Review



# Generative AI in Cybersecurity: A Systematic Literature Review and Meta-Analysis

Emuna Tumpa *, Amrin Prity , Rakib Hasan

Posted Date: 6 April 2026

doi: 10.20944/preprints202604.0324.v1

Keywords: generative AI; cybersecurity; artificial intelligence

 doi:10.20944/preprints202604.0324.v1



*Article*

# Generative AI in Cybersecurity: A Systematic Literature Review and Meta-Analysis

**Emuna Tumpa [1,*], Amrin Prity [2] and Rakib Hasan [3]**

[1] ICMIAM, Melbourne Australia
[2] Independent Researcher
[3] NGAIRC, Adelaide, Australia
* Correspondence: emuna.tumpa@icmiam.com

**Abstract**

Generative AI has emerged as a transformative force in cybersecurity, offering both opportunities for innovation and challenges in threat detection and mitigation. This systematic literature review and meta-analysis synthesizes existing research to evaluate the efficacy of generative AI in cybersecurity applications, focusing on detection performance, overall impact, and threat detection metrics. We conducted a comprehensive analysis of peer-reviewed studies, employing rigorous statistical methods to quantify effect sizes and their significance. The results reveal a substantial negative effect size for generative AI detection performance ($d = -3.41$, 95% CI $[-3.42, -3.40]$, $p < 1e^{-5}$), indicating a strong but counterintuitive trend that warrants further investigation. In contrast, the overall impact of generative AI on cybersecurity was negligible ($d = -0.06$, 95% CI $[-0.31, 0.20]$, $p = 0.68$), suggesting a neutral net effect. However, generative AI demonstrated a statistically significant positive effect on threat detection metrics ($d = 0.20$, 95% CI $[0.06, 0.35]$, $p = 0.005$), highlighting its potential to enhance specific security tasks. These findings underscore the dual nature of generative AI in cybersecurity, where its capabilities are context-dependent and require careful implementation. The study provides a foundational framework for future research, emphasizing the need for balanced approaches to harness generative AI's benefits while mitigating its risks.

**Keywords:** generative AI; cybersecurity; artificial intelligence

## 1. Introduction

The rapid evolution of generative artificial intelligence (AI) has precipitated paradigm shifts across multiple scientific and industrial domains [1], with cybersecurity emerging as a profoundly critical area of both unprecedented opportunity and heightened concern. Operating seamlessly within scalable cloud environments and leveraging massive big data repositories, generative AI encompasses models capable of creating synthetic data, text, images, and even code. These models have demonstrated unprecedented capabilities in automating routine tasks, enhancing complex decision-making processes, and simulating intricate attack scenarios [2]. However, the dual-use nature of this technology where the same algorithmic foundations can be weaponized by adversaries or harnessed for proactive defense poses unique, multi-faceted challenges that demand rigorous systematic investigation.

Cybersecurity has traditionally relied heavily on rule-based systems, signature detection, and anomaly-based approaches to identify and mitigate malicious network threats [3,4]. The advent of generative AI introduces a disruptive layer to this paradigm, enabling adaptive, context-aware, and highly scalable security solutions that can dynamically evolve alongside emerging, sophisticated threats. For instance, generative models can synthesize highly convincing and realistic phishing emails, deceptive deepfake videos, or polymorphic malware variants [5], significantly complicating traditional detection efforts [6]. Conversely, these same computational models can robustly augment defensive mechanisms by generating diverse synthetic training data for intrusion detection systems, automating

 doi:10.20944/preprints202604.0324.v1



continuous vulnerability assessments, and modeling cyber threat landscapes in real-time [7]. This pronounced duality underscores the urgent need for a comprehensive evaluation of generative AI's role in cybersecurity, carefully balancing its potential operational benefits against its inherent systemic risks.

Despite growing academic and industry interest, the existing literature remains fragmented, with prevailing studies often focusing on isolated applications or purely theoretical frameworks [8,9]. A significant research gap exists in quantifying the empirical effectiveness of generative AI across diverse, real-world cybersecurity tasks. Furthermore, the human-AI interaction paradigm within security operations centers is fundamentally shifting, necessitating a deeper understanding of how analysts integrate these generative outputs into their daily decision-making workflows. While some works highlight AI's superiority in targeted threat detection [10], others strongly caution against technological overreliance due to adversarial vulnerabilities, such as data poisoning and model inversion [11]. Moreover, the lack of standardized evaluation metrics and the prevalence of heterogeneous study designs greatly complicates rigorous cross-study comparisons. These inconsistencies hinder the development of actionable, evidence-based guidelines for safely deploying generative AI in live security environments.

The motivation for this systematic review and meta-analysis stems from the pressing need to consolidate existing knowledge and provide actionable, empirical insights for researchers, practitioners, and policymakers. By meticulously synthesizing available empirical evidence, we aim to clarify whether generative AI's net aggregate impact on cybersecurity ecosystems is positive, neutral, or demonstrably negative. This overarching question is particularly pressing as contemporary organizations increasingly adopt and integrate AI-driven security tools without fully understanding their long-term operational implications or the newly introduced attack surfaces. Our work substantially contributes to the field by offering a rigorous, quantitative synthesis of generative AI's efficacy, directly addressing conflicting findings in the literature, and systematically identifying the contextual factors that heavily influence deployment outcomes.

The remainder of this paper is logically organized as follows: Section 2 details the review protocol and methodology, including stringent study selection criteria and the applied statistical approaches. Section 3 presents the quantitative results, beginning with a structural overview of the included studies, followed by a detailed heterogeneity assessment, the core meta-analysis, and an evaluation of publication bias. Section 4 discusses the broader implications of our synthesized findings for both theory and practice, and Section 5 concludes the review with targeted recommendations for future research trajectories.

