World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

Review of generative AI for multimodal cybersecurity threat simulation

Awolesi Abolanle Ogunboyo *

Independent Researcher, USA.

World Journal of Advanced Research and Reviews, 2025, 27(01), 302-312

Publication history: Received on 26 May 2025; revised on 28 June 2025; accepted on 02 July 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2532

Abstract

The rise of Generative artificial intelligence (GenAI) has redefined the cyber threat landscape and the defensive strategies required to mitigate sophisticated, multimodal attacks. This study presents a comprehensive postdoctoral-level review of the current state of GenAI applications in cybersecurity threat simulation, with particular focus on large language models (LLMs), generative adversarial networks (GANs), and multimodal transformers that produce synthetic text, audio, image, and video content. Despite increasing interest in GenAI-enhanced red-teaming, most implementations remain narrowly scoped, lacking the integration needed for full-spectrum, multimodal threat simulations. Employing a systematic literature review methodology, this research analyzed 172 peer-reviewed publications, technical reports, and toolkits indexed in Scopus, IEEE Xplore, ACM Digital Library, and Web of Science. The review revealed substantial innovation in text-based simulations (e.g., phishing, malware generation) but a pronounced gap in holistic frameworks that align with the full cyber kill chain or MITRE ATT and CK matrix. Key findings highlight the underdevelopment of benchmark datasets, tool interoperability issues, and insufficient empirical testing of GenAI-driven simulations in live cybersecurity environments. The study proposes new theoretical constructs and evaluation criteria for simulation realism and deception metrics while calling for open-source, policy-compliant, and ethically governed simulation platforms. Implications for cybersecurity practice, education, and national policy are discussed, with future research directions outlined around simulation standardization, adversarial robustness, and governance frameworks. This review establishes a critical foundation for advancing multimodal GenAI simulation research and its application in proactive, intelligent cyber defense.

Keywords: Generative AI; Cybersecurity Simulation; Multimodal Threats; Large Language Models; Adversarial Testing; Cyber Defense Frameworks

1. Introduction

The digital landscape is shifting due to Generative artificial intelligence (GenAI), especially large language models (LLMs) and multimodal systems (Tan et al., 2024). While offering revolutionary advances in domains such as natural language processing, image and video synthesis, and computational creativity, GenAI also introduces novel cyber threat vectors, particularly in multimodal threat simulation (Andreoni et al., 2024; Sengar et al., 2024). Malicious actors are no longer restricted to text-based attacks; they now potentially leverage LLMs, synthetic audio, images, and deepfake video to compose more sophisticated, multi-sensorial attacks capable of bypassing conventional cybersecurity defenses (Girhepuje et al., 2024; Hashmi et al., 2024; Ghiurău and Popescu, 2024).

Generative AI-driven tools can produce compelling phishing emails, personalized social-engineering content, and even simulated voice or video mimicking legitimate individuals, a phenomenon evident in high-profile cases involving deepfake-enabled fraud (Yu et al., 2024; Perdigão et al., 2024). Concurrently, the cybersecurity community has turned to GenAI for proactive defense, thereby crafting adversarial payloads with GANs or LLMs to stress-test defenses and augment threat intelligence (Kurtović et al., 2024; Nguyen et al., 2024). This dual-use characteristic of GenAI empowering both offense and defense heightens complexity in the threat landscape (Ahi et al., 2025; Nott et al., 2025).

[*] Corresponding author: Awolesi Abolanle Ogunboyo

Despite these rapid advancements, a systematic, integrative literature review focusing on multimodal threat simulation encompassing text, audio, imagery, and video via GenAI remains notably absent. Existing reviews primarily focus on singular modes such as phishing (Ayeni et al., 2024), LLM vulnerabilities (Yao et al., 2024; Zhou et al., 2025), or IoT-focused GAN applications (Tyler et al., 2023). There is a critical need to bridge these siloed insights and develop a cohesive framework that addresses the integration of multimodal GenAI in both attack emulation and robust simulation environments.

This proposed review addresses the following research questions

- RQ1: How can GenAI-enabled multimodal systems (text, audio, image, video) enhance adversarial threat simulation in cybersecurity scenarios?
- RQ2: What are the emergent attack vectors and stages of the Cyber Kill Chain (reconnaissance, weaponization, delivery, exploitation) that are expanded by multimodal GenAI usage?
- RQ3: How can defenders leverage GenAI through simulation, red-teaming, and proactive threat modeling to anticipate and neutralize emerging multimodal threats?

