https://doi.org/10.58086/w1yd-0829

GENERATIVE AI MUTABILITY IN CYBERSECURITY: A BIBLIOMETRIC REVIEW

Pedro Oliveira[1]✉, **Mário Dias Lousã**[1,2], **José Carlos Morais**[1,3]

[1] Instituto Superior Politécnico Gaya (ISPGAYA), Portugal.
[2] Insight - Piaget Research Center for Ecological Human Developement, Portugal.
[3] CEOS.PP, ISCAP, Polytechnic of Porto, Portugal.
✉ Corresponding authors**:** ispg2024103478@ispgaya.pt

Abstract

The expansion of generative AI (GenAI) is forcing us to rethink cybersecurity, expanding both defensive automation and scalable offensive techniques. This bibliometric review maps the change driven by GenAI in cybersecurity through a PRISMA-guided selection of 154 documents from The Lens (20 December 2025). The current state is summarized by scientific mapping results (co-authorship, co-word, and co-citation networks, and thematic evolution) to identify dominant architectures, thematic clusters, and collaboration patterns, and implications for governance and auditing. We note the exponential growth of publications in 2022. We notice the trend. The authors group publications into several architectures: large language models (LLMs), generative networks (GANs), and diffusion models. These focus on common topics, (i) large-scale phishing and social engineering, (ii) mutability, obfuscation, and adversarial evasion of malware, and (iii) intrusion detection and cyber threat intelligence using synthetic data. Co-citation networks and keywords show that adversarial robustness, red teaming, and benchmarking are interconnected. We find that explainability and human-in-the-loop defense exist as minor but growing topics. One risk is the BlackMamba case, which transmits an LLM-assisted pipeline capable of generating more than 10,000 semantically identical but structurally distinct mutations per hour and achieving a 98.2% evasion rate against commercial EDR solutions. Risk mitigation should prioritize benchmarking and standardized reporting, continuous red teaming, and telemetry monitoring, incorporated into dynamic audit frameworks, supported by explicit international governance for high-risk GenAI cybersecurity applications.

Keywords: Adversarial AI; AI-driven cyber threats; Polymorphic and metamorphic malware; Synthetic data; Intrusion detection systems.

1. Introduction

Cybersecurity faces a rapidly evolving threat landscape in which adversaries iterate faster than defenders. GenAI [including GANs, large language models (LLMs), and diffusion models] alters this balance by reducing the cost of content generation, code transformation, and scalable automation (Gupta et al., 2023), thereby intensifying routine and advanced cyber operations. A concrete illustration of the mutability enabled by GenAI is BlackMamba, in which an LLM is used to continuously generate and rewrite a keylogger, helping it evade signature- and behavior-based controls such as endpoint detection and response (EDR) (Ibrar et al., 2025). Such cases exemplify how GenAI supports polymorphic and metamorphic behaviors that undermine traditional detection pipelines (Javaheri et al., 2021). This bibliometric review maps the research landscape on GenAI and mutability in cybersecurity, using n=154 documents (2018–20 December 2025) indexed in The Lens and selected under PRISMA. The objective is to quantify growth, identify influential sources and actors, and characterize dominant themes in defensive and offensive research trajectories. This bibliometric analysis investigates the following research questions:

RQ1: Identify scientific production over time, main sources, authors, and countries leading the field?

RQ2: Which thematic clusters and GenAI architectures dominate GenAI research in cybersecurity?

RQ3: How is the literature distributed between defensive applications and offensive/evasion-oriented research?

RQ4: What limitations and research gaps emerge for future work on GenAI mutability in cybersecurity?

This article is structured as follows. Section two summarizes the conceptual background on GenAI architectures and mutability. Section three details the PRISMA-guided bibliometric methodology, including the search strategy, inclusion criteria, and analytical tools. Section four presents the bibliometric results (production trends, sources, actors, and thematic structures). Section five discusses the implications, including attack-defense asymmetry and governance challenges. Section six concludes with limitations and future directions.

2. Literature review

GenAI is characterized by scalable content and code generation, enabling both defensive analytics and offensive capability. In cybersecurity, the same generative mechanisms that support detection and simulation can also be repurposed to evade controls through continuous variation (mutability) (Gazzan et al., 2025).