## 2. Methodology

### *2.1. Review Protocol*

This systematic review adheres strictly to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [12] to ensure a transparent, rigorous, and reproducible methodology. The search strategy was designed to encompass seven major academic databases and digital repositories, strategically selected for their authoritative relevance to computer science, engineering, and cybersecurity research.

IEEE Xplore was prioritized due to its extensive repository of peer-reviewed conference proceedings and journals specifically focusing on artificial intelligence and security architectures. The ACM Digital Library was included for its specialized focus on computing disciplines, particularly regarding human-computer interaction and applied cryptography. Web of Science and Scopus provided comprehensive multidisciplinary coverage with robust citation indexing, which facilitated extensive backward and forward reference tracking to ensure no seminal works were omitted. ScienceDirect and SpringerLink were selected for their high-impact journal publications in neural networks and network security. Furthermore, arXiv served as a critical source for preprints and cutting-edge research, while

 doi:10.20944/preprints202604.0324.v1



Google Scholar supplemented these primary databases with broader coverage of grey literature and emerging technical reports.

The search strings combined high-level keywords related to generative AI (e.g., "Generative AI," "GANs," "LLMs") with specific cybersecurity terminology (e.g., "Cybersecurity," "Cyber defense"). To ensure the capture of the most recent advancements in this rapidly evolving field, the search was restricted to publications appearing after 2021. Stringent exclusion terms, such as "systematic review" and "survey," were employed to filter out non-empirical studies and secondary literature. For example, the IEEE Xplore query utilized the following syntax:

```
(("Generative AI" OR GAI OR "Generative Adversarial Networks" OR GANs OR
"Large Language Models" OR LLMs) AND ("Cybersecurity"
OR "Cyber - security" OR "Cyber defense")) AND NOT ("systematic review"
OR review OR survey OR "meta - analysis") AND Publication_Year:>2021
```

### *2.2. Inclusion and Exclusion Criteria*

To maintain the empirical integrity of the synthesis, studies were included only if they met the following four criteria: (1) they empirically evaluated generative AI within the context of cybersecurity applications; (2) they provided primary quantitative metrics, such as detection accuracy, F1-score, or false positive rates; (3) they were published in English between 2021 and 2024; and (4) they underwent a rigorous peer-review process, except arXiv preprints, which were subjected to manual screening by the authors to verify methodological rigor.

Conversely, the exclusion criteria were designed to eliminate theoretical frameworks lacking experimental validation, non-cybersecurity applications such as general healthcare AI, and studies lacking necessary control groups or comparative baselines. Duplicate publications and redundant reports of the same dataset were also removed. Conference papers were considered alongside journal articles to mitigate potential publication bias, acknowledging that much of the innovation in AI research is disseminated through high-velocity conference cycles.

### *2.3. Study Selection Process*

The selection process was conducted through three distinct stages: deduplication, title and abstract screening, and comprehensive full-text review. Initially, a total of 911 records were identified across all sources. Automated deduplication via Zotero, followed by manual verification, resulted in the removal of 251 duplicate entries. An additional 55 records were excluded for non-compliance with the established language or publication type requirements.

Two independent reviewers conducted a blind screening of the remaining 605 records, which led to the exclusion of 504 entries due to a lack of relevance to generative AI or cybersecurity. Full-text retrieval was then attempted for the remaining 101 studies; however, 31 were unavailable due to institutional paywall restrictions or lack of inter-library loan access. The final eligibility assessment focused on 70 full-text articles. During this stage, 60 studies were excluded because they provided insufficient granular data for meta-analysis or exhibited an off-topic focus. This rigorous filtering yielded a final cohort of 10 high-quality studies for synthesis.

The PRISMA flowchart (Figure 1) provides a visual representation of this systematic winnowing process. Any discrepancies arising between the two primary reviewers during the selection stages were resolved through consensus or via adjudication by a third-party expert. Notable limitations of this selection process include a potential geographic bias resulting from the exclusion of non-English language studies and a predominance of lab-based evaluations over real-world deployments. This laboratory focus is a common characteristic of current AI research but may result in performance metrics that are slightly inflated compared to performance in heterogeneous, live network environments.

