techniques. The different use of datasets, varying evaluation methods, and dissimilar testing procedures prevent objective comparisons being drawn between security systems. Research needs to develop standardized evaluation frameworks that test for the accuracy, consistency, robustness, and trustworthiness of AI security tools. [1]

C. Human-in-the-loop Security Validation

The literature review shows consistently that human input is still a crucial part of any security system in place. Therefore future security operations would greatly benefit from integrating a "human-in-the-loop" model which utilizes the strengths of both humans and automated security systems to validate AI output and reduce false positives and hallucinations. [7] [13]

D. Security in deployment of Generative AI

Large Language models used for security are introducing a range of new security risks such as prompt injection, data exfiltration and adversarial attacks. Future work must establish frameworks for secure deployment, appropriate input validation, and privacy-preserving models in order to utilize these advanced models safely within organizational systems. [4] [8] [9]

E. Trustworthiness Centered AI frameworks

The research to date has mostly centered on detecting vulnerabilities rather than on evaluating whether such detection is trustworthy. Future work should examine and develop frameworks which provide assessments for all of the factors for trustworthiness in security; reliability, transparency, security, accountability, human oversight, etc. [13]

VI. CONCLUSION

As of the present, artificial intelligence (AI) can serve as an important tool to assist vulnerability detection and reporting in current cybersecurity operations. Recent developments in machine learning, deep learning, and large language models have made it possible for automatically detect software vulnerabilities, generate security reports, and prioritize their remediation. The abovementioned capabilities are beneficial in terms of efficiency, scalability, and operational support for cybersecurity experts.

Nonetheless, this SLRs findings show that significant challenges still exist in relation to trustworthiness of AI-driven vulnerability assessment system. There still exist false positives, false negatives, hallucinated reports, unreliable severity judgments issues impacting accuracy of AI security outputs. Security issues like prompt injection attack, model manipulation, model poisoning, data exfiltration were also mentioned with significant concerns regarding safe deployment of AI within the cybersecurity field. Important research gaps found include no standardized trustworthiness framework, no real-world validations studies, no human-understandable explanations and no standardized benchmarks to test AI generated vulnerability reports. To build trust, we need to solve the aforementioned challenges.

More works on explainable AI technologies, trustworthiness benchmarks, robust deployment architectures, and human-in-the-loop approaches should be conducted in the future. Combined with reliability, transparency, security, and human intervention, the future of AI-driven vulnerability detection and reporting systems will certainly earn more trusts from cybersecurity professionals.

REFERENCES

[1] S. Shimmi, H. Okhravi, and M. Rahimi, “AI-Based Software Vulnerability Detection: A Systematic Literature Review,” *arXiv preprint arXiv:2506.10280*, 2025.

[2] A. Malkawi et al., “AI-Powered Vulnerability Detection and Patch Management in Cybersecurity: A Systematic Review of Techniques, Challenges and Emerging Trends,” *AI Journal*, vol. 8, no. 1, 2026.

[3] M. Mirtaheri et al., “Cybersecurity in the Age of Generative AI: A Systematic Review,” *Future Generation Computer Systems*, 2025.

[4] M. A. Ferrag, “Generative AI in Cybersecurity: A Comprehensive Review of Large Language Models Applications and Vulnerabilities,” *Array*, vol. 25, 2025.

[5] A. Tumpa et al., “Generative AI in Cybersecurity: A Systematic Literature Review and Meta-Analysis,” *Preprints*, 2026.

[6] D. Nott, “Organizational Adaptation to Generative AI in Cybersecurity: A Systematic Review,” *arXiv preprint*, 2025.

[7] B. Yiğit and M. Alkan, “Review of Generative AI Methods in Cybersecurity,” *Array*, vol. 26, 2026.

[8] M. Ibrar et al., “Generative AI: A Double-Edged Sword in the Cyber Threat Landscape,” *Artificial Intelligence Review*, 2025.

[9] M. Uddin et al., “Generative AI Revolution in Cybersecurity: Opportunities and Challenges,” *Artificial Intelligence Review*, 2025.

[10] P. Kaniewski et al., “A Systematic Literature Review on Detecting Software Vulnerabilities with Large Language Models,” *arXiv preprint arXiv:2507.22659*, 2025.

[11] M. Naseer et al., "A Systematic Literature Review for Transformer-Based Software Vulnerability Detection," *arXiv preprint arXiv:2604.24822*, 2026.

[12] Vincent Tamaramiebi Daniel, Biralatei Fawei, "A Systematic Review of AI-Driven Automated Software Vulnerability Detection Systems," *International Journal of Computer Science and Mathematical Theory*, 2025.

[13] M. Lezzi et al., "A Systematic Literature Review on AI-Based Cybersecurity," *Cybersecurity*, vol. 5, no. 4, 2025.