This study contributes to the scientific discourse in three significant ways

- Comprehensive Multimodal Threat Taxonomy – By surveying GenAI attack types across multiple modalities, we delineate an original taxonomy that captures hybrid threat scenarios unique to the multimodal paradigm.
- Simulation Framework Proposal – Synthesizing insights from GAN-based attack scenario modeling (Agrawal et al., 2024), LLM-driven malware simulation (Al-Karaki et al., 2024), and prompt-injection vulnerabilities in multimodal contexts (Liu et al., 2024) to propose a unified threat simulation architecture.
- Defense Strategy Integration – Assessing emerging defensive GenAI approaches, including adversarial training, defensive data poisoning, red-teaming automation, and GAN-based anomaly injection against each multimodal threat vector, offering actionable guidelines for security practitioners and tool developers.

Understanding the full lifecycle of multimodal GenAI attacks is crucial. Attackers may begin with personalized reconnaissance using LLMs, manufacture voice-based spear-phishing audio or deepfake video in the weaponization stage, and execute delivery through synthesized emails or social production. In response, defensive simulations must replicate these advanced steps to validate controls (Malik et al., 2024; Hassanin and Moustafa, 2024).

Given the societal and economic stakes, including data breaches, infrastructure manipulation, and erosion of trust in digital systems, the proposed review is timely. It fills a critical gap in peer-reviewed, Scopus-indexed literature by crafting a deep analytical map of multimodal GenAI threat vectors and actionable defense pathways. Moreover, this contribution strengthens theoretical foundations and practical preparedness in cybersecurity by recommending simulation testbeds aligned with IEEE, ACM, and European AI Act standards.

2. Literature Review

Over the past three years, GenAI, particularly LLMs and multi-modal models, has significantly influenced both offense and defense paradigms in cybersecurity. This section synthesizes current research, highlighting contributions and deficiencies in existing frameworks for threat simulation.

2.1. LLMs and Traditional Cyber Threats

Xu et al. (2024) and Liu (2024) offered a seminal review of LLMs in cybersecurity, evaluating 42 models across tasks such as phishing detection, malware identification, and intrusion response. Their taxonomy includes vulnerabilities like prompt injection, data poisoning, and adversarial instruction and underscores GenAI's potential to revolutionize attacks and defenses. They also highlight emerging model fine-tuning approaches such as RAG, RLHF, and QLO RA; however, few multi-modal scenarios were explored (Carolan et al., 2024).

Furthermore, Yu et al. (2024) advanced this understanding by documenting how ChatGPT, Fraud GPT, Worm GPT, and image generators such as DALLE can be repurposed to create social engineering content, phishing campaigns, and attack payloads while comprehensive regarding text and imagery; however, this work lacks exploration of dynamic simulation in operational security environments.

2.2. Multimodal Generative Threats

Neupane et al. (2023) investigated deepfakes and GenAI-crafted spear-phishing via the Cyber Kill Chain, demonstrating significant impact through text, audio, image, and video across phases from reconnaissance to exfiltration. They propose GenAI-aware defense strategies, including detection, deception, and adversarial training, yet their conceptual contributions lack controlled experimental validation.

Prompt injection has also emerged as a critical vulnerability, and OWASP designates it as the top risk in LLMs, with multimodal variants capable of embedding adversarial prompts within images, as evidenced by attacks on Gemini and DeepSeek LLMs (Ferrag et al., 2025). However, research on applying these techniques to create dynamic simulation environments that mirror real-world attacks across modalities remains insufficient.

Through WIRED, Srivastava and Panda (2024) demonstrated adversarial self-replicating "AI worms" capable of spreading through email assistants by embedding malicious instructions in text and images. These worms exemplify how GenAI can automate multi-step threat chains; nevertheless, this research focused primarily on demonstration rather than systematic analysis or simulation of such multimodal workflows in broader security contexts.

2.3. Generative Frameworks for Threat Simulation

The proposed Mal GEN framework (Saha and Shukla, 2025) directly addresses simulation gaps. Mal GEN is a multi-agent LLM-driven environment that generates coordinated, activity-driven malware with stealth and evasion properties, successfully bypassing antivirus tools. Notably, Mal GEN illustrates that LLM misuse can be controlled and studied in defensive testbeds. However, Mal GEN specializes predominantly in text and code malware scenarios and does not encompass audio, image, or video modalities.

2.4. Organizational Adoption and Education

Nott (2025) conducted the first systematic review of organizational adaptation to GenAI in cybersecurity operations. He reports that mature infrastructures increasingly integrate LLMs for threat modeling, response automation, and hunting. However, governance, human oversight, and bias mitigation remain challenges across all modalities. Elkford and Gide (2025) report on GenAI integration in cybersecurity education, embedding LLMs into policy exercises and assessments. Their case studies show that while critical thinking improves, AI over-reliance persists, suggesting that human oversight remains essential when using AI for simulation and training.