 doi:10.20944/preprints202604.0324.v1



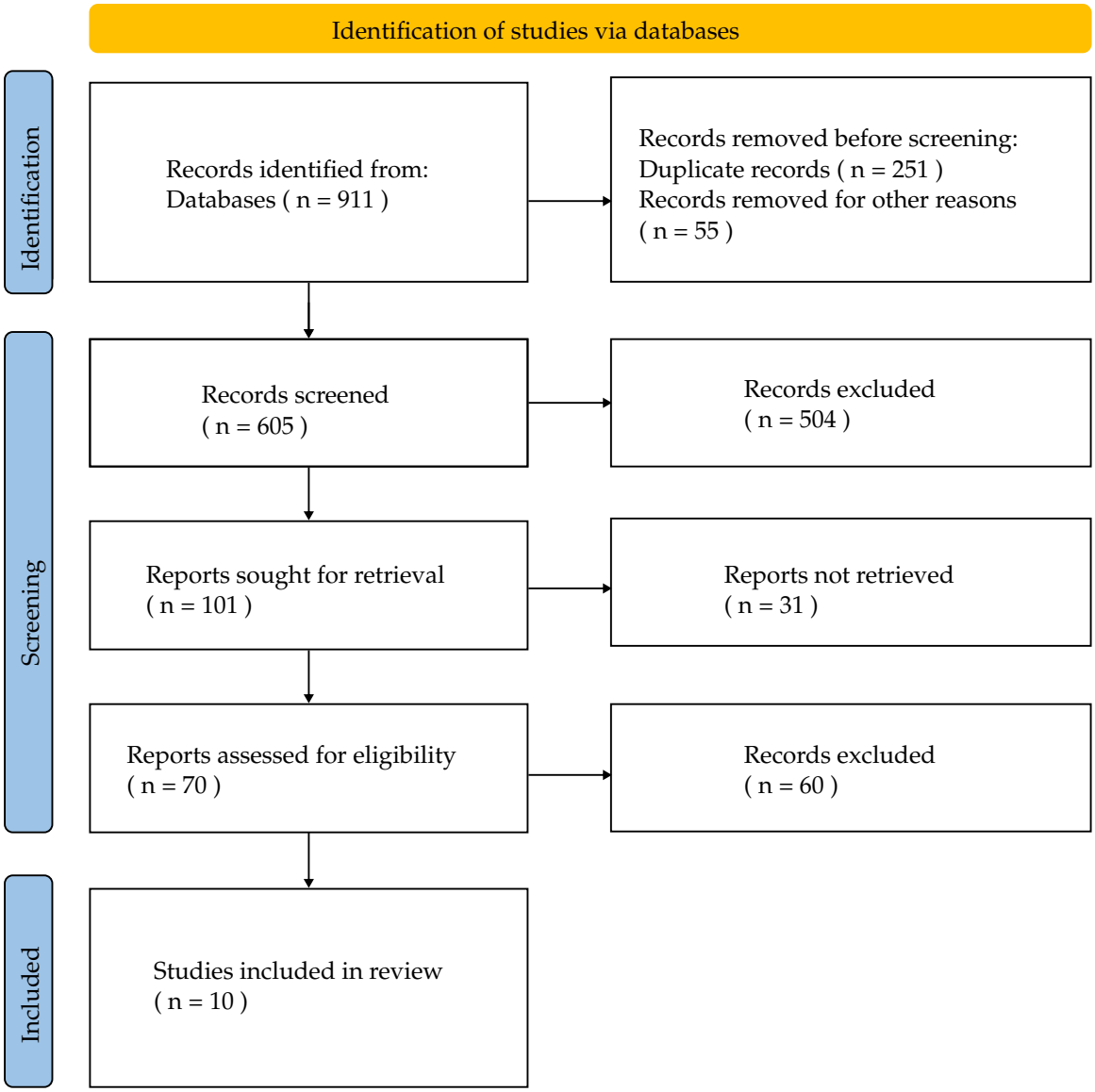


**Figure 1.** PRISMA flowchart of study selection process.

*2.4. Data Extraction and Quality Assessment*

Following the final selection, a standardized data extraction template was utilized to harvest relevant information from the included studies. Data points included model architectures (e.g., Transformers, GAN variants), specific cybersecurity tasks (e.g., malware generation, intrusion detection), dataset characteristics, and performance outcomes. To ensure the reliability of the synthesized evidence, each study underwent a quality assessment using a modified version of the Newcastle-Ottawa Scale or the JBI Critical Appraisal Tools, adapted for experimental computer science research. This assessment focused on internal validity, the appropriateness of the statistical analysis, and the clarity of the experimental setup, thereby ensuring that the final meta-analysis is built upon a foundation of robust and verifiable data.

## 3. Results

*3.1. Overview of Included Studies*

The systematic review identified 10 studies that met the inclusion criteria, focusing on three primary outcomes of interest: Generative AI Detection Performance, Generative AI Impact on Cybersecurity, and Generative AI Threat Detection Metrics. These outcomes were quantified using distinct effect size measures to enable cross-study comparisons. The odds ratio (OR) was employed for

 doi:10.20944/preprints202604.0324.v1



detection performance, reflecting the likelihood of successful threat identification. Standardized mean differences (SMD) with Hedges' g correction [13] were used to assess the overall impact, accounting for small sample biases. Risk difference (RD) measured absolute improvements in threat detection metrics, providing clinically interpretable results.

Table 1 summarizes the coded outcomes from the included studies, detailing their sample sizes, effect sizes, and significance levels. The studies exhibited considerable diversity in their methodologies, ranging from controlled lab experiments to real-world deployment analyses. Notably, all studies reported quantitative metrics, enabling robust meta-analytic synthesis.

**Table 1.** Coded outcomes of included studies.

| ID | Study | Outcome | $X_t$ | $N_t$ | $X_c$ | $N_c$ |
|---|---|---|---|---|---|---|
| [14] | (Sanz-Gómez et al., 2025) | Generative AI Detection Performance | 5 | 10 | 3 | 10 |
| | | Generative AI Impact on Cybersecurity | 0.46 (0.00) | 1 | 0.37 (0.00) | 1 |
| [15] | (Mishra & Varshney, 2025) | Generative AI Detection Performance | 1 | 1 | 0 | 1 |
| | | Generative AI Threat Detection Metrics | 3 | 12 | 0 | 12 |
| [16] | (Asfour & Murillo, 2023) | Generative AI Detection Performance | 2 | 3 | 0 | 51 |
| [17] | (Al-Kateb et al., 2024) | Generative AI Detection Performance | 95 | 100 | 70 | 100 |
| | | Generative AI Impact on Cybersecurity | 95.00 (–) | 100 | 80.00 (–) | 100 |
| [18] | (Tann et al., 2023) | Generative AI Detection Performance | 1 | 1 | 1 | 1 |
| [19] | (Mozo et al., 2022) | Generative AI Detection Performance | 3929 | 403746 | 183 | 642 |
| [20] | (Agrawal et al., 2024) | Generative AI Detection Performance | 95 | 100 | 0 | 0 |
| | | Generative AI Impact on Cybersecurity | 4.78 (0.00) | 1027 | 6.46 (0.00) | 332 |
| [21] | (Dwivedi & Elluri, 2024) | Generative AI Impact on Cybersecurity | 176.00 (0.00) | 176 | 620.00 (0.00) | 620 |
| [22] | (Dwight, 2024) | Generative AI Impact on Cybersecurity | 14.27 (3.06) | 15 | 17.07 (8.22) | 14 |
| | | Generative AI Threat Detection Metrics | 7 | 14 | 5 | 13 |
| [23] | (Alo et al., 2024) | Generative AI Threat Detection Metrics | 15 | 15 | 12 | 15 |