GenAI architectures appear in the corpus primarily as (i) GANs, which learn to synthesize realistic samples via generator–discriminator competition and are frequently used for synthetic security data and adversarial example generation (Mahmoudi et al., 2025; Pei et al., n.d.); (ii) LLMs and Transformer variants, which support Cyber Threat Intelligence (CTI), phishing and social-engineering automation, and code transformation/rewriting (Balasubramanian et al., 2025); and (iii) diffusion models, an emerging class increasingly used for higher-fidelity synthetic generation and augmentation (Yazdani et al., 2025). Across architectures, the security relevance is operational: models enable automation, scalability, and rapid iteration under constrained defender visibility. GANs: Composed of a generator and a discriminator in competition, GANs are widely used to generate adversarial attack traffic that "fools" the detector, forcing it to learn more robust patterns. In Internet of Vehicles (IoV) environments (Dunmore et al., 2023), GANs are crucial for simulating attacks and strengthening intrusion detection. GenAI facilitates the creation of malware that alters its signature to evade detection (Labaca-Castro, 2023).

Metamorphism via LLMs: "Uncensored" or jailbroken LLMs can rewrite the syntax of malicious code (e.g., renaming variables, inserting dead code) while maintaining the original semantics. This automates the creation of zero-day variants (Acosta-Bermejo et al., 2025).

Mutability refers to the capability of malicious artifacts to change form while preserving function. Polymorphism typically modifies surface-level representations (e.g., encryption/packing and variable changes) to break signatures, whereas metamorphism rewrites code structure to alter control flow and semantics (Guo, 2023). LLMs can accelerate metamorphic behavior by rewriting payloads on demand, including renaming, refactoring, and generating variant implementations. BlackMamba exemplifies LLM-enabled metamorphism: a keylogger is generated and iteratively modified to reduce detection likelihood, challenging EDR pipelines that rely on known behavioral patterns (Ibrar et al., 2025). In parallel, GAN-based approaches support adversarial attacks by producing near-indistinguishable perturbations that degrade malware or intrusion classifiers, exposing vulnerabilities in AI-based

defense (Liu et al., 2025). Defensively, GenAI is used to address data scarcity and imbalance through synthetic data generation for IDS training, attack simulation, and IoT/edge scenarios where labeled telemetry is limited (Kumar et al., 2025). However, dual-use nature implies that improved generation and simulation capabilities must be paired with rigorous evaluation, adversarial testing, and governance to prevent the same tooling from scaling offensive operations (Radanliev, 2025).

Adversarial attacks: GANs are used to create adversarial examples — samples of malware with minimal perturbations, invisible to humans but causing misclassification by deep learning models (Gaber et al., 2024).

3. Methodology

The final dataset consists of n=154 records after applying the PRISMA protocol (Figure 1). The data served as the basis for obtaining descriptive indicators, co-occurrence structures, and co-citation networks using Bibliometrix (R) and VOSviewer.

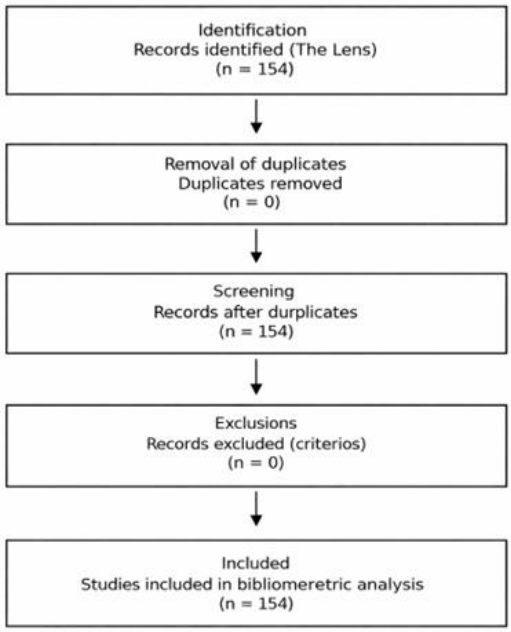

Fig 1. PRISMA diagram
Source: Authors' elaboration.

3.1. Data source and analysis period

The search was executed in The Lens on 20 December 2025. Data collection was carried out on The Lens platform, which aggregates scientific publications from multiple publishers and indexing databases (IEEE Xplore, ACM Digital Library, Scopus, SpringerLink), and the records were exported on the same date, yielding N=154 documents published between 2018 and 20 December 2025.