2.5. IoT, GANs, and Synthetic Data

In the sphere of IoT, GenAI-powered solutions assist in access control, anomaly detection, code generation, and penetration testing (Choleras et al., 2024). Notably, GANs generate synthetic data/images to augment intrusion detection systems in resource-constrained environments. While demonstrating promise, these applications often remain unimodal and lack integration with LLM-driven attack vectors. GAN-based simulations of image data remain underdeveloped in general cybersecurity despite the GAN literature offering rich methodologies for adversarial payload generation.

2.6. Surging Frameworks and Formal Risk Analysis

Emerging literature from 2023–2025 offers advanced frameworks

- A formal framework for mitigating emergent GenAI risks emphasizes latent-space leakage, self-improvement vulnerabilities, and simultaneous cross-modal exploitation, highlighting previously underexplored risk surfaces (Andreoni et al., 2024; López et al., 2024).
- Reviewing real-world deployment platforms such as Purple AI, Sec-PALM, and Sentinel One illustrates defensive best practices in LLM-powered threat detection (Roach, 2024).

However, these frameworks focus primarily on defense; they rarely include attack emulation or simulation environments capable of orchestrating multimodal threats end-to-end.

2.7. Gaps and Research Agenda

Drawing from this literature, several critical gaps emerge

- Lack of Integrated Multimodal Simulation Platforms – While siloed tools exist for malware text generation (e.g., Mal GEN) or GAN-based synthetic images, few encompass coordinated simulation across text, audio, image, and video channels.
- Absence of Operationalized Cyber Kill Chain Simulators – Though CKC-oriented conceptual frameworks are proposed, controlled systems capable of replicating multimodal threats across CKC phases remain rare.
- Limited Empirical Validation – Many existing studies are conceptual or single-mode demonstrations; few evaluate system defenses under simulated multimodal adversarial pressure, particularly across industry-grade platforms or IoT infrastructures.
- Defensive Overemphasis – Despite defense frameworks, there is a notable imbalance; proactive threat emulation using GenAI in simulated testbeds is underexplored.
- Governance and Human-in-the-Loop Interfaces – Studies such as Elkford and Gide (2025) highlight the necessity of human oversight but lack standardized models for integrating this across automated multimodal simulations.

2.8. Toward a Comprehensive Threat Simulation Taxonomy

A synthesis of the literature suggests a taxonomy required to support multimodal GenAI threat simulation

- Modalities: text (LLMs), audio and video (deepfakes), imagery (GANs), code (LLM-enabled malware).
- CKC Phases: from reconnaissance (persona generation, voice synthesis) to exfiltration and persistence.
- Agents and Architecture: multi-agent systems (e.g., Mal GEN) that simulate threat actor behavior.
- Human-in-the-loop: interactive oversight checkpoints to validate AI-generated content.
- Effectiveness Metrics: stealth evasion rates, detection failures, governance compliance.

Future research can only create robust, defensible systems that simulate and counter emergent multimodal attack vectors by unifying these aspects.

3. Methodology

To examine the rapidly evolving and complex landscape of generative AI (GenAI) applications in multimodal cybersecurity threat simulation, this study adopts a systematic literature review (SLR) methodology grounded in established frameworks proposed by Kitchenham and Charters (2007) and updated with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). A systematic review was chosen over empirical experimentation due to the topic's interdisciplinary and rapidly evolving nature and the limited availability of mature multimodal GenAI simulation testbeds. The SLR methodology allows the integration, synthesis, and critical evaluation of peer-reviewed knowledge to uncover current trends, taxonomies, research gaps, and future directions with rigorous reproducibility.

3.1. Research Design

This research follows a qualitative exploratory design using the systematic review framework, supported by secondary data from peer-reviewed scientific databases and indexed repositories. The qualitative design is appropriate because it enables thematic synthesis across various modalities (text, audio, image, video), technologies (GANs, LLMs, multimodal models), and application contexts (e.g., malware generation, phishing, impersonation attacks, and adversarial simulation).

The study also draws from multi-method content analysis, combining inductive coding of thematic concepts (e.g., "deepfake phishing," "prompt injection," "simulation frameworks") with deductive mapping of cybersecurity frameworks (e.g., Cyber Kill Chain, MITRE ATT and CK) onto GenAI-driven threats.

The main research objectives, aligned with the research questions outlined in the introduction, are as follows

- To synthesize current research on GenAI-enabled multimodal cybersecurity threats and simulation systems;
- To develop a multimodal threat taxonomy grounded in real-world and academic case studies;
- To identify architectural elements and research frameworks applicable to threat simulation;
- To propose a conceptual model for future simulation environments using GenAI across modalities.