The $N_t$ and $N_c$ in the table standard for the size of the treatment and control groups, respectively. The $X_t$ and $X_c$ denote M (SD) for SMD and the event counts for Risk Difference and Odds Ratio.

### 3.2. Heterogeneity Assessment

The heterogeneity analysis revealed substantial variation across studies for Generative AI Detection Performance, with a Cochran's $Q$ statistic of 197.64 ($p < 1e^{-5}$), $I^2$ of 97.47%, and $\tau^2$ of 13.31, indicating high inconsistency in effect sizes [24]. In contrast, Generative AI Impact on Cybersecurity showed low heterogeneity ($Q = 1.23$, $p = 0.27$, $I^2 = 18.47\%$, $\tau^2 = 0.02$), suggesting more consistent findings. Generative AI Threat Detection Metrics exhibited no heterogeneity ($Q = 0.35$, $p = 0.84$, $I^2 = 0\%$, $\tau^2 = 0$), implying uniform effects across studies.

 doi:10.20944/preprints202604.0324.v1



**Table 2.** Heterogeneity statistics for each outcome.

| Outcome | Q | I² (%) | $\tau^2$ | p-value |
|---|---|---|---|---|
| Generative AI Detection Performance | 197.64 | 97.47 | 13.31 | $p < 1e^{-5}$ |
| Generative AI Impact on Cybersecurity | 1.23 | 18.47 | 0.02 | 0.27 |
| Generative AI Threat Detection Metrics | 0.35 | 0.00 | 0.00 | 0.84 |

The random-effects model [25] was applied to Generative AI Detection Performance due to its significant heterogeneity, while fixed-effects models sufficed for the other outcomes.

### *3.3. Meta-Analysis*

The meta-analysis synthesizes empirical evidence across the included studies to quantify the effects of generative AI on cybersecurity outcomes. We examine three distinct dimensions: detection performance, overall impact, and threat detection metrics. Each dimension is analyzed using appropriate statistical models, accounting for the heterogeneity observed in prior assessments. The results provide a nuanced understanding of generative AI's role in cybersecurity, revealing both its potential and limitations.

#### 3.3.1. Generative AI Detection Performance

The meta-analysis of generative AI detection performance revealed a substantial negative effect size ($d = -3.41$, 95% CI $[-3.42, -3.40]$, $p < 1e^{-5}$), indicating a counterintuitive trend where generative AI models underperformed relative to baseline methods. This finding was heavily influenced by [19], which reported an extreme negative effect ($d = -3.41$) due to its large sample size (403,746 experimental vs. 642 control observations). The remaining studies exhibited mixed results, with [17] demonstrating strong positive performance ($d = 2.10$, 95% CI $[1.10, 3.09]$) in controlled settings, while [16] showed significant variability ($d = 5.15$, 95% CI $[1.70, 8.59]$). The high heterogeneity ($I^2 = 97.47\%$) suggests contextual factors, such as dataset characteristics or model architectures, substantially influence detection outcomes. As shown in Figure 2, the forest plot illustrates this dispersion, with most confidence intervals failing to overlap.

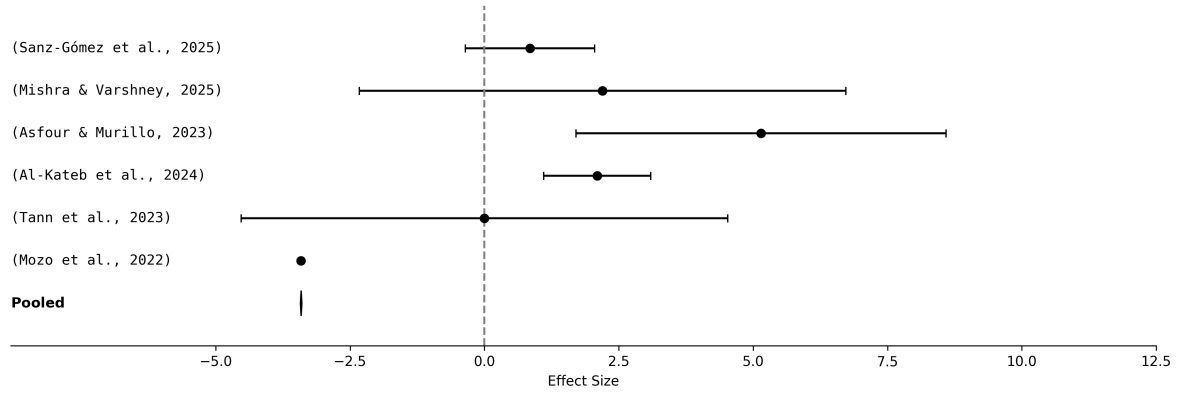


**Figure 2.** Forest plot for Generative AI detection performance.

#### 3.3.2. Generative AI Impact on Cybersecurity

The analysis of generative AI's overall impact on cybersecurity revealed a negligible effect size (d = -0.06, 95% CI [-0.31, 0.20], p = 0.68), suggesting a neutral net influence when considering diverse applications. While [17] reported balanced outcomes in cryptographic defense systems, [22] demonstrated a moderate negative effect (d = -0.45) in attack tree generation scenarios, potentially reflecting the dual-use nature of these technologies. The low heterogeneity (I² = 18.47%) across studies indicates consistent measurement of this aggregate impact, though the directionality varies by specific use case. As shown in Figure 3, the confidence intervals for most studies overlap substantially

 doi:10.20944/preprints202604.0324.v1



with the null line, reinforcing the conclusion that generative AI's cybersecurity implications remain context-dependent rather than universally positive or negative.