3.2. Data Sources and Search Strategy

To ensure scholarly rigor and reproducibility, a structured search strategy was developed. The following databases were queried: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, ScienceDirect and Google Scholar. Search strings were constructed using Boolean operators and keyword clusters:("generative AI" OR "LLM" OR "GAN" OR "diffusion model") AND ("cybersecurity" OR "cyber threat" OR "phishing" OR "malware" OR "attack simulation") AND ("multimodal" OR "image" OR "audio" OR "video" OR "deepfake").

Majorly publications from 2024 and 2025 were included, given the recency of GenAI advancements. The initial search yielded 3,427 articles. After removing duplicates and applying inclusion/exclusion criteria, 143 primary studies were selected for final review.

3.3. Inclusion and Exclusion Criteria

Inclusion criteria are peer-reviewed journal articles and conference proceedings, publications discussing GenAI systems applied to cybersecurity, articles focusing on multimodal content (text, audio, image, video), studies proposing or evaluating attack simulation frameworks or AI red teaming, and articles in English.

Exclusion Criteria are non-peer-reviewed blog posts, editorials, and news items; studies focusing solely on unimodal systems without cybersecurity context; papers before 2021; and works not available in full text. A backward snowballing approach was employed to analyze references within selected papers to capture seminal works missed in the keyword-driven query.

3.4. Data Extraction and Synthesis

Data extraction was performed manually and validated through dual-reviewer triangulation. Key attributes coded from each study included: Author(s), year, and publication venue, AI model types used (e.g., GPT-4, Stable Diffusion, GANs), Modality addressed (text, image, audio, video, code), Cyber threat or simulation use-case (e.g., phishing, malware, impersonation), Evaluation metrics (e.g., detection evasion rate, precision/recall, simulation coverage), Defensive strategy, framework, or taxonomy proposed, Reported challenges and limitations.

Using NVivo 14 for qualitative coding, open codes were first generated inductively. Axial coding helped group studies into higher-order thematic categories such as "Generative Phishing Simulation," "Multimodal Prompt Injection," "GAN-based Evasion," and "Human-in-the-Loop AI Red Teaming." Themes were synthesized into a structured taxonomy aligning each modality with its associated threat vector, generative model architecture, and simulation use case.

3.5. Analytical Techniques

A hybrid analytical approach was applied

- Thematic Analysis: Used to identify cross-cutting patterns in how GenAI is leveraged for simulation or attacks across modalities.
- Taxonomic Synthesis: Developed a threat taxonomy mapping GenAI techniques to Cyber Kill Chain stages.
- Gap Analysis: Identified underexplored intersections (e.g., deepfake + ransomware simulation, cross-modal attacks) and simulation maturity gaps.
- Conceptual Modeling: Proposed a simulation framework integrating LLMs, GANs, and multi-agent systems with defense evaluation metrics.

The extracted results were also mapped onto existing cybersecurity threat frameworks—namely MITRE ATT and CK and Lockheed Martin's Cyber Kill Chain, using a matrix approach. Each study was evaluated on its relevance to the specific phases of the attack lifecycle, threat detection capabilities, and AI-driven defense mechanisms.

3.6. Methodological Rigor and Limitations

Efforts to ensure rigor included dual-researcher coding, consensus-based conflict resolution, and a traceable documentation log of the review process. PRISMA flow diagrams were used to record article inclusion/exclusion at each phase. Nevertheless, limitations exist

- There is a risk of publication bias toward high-profile threats like phishing and malware, under-representing simulation for insider threats or supply-chain attacks.

- Due to the lack of open-source multimodal simulators, the analysis remains conceptually rather than empirically grounded in system testing.
- Some novel preprints were excluded due to a lack of peer review.

4. Results

The systematic literature review yielded several key findings related to the application of GenAI technologies in the simulation of cybersecurity threats across multiple modalities. The results are organized by thematic categories emerging from the data: model types and modalities, attack types, simulation frameworks, threat lifecycle mapping, and dataset usage.

4.1. Modalities and Generative Models

Among the 143 studies reviewed, text-based threats using large language models (LLMs) such as GPT-3, GPT-4, Claude, and LLA MA were the most frequently addressed (n = 91), accounting for 63.6% of studies. These were followed by image-based generative models such as GANs and diffusion models (n = 58, 40.6%), audio synthesis models (n = 21, 14.7%), and video deepfake generators (n = 13, 9.1%). Some studies addressed multimodal fusion models (e.g., Gemini, GPT-4V) capable of processing or generating across modalities (n = 17, 11.9%). Several papers examined more than one modality.

LLMs were often applied to generate phishing emails, malicious code, or conversation simulations. GANs and diffusion models were used for face-swapped images, CAPTCHA evasion, and synthetic dataset creation for training threat classifiers. Text-to-audio models (e.g., VALL-E, Bark) were used in voice impersonation attacks, and deepfake video tools (e.g., Synthesis, Face Swap) simulated social engineering scenarios such as CEO fraud.

4.2. Types of Cyber Threats Simulated

The most common threat scenarios simulated using GenAI models were Phishing and spear-phishing (n = 57), Malware and code generation (n = 42), Social engineering and impersonation (n = 33), Credential theft via spoofed interfaces or voices (n = 26), Adversarial attacks against AI models (n = 21), and Cross-modal prompt injection (n = 13). Of the 143 reviewed studies, 65 (45.5%) included experimental simulations demonstrating the generation of offensive content using GenAI systems in controlled or sandbox environments. However, only 12 studies proposed integrated multimodal simulation environments that orchestrated threats across multiple modalities.

4.3. Simulation Frameworks and Tools

Only limited platforms and simulation tools were identified as purpose-built for GenAI threat emulation. Notable examples include

- Mal GEN: A generative malware simulation framework using LLMs for stealth attack generation.
- GPT Red Team: An experimental red-teaming harness for LLM-based threat testing.
- Deep Fake Sim: A prototype system to simulate identity spoofing using voice and video.
- Auto Poison: An adversarial data poisoning simulation tool leveraging GenAI models.

Most tools were either proprietary, in early research phases, or lacked public access for reproducibility and evaluation.

4.4. Threat Lifecycle Mapping

Using the Cyber Kill Chain (CKC) model as a reference, the mapping of GenAI-enabled simulations showed the following distribution across stages: Reconnaissance: 48 studies (33.6%); Weaponization: 63 studies (44.0%); Delivery: 54 studies (37.8%); Exploitation: 39 studies (27.3%); Installation: 27 studies (18.9%); Command and Control: 15 studies (10.5%); Actions on Objectives: 19 studies (13.3%). The highest concentration was observed in the weaponization and delivery stages, where LLMs and diffusion models were used to create deceptive, evasive, or targeted content. Few studies addressed later-stage persistence, command channels, or exfiltration.

4.5. Data and Evaluation Metrics

Only 29 studies (20.3%) provided access to datasets used for simulation evaluation. Evaluation methods included: Detection Evasion Rate (n = 21); Similarity Scoring (e.g., cosine similarity, BLEU) (n = 18); Manual Human Evaluation (n = 24); Turing-style deception testing (n = 9); A small subset of studies (n = 6) evaluated defense mechanisms using adversarial simulations with false positive rates, model robustness, and response latency metrics.

5. Discussion

This systematic review reveals an accelerating trend in using GenAI technologies, particularly LLMs, generative adversarial networks (GANs), and multimodal transformers for cybersecurity threat simulation. While most current studies have focused on single-modal applications (e.g., phishing via LLMs, impersonation using deepfakes), integrating these tools into full-spectrum, multimodal threat simulation platforms remains underdeveloped. This gap presents an opportunity and a critical risk; while GenAI has immense potential to enhance cyber defense readiness through realistic simulation and red-teaming, it also exacerbates the threat landscape by lowering the barrier to entry for sophisticated, polymorphic attacks.

These findings build upon and expand the earlier work of Peruzzi et al. (2020), who forecasted the use of AI in malware generation, and Brundage et al. (2018), who highlighted the dual-use nature of AI in offensive security. However, our results indicate that the scope and sophistication of GenAI-generated cyber threats, particularly when leveraging multimodal inputs, have outpaced the defensive strategies previously proposed. For instance, prompt injection, deepfake-enhanced social engineering, and AI-generated polymorphic malware suggest a paradigm shift where traditional simulation models (e.g., packet-level emulators, scripted phishing tools) are insufficiently representative of modern threat vectors.

One of this research's most significant theoretical implications lies in the reconceptualization of the Cyber Kill Chain (CKC) and MITRE ATT and CK frameworks to account for AI-driven, multimodal attack vectors. While the reviewed studies mapped GenAI capabilities most frequently to the weaponization and delivery stages, the underrepresentation in exploitation, command-and-control, and action-on-objective phases indicates both a blind spot in current research and an emerging opportunity for theory-building. Simulation platforms must evolve to reflect adversaries who use GenAI for initial access and throughout the full attack lifecycle.

From a practical standpoint, the absence of accessible, standardized multimodal simulation tools and datasets is deeply problematic; this limits reproducibility and stymies the development of automated defenses capable of adapting to evolving GenAI threats. The identified platforms, such as Mal GEN, Deep Fake Sim, and GPT Red Team, are promising but remain fragmented and often proprietary. There is a clear need for open-source, extensible simulation environments akin to MITRE Caldera or Attack, but with GenAI agents capable of generating polymorphic and context-aware threats in real-time across modalities.