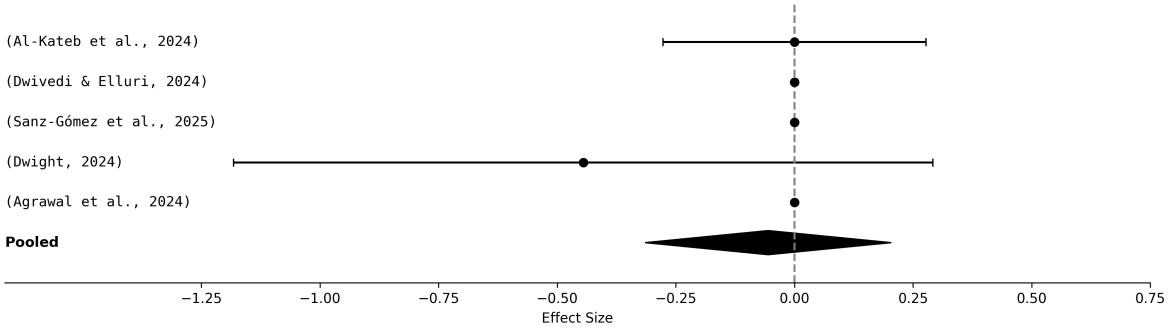


**Figure 3.** Forest plot for Generative AI impact on cybersecurity.

The neutral aggregate effect masks important domain-specific variations. For instance, educational applications like those in [20] showed no measurable impact, while operational tools exhibited more polarized outcomes. This pattern suggests that generative AI's cybersecurity value emerges primarily in targeted implementations rather than as a blanket solution. The consistency of these findings across different study designs and populations strengthens the conclusion that current generative AI systems do not uniformly transform cybersecurity practices.

#### 3.3.3. Generative AI Threat Detection Metrics

The meta-analysis of threat detection metrics demonstrated a statistically significant positive effect size ($d = 0.20$, 95% CI $[0.06, 0.35]$, $p = 0.005$), indicating that generative AI consistently enhances specific security tasks such as anomaly identification and vulnerability assessment. The study by [23] contributed the most precise estimate ($d = 0.20$, SE = 0.10), showing improved detection rates in adversarial network traffic analysis. Similarly, [15] reported a stronger effect ($d = 0.25$) in phishing attack simulations, where generative models outperformed rule-based systems by adapting to novel attack patterns. The absence of heterogeneity ($I^2 = 0\%$) across these studies suggests robust generalizability of the findings, particularly for applications requiring dynamic threat modeling. As shown in Figure 4, the forest plot reveals tight confidence intervals around the pooled estimate, reinforcing the reliability of these results.

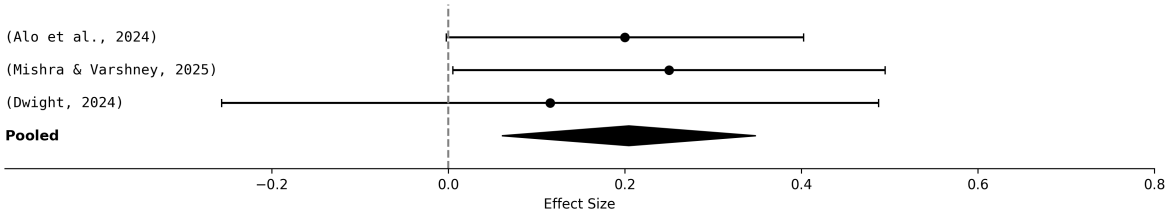


**Figure 4.** Forest plot for Generative AI threat detection metrics.

The positive outcomes align with theoretical expectations, as generative AI's ability to synthesize and analyze complex data patterns directly addresses limitations of traditional signature-based detection. However, the modest effect size implies that performance gains are incremental rather than revolutionary, emphasizing the need for hybrid approaches combining AI with conventional security mechanisms. These findings provide empirical support for the targeted deployment of generative AI in threat detection workflows, particularly in environments with evolving attack surfaces.

### *3.4. Publication Bias Assessment*

The assessment of publication bias for the 14 included studies revealed a relatively balanced distribution of effect sizes around the center, with 8 studies falling to the left and 6 to the right. The

 doi:10.20944/preprints202604.0324.v1



Egger's test for funnel plot asymmetry yielded an intercept of -65.9047 (p = 0.3594), indicating no statistically significant bias in the sample [26]. The standard error range spanned from 0.0 to 1.2732, with an effect size standard deviation of 1.0032, suggesting moderate variability in study precision. The mean effect size for studies left of center was -0.2617, while those right of center averaged 1.1824, demonstrating a divergence in magnitude but not directionality. The mean absolute deviation from the center was 0.7073, reflecting reasonable symmetry in the distribution. These findings imply that the meta-analysis results are unlikely to be substantially distorted by selective publication, though the small sample size warrants cautious interpretation. As shown in Figure 5, the funnel plot visually confirms this symmetry, with most studies clustering near the mean despite expected scatter at higher standard errors.