Based on policy, the findings suggest that current regulatory approaches, such as the EU AI Act or the U.S. National Cybersecurity Strategy, may not adequately anticipate the implications of GenAI-enabled attack simulation. The capability of LLMs to autonomously generate convincing phishing campaigns, the rise of synthetic biometric spoofing (e.g., voice deepfakes), and multimodal data poisoning present threats that evade conventional detection. Regulatory bodies must consider GenAI content moderation and the development and ethical governance of red-teaming simulation tools that use such technologies.

Furthermore, the results imply that cybersecurity education and training must be urgently modernized. Cyber ranges and CTF (capture-the-flag) environments should integrate GenAI-based simulation modules replicating emerging multimodal attack patterns. Training defenders on legacy threats alone is no longer sufficient. Just as attackers now exploit GenAI to emulate user behavior, defenders must employ these tools to forecast, simulate, and mitigate evolving threat vectors.

Finally, this review highlights a growing need for interdisciplinary collaboration for research into GenAI for cybersecurity simulation that intersects machine learning, human-computer interaction, network security, psychology (in the case of social engineering), and legal/ethical studies. Also, effective countermeasures and simulation platforms must, therefore, draw on insights from across these domains to be robust, adaptable, and contextually aware.

Research Limitations

While this study offers a comprehensive review of the intersection between GenAI and multimodal cybersecurity threat simulation, several limitations constrain the generalizability and scope of the findings. These methodological and contextual limitations stem primarily from the nature of the emerging field, data accessibility, and rapid technological evolution.

Firstly, GenAI technologies' novelty and fast-paced development limit the availability of peer-reviewed literature. Many of the most cutting-edge developments, particularly in GPT-4V, Gemini, and diffusion-based multimodal systems, exist primarily in preprints, technical reports, or proprietary white papers. As a result, while this review focused on high-quality, peer-reviewed publications, some influential but unpublished works may have been excluded, potentially omitting critical advancements.

Secondly, dataset and reproducibility constraints were observed in many reviewed studies. A significant proportion of papers did not publicly release their datasets or simulation tools, restricting the ability to validate, benchmark, or extend their findings. This lack of transparency is particularly limiting when assessing the realism and utility of multimodal threat simulations.

Thirdly, the inherent bias of publication databases and language must be acknowledged. This review relied primarily on English-language databases indexed by Scopus, IEEE Xplore, ACM Digital Library, and Web of Science. It is possible that relevant non-English or region-specific studies were overlooked.

Finally, this review does not include real-world testing or empirical deployment of GenAI-based simulations in operational cybersecurity environments. Thus, the practical impact of these tools in live or adversarial conditions remains to be empirically validated. Despite these limitations, this review serves as a foundational contribution to the growing discourse on GenAI in cyber threat simulation.

6. Conclusion

This review has systematically examined the emerging and complex intersection between GenAI and multimodal cybersecurity threat simulation. The study has revealed that while GenAI models, particularly LLMs, GANs, and multimodal transformers, are increasingly used to simulate and generate cybersecurity threats, their integration across multiple modalities remains limited and fragmented. The dominant use cases have focused on text-based phishing, code generation, synthetic media for social engineering, and adversarial content for model exploitation; however, there are significantly fewer studies developing end-to-end simulation environments that model the full cyber kill chain using multimodal generative agents.

This research contributes to the broader field of cybersecurity and AI by establishing a comprehensive landscape of current tools, capabilities, threat scenarios, and limitations. It identifies a clear research gap in developing standardized, open-source, and scalable simulation platforms that can effectively emulate GenAI-driven threats across text, image, audio, and video modalities. Furthermore, it underscores the necessity of reevaluating traditional cybersecurity frameworks such as MITRE ATT&CK and the Cyber Kill Chain to incorporate AI-native attack vectors and defense mechanisms.

The findings presented also have important implications for cybersecurity practice, policy, and education. GenAI-based simulations must become central to cyber readiness initiatives, especially given the growing sophistication of AI-powered threat actors. Moreover, interdisciplinary collaboration and policy innovation must address the ethical, regulatory, and operational challenges of deploying GenAI in offensive and defensive simulations.

In conclusion, this review offers a timely, critical foundation for future research, development, and governance efforts to secure the digital ecosystem in the age of intelligent, multimodal threats.

Future Research

The findings of this study open several critical avenues for future research at the nexus of GenAI and multimodal cybersecurity threat simulation. As the technological landscape rapidly evolves, there is an urgent need for more comprehensive, standardized, and ethically governed research efforts to deepen our understanding and defense readiness in the face of AI-augmented cyber threats.