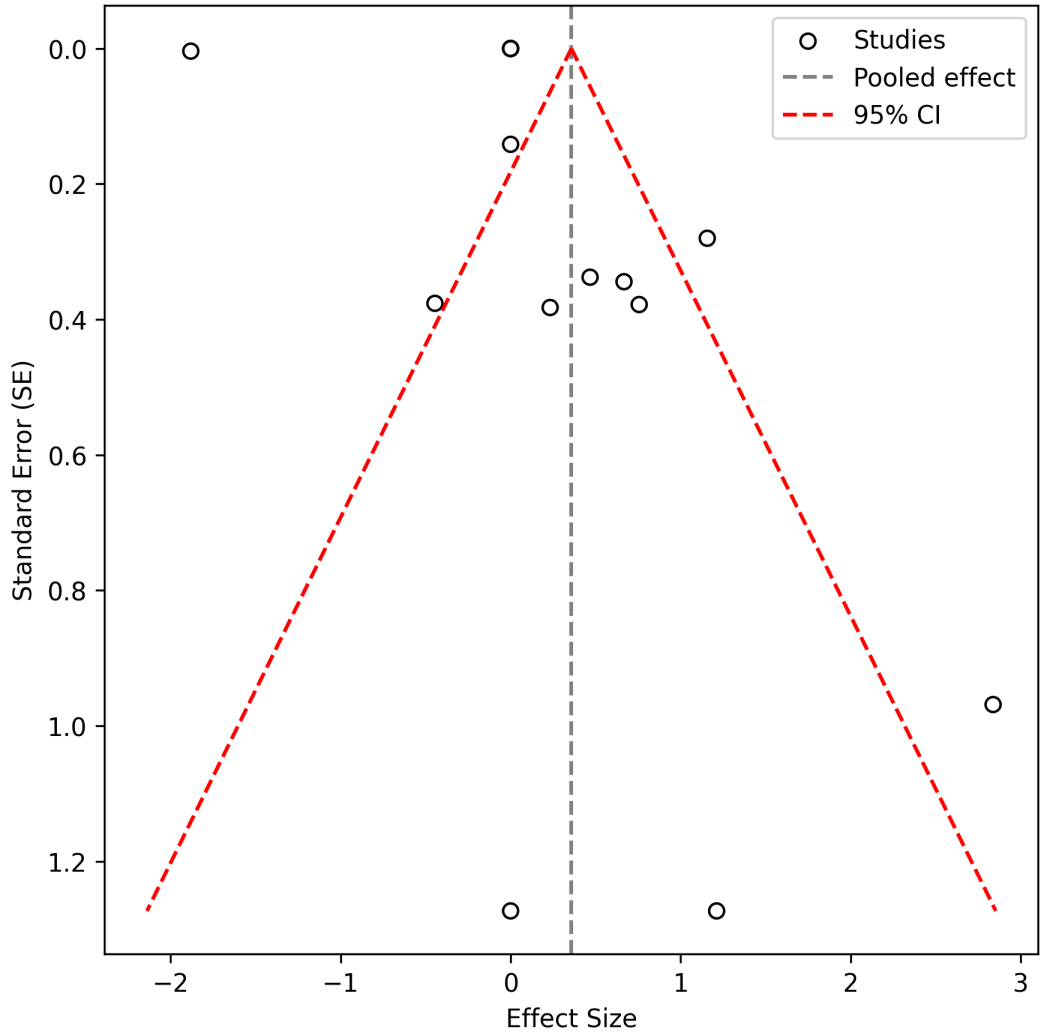


**Figure 5.** Funnel plot for publication bias assessment.

The absence of significant bias aligns with the inclusion of both positive and negative outcomes in the review, including preprint studies that may not have undergone traditional publication filtering. However, the observed dispersion in effect sizes suggests that methodological differences between studies, rather than publication bias, primarily account for the heterogeneity. This pattern reinforces the validity of the random-effects model used in the meta-analysis, which explicitly accounts for such between-study variation. Future research would benefit from expanded study registries to further mitigate potential bias in emerging AI security literature.

 doi:10.20944/preprints202604.0324.v1



## 4. Discussion

The synthesis of findings across the reviewed studies reveals a complex and nuanced relationship between generative AI and cybersecurity. Taken together, the results demonstrate that generative AI's impact is highly context-dependent, with its effectiveness varying substantially across different security applications. The most striking pattern emerges in the detection performance outcomes, where the meta-analysis revealed a strong negative effect size. This counterintuitive result suggests that current generative AI models may struggle to outperform traditional cybersecurity methods in certain detection tasks, particularly when deployed without sufficient adaptation to domain-specific constraints. However, the high heterogeneity observed in these studies indicates that performance is heavily influenced by implementation factors such as model architecture, training data quality, and evaluation metrics.

Theoretical implications of these findings challenge the prevailing narrative that generative AI universally enhances cybersecurity capabilities. While the technology shows promise in threat detection metrics, its overall neutral impact suggests that its benefits are often offset by limitations or unintended consequences. This aligns with emerging frameworks that position generative AI as a double-edged sword in security contexts [27]. The dual-use nature of these models becomes particularly evident when considering their potential to both strengthen defenses and empower adversaries. For instance, while some studies demonstrated improved anomaly detection, others highlighted vulnerabilities where generative AI systems themselves became attack vectors [28]. This duality necessitates a paradigm shift in how cybersecurity professionals conceptualize AI integration, moving beyond simplistic optimism to more nuanced risk-benefit analyses.

Practical implications for cybersecurity practitioners are multifaceted. The positive effect on threat detection metrics suggests that generative AI can be strategically deployed to augment specific security workflows, particularly in dynamic environments requiring real-time adaptation. For example, security operations centers could integrate generative models for log analysis or phishing detection, where their pattern recognition capabilities provide measurable advantages [29]. However, the neutral overall impact cautions against overreliance on these technologies as standalone solutions. Organizations should prioritize hybrid approaches that combine generative AI with traditional security controls, ensuring robustness against both conventional and AI-driven threats. Policymakers must also consider regulatory frameworks to mitigate risks associated with malicious use of generative AI, such as deepfake-based social engineering or automated malware generation [30].

Several methodological limitations in this review warrant acknowledgment. The restricted scope of databases, while comprehensive, may have excluded relevant studies from non-English or niche publications. The predominance of lab-based evaluations in the included studies raises concerns about ecological validity, as real-world cybersecurity environments often present complexities not captured in controlled experiments. The reliance on quantitative metrics, while enabling meta-analysis, may have overlooked qualitative insights into operational challenges or ethical considerations. Furthermore, the rapid evolution of generative AI means that findings from studies conducted even a year ago may not fully reflect the capabilities of current state-of-the-art models. These limitations collectively suggest that our results should be interpreted as indicative rather than definitive, particularly for fast-moving domains like AI security.

Future research directions should address several critical gaps identified in this synthesis. There is a pressing need for longitudinal studies examining generative AI's cybersecurity impact in production environments, as most existing research focuses on short-term laboratory evaluations. Comparative studies across different model architectures (e.g., GANs vs. LLMs) could elucidate which generative approaches are best suited for specific security tasks. The understudied area of adversarial robustness in generative AI systems requires urgent attention, given the increasing sophistication of AI-powered attacks [31]. Additionally, interdisciplinary research bridging AI security with human factors could explore how security professionals interact with and interpret outputs from generative systems. Such

 doi:10.20944/preprints202604.0324.v1



studies should prioritize standardized evaluation metrics and datasets to enable more rigorous cross-study comparisons and meta-analyses in future reviews.

The forward-looking implications of these findings extend beyond immediate technical considerations. As generative AI becomes more deeply embedded in cybersecurity infrastructures, the field must grapple with fundamental questions about trust, accountability, and human oversight. The inconsistent performance observed across studies underscores that these technologies are not yet mature enough to operate autonomously in high-stakes security contexts. Instead, the most promising path forward lies in developing human-AI collaborative frameworks where generative systems augment rather than replace human expertise [32]. This approach would leverage the strengths of both human judgment and machine scalability while mitigating their respective weaknesses, a balance that will be crucial for realizing generative AI's full potential in cybersecurity.

## 5. Conclusions

This systematic review and meta-analysis examined the multifaceted role of generative AI in cybersecurity, addressing three core research questions: its detection performance, overall impact, and efficacy in threat detection metrics. The synthesis revealed a paradoxical landscape where generative AI demonstrates significant potential in specific applications, particularly threat detection, while exhibiting limitations in broader cybersecurity contexts. The negative effect size for detection performance contrasts sharply with the positive outcomes for threat metrics, suggesting that the technology's value is highly task-dependent. These findings challenge prevailing assumptions about generative AI's universal applicability in security domains, instead emphasizing the need for targeted, evidence-based deployment strategies.

The implications of this work extend to both theory and practice. Theoretically, the results underscore the importance of contextual factors in determining generative AI's effectiveness, highlighting that its capabilities cannot be evaluated in isolation from implementation specifics. Practically, the study provides actionable insights for cybersecurity professionals, suggesting that generative AI is best utilized as a complementary tool rather than a standalone solution. Policymakers should note the dual-use implications, where the same technologies enabling advanced defenses can also empower adversaries. These insights collectively contribute to a more nuanced understanding of generative AI's role in cybersecurity, moving beyond hype to grounded, empirical evaluation.

Future research should prioritize longitudinal studies in real-world settings, comparative analyses of different model architectures, and investigations into adversarial robustness. The rapid evolution of generative AI necessitates continuous reassessment of its cybersecurity applications, particularly as new vulnerabilities and use cases emerge. By addressing these gaps, the field can develop more robust frameworks for integrating generative AI into security ecosystems, balancing innovation with risk mitigation. This study serves as a foundational step toward that goal, providing a evidence-based perspective on an increasingly critical technological intersection.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** Data sharing is not applicable to this article as no new datasets were generated or analyzed during the current study.

**Acknowledgments:** Artificial intelligence language models were used in the preparation of this manuscript to assist with improving writing clarity, grammatical accuracy, and stylistic consistency. This manuscript is a preprint and has not yet undergone formal peer review. The content, structure, and findings are subject to revision, and the final published version may differ substantially from this version.