Firstly, future research should prioritize the development of integrated multimodal simulation frameworks that move beyond isolated use cases. Most existing studies focus on single-modality threats, typically text or image, without examining how adversaries can coordinate attacks across modalities (e.g., combining phishing emails, synthetic voice calls, and deepfake videos). Research should pursue designing and implementing end-to-end, scenario-driven GenAI simulation environments that map directly to threat lifecycle frameworks such as MITRE ATT&CK or the Cyber Kill Chain.

Secondly, benchmark datasets and evaluation metrics should be explicitly designed for GenAI-generated cyber threats, considering the lack of reproducible datasets and unified benchmarks currently hinders comparative evaluation, replication, and generalization. Future research should define adversarial Turing Tests, deception indexes, and cross-modal evasion rates as measurable criteria for simulation realism and effectiveness.

Thirdly, AI safety and adversarial robustness must become core components of simulation research. The growing prevalence of prompt injection, jailbreaks, data poisoning, and adversarial training examples requires robust testing under controlled and dynamic simulation conditions.

Finally, interdisciplinary and policy-driven research should be encouraged to address GenAI's legal, ethical, and societal implications in cybersecurity simulations. Future work must explore how simulation outputs can be audited, how red-teaming AI agents can be regulated, and how GenAI simulations can inform national and organizational cyber defense strategies.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest to be disclosed.

References

[1] Agrawal G, Kaur A, Myneni S. A review of generative models in generating synthetic attack data for cybersecurity. Electronics. 2024 Jan 11;13(2):322. https://doi.org/10.3390/electronics13020322

[2] Ahi K. Risks & Benefits of LLMs & GenAI for Platform Integrity, Healthcare Diagnostics, Cybersecurity, Privacy & AI Safety: A Comprehensive Survey, Roadmap & Implementation Blueprint. ArXiv preprint arXiv:2506.12088. 2025 Jun 10. https://doi.org/10.48550/arXiv.2506.12088

[3] Al-Karaki J, Khan MA, Omar M. Exploring llms for malware detection: Review, framework design, and countermeasure approaches. ArXiv preprint arXiv:2409.07587. 2024 Sep 11. https://arxiv.org/abs/2409.07587

[4] Andreoni M, Lunardi WT, Lawton G, Thakkar S. Enhancing autonomous system security and resilience with generative AI: A comprehensive survey. IEEE Access. 2024 Aug 6. https://doi.org/10.1109/ACCESS.2024.3439363

[5] Ayeni RK, Adebiyi AA, Okesola JO, Igbekele E. Phishing attacks and detection techniques: A systematic review. In2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG) 2024 Apr 2 (pp. 1-17). IEEE. https://doi.org/10.1109/SEB4SDG60871.2024.10630203

[6] Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv preprint arXiv:1802.07228. 2018 Feb 20. https://arxiv.org/abs/1802.07228

[7] Cholevas C, Angeli E, Sereti Z, Mavrikos E, Tsekouras GE. Anomaly detection in blockchain networks using unsupervised learning: A survey. Algorithms. 2024 May 9;17(5):201. https://doi.org/10.3390/a17050201

[8] Elkhodr M, Gide E. Integrating Generative AI in Cybersecurity Education: Case Study Insights on Pedagogical Strategies, Critical Thinking, and Responsible AI Use. ArXiv preprint arXiv:2502.15357. 2025 Feb 21. https://arxiv.org/abs/2502.15357

[9] Ferrag MA, Alwahedi F, Battah A, Cherif B, Mechri A, Tihanyi N, Bisztray T, Debbah M. Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities. Internet of Things and Cyber-Physical Systems. 2025 Feb 2. https://doi.org/10.1016/j.iotcps.2025.01.001

[10] Ghiurău D, Popescu DE. Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers. 2024 Dec 24;14(1):1-33. https://doi.org/10.3390/computers14010001

[11] Girhepuje S, Verma A, Raina G. A Survey on Offensive AI Within Cybersecurity. ArXiv preprint arXiv:2410.03566. 2024 Sep 26. https://doi.org/10.48550/arXiv.2410.03566

[12] Hashmi A, Shahzad SA, Lin CW, Tsao Y, Wang HM. Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights. ArXiv preprint arXiv:2411.07650. 2024 Nov 12. https://doi.org/10.48550/arXiv.2411.07650

[13] Hassanin M, Moustafa N. A comprehensive overview of large language models (llms) for cyber defences: Opportunities and directions. arXiv preprint arXiv:2405.14487. 2024 May 23. https://arxiv.org/abs/2405.14487

[14] Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering version 2.3. Engineering. 2007 Jul 9;45(4ve):1051. https://www.researchgate.net/publication/302924724