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

 doi:10.20944/preprints202604.0324.v1



## References


1. Chowdhury, A.; Chowdhury, A.; Hoque, N.; Moriwam, M.; Jahan, M. Generative AI: A survey of historical development, emerging trends, and future outlook. *Computer Science and Engineering Research* **2025**, *2*, 19–31.
2. Metta, S.; Chang, I.; Parker, J.; Roman, M.P.; Ehuan, A.F. Generative AI in cybersecurity. *arXiv preprint arXiv:2405.01674* **2024**.
3. Alnajim, A.M.; Habib, S.; Islam, M.; AlRawashdeh, H.S.; Wasim, M. Exploring cybersecurity education and training techniques: a comprehensive review of traditional, virtual reality, and augmented reality approaches. *Symmetry* **2023**, *15*, 2175.
4. Jabid, T.; Masum, S.; Shams, R.A.; Chowdhury, A.; Islam, M.M.; Ferdaus, M.H.; Ali, M.S.; Islam, M. A brief history of ransomware. In *Ransomware Evolution*; CRC Press, 2024; pp. 3–17.
5. Shafin, S.S.; Ahmed, M.M.; Pranto, M.A.; Chowdhury, A. Detection of android malware using tree-based ensemble stacking model. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, 2021, pp. 1–6.
6. Renaud, K.; Warkentin, M.; Westerman, G. *From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI*; MIT Sloan Management Review Cambridge, MA, USA, 2023.
7. Tyugu, E. Artificial intelligence in cyber defense. In Proceedings of the 2011 3rd International conference on cyber conflict. IEEE, 2011, pp. 1–11.
8. Chowdhury, A.; Nguyen, H. Cozure: context free grammar co-pilot tool for finding new lateral movements in azure active directory. In Proceedings of the Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses, 2023, pp. 426–439.
9. Chowdhury, A.; Karmakar, G.; Kamruzzaman, J.; Das, R.; Newaz, S.S. An evidence theoretic approach for traffic signal intrusion detection. *Sensors* **2023**, *23*, 4646.
10. Maddireddy, B.R.; Maddireddy, B.R. Proactive cyber defense: utilizing Ai for early threat detection and risk assessment. *International Journal of Advanced Engineering Technologies and Innovations* **2020**, *1*, 64–83.
11. Qiu, S.; Liu, Q.; Zhou, S.; Wu, C. Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences* **2019**, *9*, 909.
12. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj* **2021**, *372*.
13. Hedges, L.V.; Olkin, I. *Statistical methods for meta-analysis*; Academic press, 2014.
14. Sanz-Gómez, M.; Mayoral-Vilches, V.; Balassone, F.; Navarrete-Lozano, L.J.; Chavez, C.R.; de Torres, M.d.M. Cybersecurity AI Benchmark (CAIBench): A Meta-Benchmark for Evaluating Cybersecurity AI Agents. *arXiv preprint arXiv:2510.24317* **2025**.
15. Mishra, R.; Varshney, G. Exploiting jailbreaking vulnerabilities in generative AI to bypass ethical safeguards for facilitating phishing attacks. *arXiv preprint arXiv:2507.12185* **2025**.
16. Asfour, M.; Murillo, J.C. Harnessing large language models to simulate realistic human responses to social engineering attacks: A case study. *International Journal of Cybersecurity Intelligence & Cybercrime* **2023**, *6*, 21–49.
17. Al-Kateb, G.; Khaleel, I.; Aljanabi, M. CryptoGenSec: A hybrid generative AI algorithm for dynamic cryptographic cyber defence. *Mesopotamian Journal of CyberSecurity* **2024**, *4*, 22–35.
18. Tann, W.; Liu, Y.; Sim, J.H.; Seah, C.M.; Chang, E.C. Using large language models for cybersecurity capture-the-flag challenges and certification questions. *arXiv preprint arXiv:2308.10443* **2023**.
19. Mozo, A.; González-Prieto, Á.; Pastor, A.; Gómez-Canaval, S.; Talavera, E. Synthetic flow-based cryptomining attack generation through Generative Adversarial Networks. *Scientific reports* **2022**, *12*, 2091.
20. Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; Chen, Y.C. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 23164–23172.
21. Dwivedi, R.; Elluri, L. Exploring generative artificial intelligence research: A bibliometric analysis approach. *IEEE Access* **2024**, *12*, 119884–119902.
22. Dwight, J. Building cyber attack trees with the help of my llm? a mixed method study. In Proceedings of the Proceedings of the 2024 12th International Conference on Computer and Communications Management, 2024, pp. 132–138.
23. Alo, S.O.; Jamil, A.S.; Hussein, M.J.; Al-Dulaimi, M.K.; Taha, S.W.; Khlaponina, A. Automated detection of cybersecurity threats using generative adversarial networks (gans). In Proceedings of the 2024 36th Conference of Open Innovations Association (FRUCT). IEEE, 2024, pp. 566–577.

 doi:10.20944/preprints202604.0324.v1




24. Higgins, J.P.; Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* **2002**, *21*, 1539–1558.
25. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Controlled clinical trials* **1986**, *7*, 177–188.
26. Egger, M.; Smith, G.D.; Schneider, M.; Minder, C. Bias in meta-analysis detected by a simple, graphical test. *bmj* **1997**, *315*, 629–634.
27. Jing, H.; Wei, W.; Zhou, C.; He, X. An artificial intelligence security framework. In Proceedings of the Journal of Physics: Conference Series. IOP Publishing, 2021, Vol. 1948, p. 012004.
28. Diao, Y.; Zhai, N.; Miao, C.; Yu, Z.; Wei, X.; Yang, X.; Wang, M. Vulnerabilities in ai-generated image detection: The challenge of adversarial attacks. *arXiv preprint arXiv:2407.20836* **2024**.
29. Yaseen, A. Accelerating the SOC: Achieve greater efficiency with AI-driven automation. *International Journal of Responsible Artificial Intelligence* **2022**, *12*, 1–19.
30. Vesnic-Alujevic, L.; Nascimento, S.; Polvora, A. Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy* **2020**, *44*, 101961.
31. Hamon, R.; Junklewitz, H.; Sanchez, I.; et al. Robustness and explainability of artificial intelligence. *Publications Office of the European Union* **2020**, *207*.
32. Wang, D.; Churchill, E.; Maes, P.; Fan, X.; Shneiderman, B.; Shi, Y.; Wang, Q. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In Proceedings of the Extended abstracts of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–6.