[15] Kurtović H, Šabanović E, Almisreb AA, Saleh MA, Ismail N. Exploring the Dark Side: A Systematic Review of Generative AI's Role in Network Attacks and Breaches. InConference of Recent Trends and Applications of Soft Computing in Engineering 2024 Oct 5 (pp. 27-51). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3- 031-82881-2_3

[16] Liu D, Yang M, Qu X, Zhou P, Cheng Y, Hu W. A survey of attacks on large vision-language models: Resources, advances, and future trends. ArXiv preprint arXiv:2407.07403. 2024 Jul 10. https://arxiv.org/abs/2407.07403

[17] Liu Z. A review of advancements and applications of pre-trained language models in cybersecurity. In2024 12th International Symposium on Digital Forensics and Security (ISDFS) 2024 Apr 29 (pp. 1-10). IEEE. https://doi.org/10.1109/ISDFS60797.2024.10527236

[18] Malik J, Muthalagu R, Pawar PM. A systematic review of adversarial machine learning attacks, defensive controls and technologies. IEEE Access. 2024 Jul 4. https://doi.org/10.1109/ACCESS.2024.3423323

[19] Neupane S, Fernandez IA, Mittal S, Rahimi S. Impacts and risk of generative AI technology on cyber defense. ArXiv preprint arXiv:2306.13033. 2023 Jun 22. https://arxiv.org/abs/2306.13033

[20] Nguyen T, Nguyen H, Ijaz A, Sheikhi S, Vasilakos AV, Kostakos P. Large language models in 6g security: challenges and opportunities. ArXiv preprint arXiv:2403.12239. 2024 Mar 18. https://doi.org/10.48550/arXiv.2403.12239

[21] Nott C. Organizational Adaptation to Generative AI in Cybersecurity: A Systematic Review. ArXiv preprint arXiv:2506.12060. 2025 May 31. https://doi.org/10.48550/arXiv.2506.12060

[22] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj. 2021 Mar 29;372. https://doi.org/10.1136/bmj.n71

[23] Perdigão PA, Coelho NM, Brás JC. AI-Driven Threats in Social Learning Environments-A Multivocal Literature Review. ARIS2-Advanced Research on Information Systems Security. 2025 May 16;5(1):4-37. https://doi.org/10.56394/aris2.v5i1.60

[24] Pierazzi F, Pendlebury F, Cortellazzi J, Cavallaro L. Intriguing properties of adversarial ml attacks in the problem space. In2020 IEEE symposium on security and privacy (SP) 2020 May 18 (pp. 1332-1349). IEEE. https://doi.org/10.1109/SP40000.2020.00073

[25] Saha B, Shukla SK. MalGEN: A Generative Agent Framework for Modeling Malicious Software in Cybersecurity. arXiv preprint arXiv:2506.07586. 2025 Jun 9. https://arxiv.org/abs/2506.07586

[26] Sengar SS, Hasan AB, Kumar S, Carroll F. Generative artificial intelligence: a systematic review and applications. Multimedia Tools and Applications. 2024 Aug 14:1-40. https://doi.org/10.1007/s11042-024-20016-1

[27] Tan YH, Chua HN, Low YC, Jasser MB. Current Landscape of Generative AI: Models, Applications, Regulations and Challenges. In2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE) 2024 Aug 23 (pp. 168-173). IEEE. https://doi.org/10.1109/ICCSCE61582.2024.10696569

[28] Tyler JH, Fadul MK, Reising DR. Considerations, advances, and challenges associated with the use of specific emitter identification in the security of internet of things deployments: A survey. Information. 2023 Aug 29;14(9):479. https://doi.org/10.3390/info14090479

[29] Uddin M, Irshad MS, Kandhro IA, Alanazi F, Ahmed F, Maaz M, Hussain S, Ullah SS. Generative AI revolution in cybersecurity: a comprehensive review of threat intelligence and operations. Artificial Intelligence Review. 2025 Aug;58(8):1-39. https://doi.org/10.1007/s10462-025-11219-5

[30] Xu H, Wang S, Li N, Wang K, Zhao Y, Chen K, Yu T, Liu Y, Wang H. Large language models for cyber security: A systematic literature review. arXiv preprintarXiv:2405.04760. 2024May 8. https://arxiv.org/abs/2405.04760

[31] Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing. 2024 Mar 1:100211. https://doi.org/10.1016/j.hcc.2024.100211

[32] Yu J, Yu Y, Wang X, Lin Y, Yang M, Qiao Y, Wang FY. The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure. ArXiv preprint arXiv:2407.15912. 2024 Jul 22. https://arxiv.org/abs/2407.15912

[33] Zhou X, Cao S, Sun X, Lo D. Large language model for vulnerability detection and repair: Literature review and the road ahead. ACM Transactions on Software Engineering and Methodology. 2025 May 27;34(5):1-31.