3.2. Research strategy

The search strategy was executed in the "Full Text" field of The Lens platform, using the following query: ("generative AI" AND "cybersecurity") OR ("generative AI" AND "generative adversarial networks") OR ("metamorphic" AND "generative AI") OR ("generative AI" AND "polymorphic") OR ("generative AI" AND "red team") OR ("generative AI" and "malware") OR ("generative adversarial networks" AND "metamorphic") OR ("generative adversarial networks" AND "polymorphic")]. The following filters were applied: "Document Type: Journal Article or Conference Article or Book Chapter"; "Language: English"; and "Period: 2018-20 December 2025". The subject matter: "Computer Science Applications, Artificial Intelligence, Information Systems, Computer Networks and Communications, General Computer Science, Molecular Biology, Software, Library and Information Sciences, General Engineering, Physical and Theoretical Chemistry". The exported records included complete bibliographic metadata, citations, author keywords, and abstracts.

3.3. Inclusion and exclusion criteria

We included studies published from early 2018 to 20 December 2025 (based on The Lens database) that explicitly address the use of generative models in creating threats or defense mechanisms, with experimental validation. Articles on generative AI unrelated to security were excluded, as were cybersecurity studies that only use classical discriminative techniques. Technical reports and non-peer-reviewed documents were also excluded, unless explicitly referenced in systematic reviews. No records were excluded during screening (n=0)

because the search string and filters were highly specific, yielding a corpus already aligned with the intersection of GenAI and cybersecurity.

3.4. Processing and tools

The files were cleaned and standardized in Microsoft Excel, removing possible duplicates and unifying variants of keywords and author names. The data was then analyzed in the Bibliometrix (R) package to generate indicators of scientific output, citations, sources, authors, countries, and term co-occurrence networks. VOSviewer was used to visualize network maps, density maps, and temporal overlap, while additional graphs were generated in The Lens itself.

4. Bibliometric Results

Analysis of empirical data reveals clear trends that answer research questions.

4.1. Scientific output and life cycle of the topic

Across 2018–20 December 2025, annual output accelerates markedly after 2022 (Figure 2), consistent with the diffusion of LLMs (e.g., ChatGPT) into mainstream security practice and research.

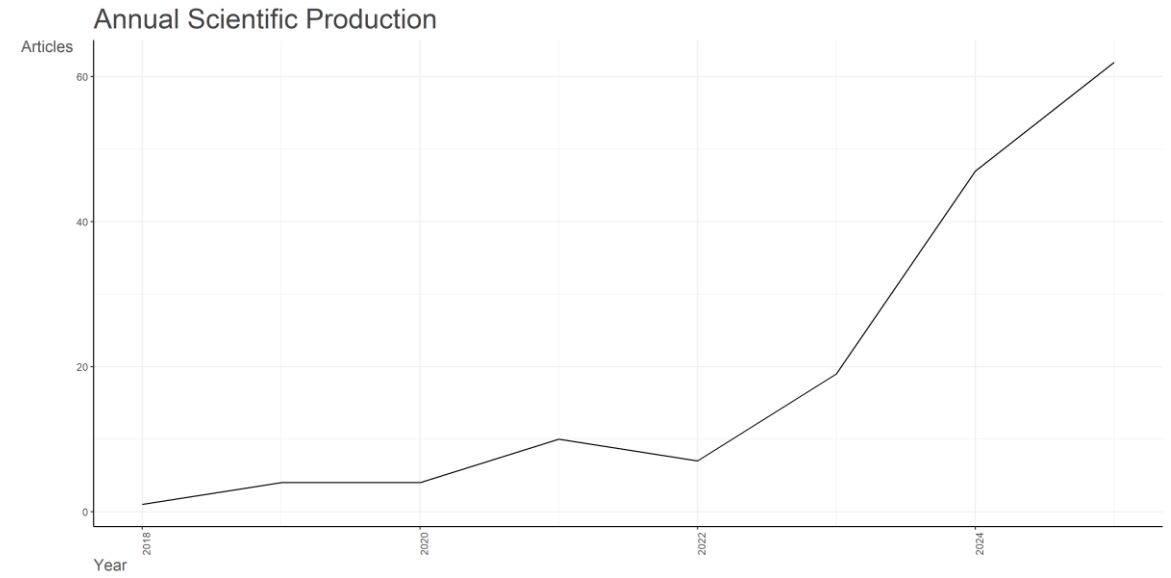

Fig 2. Annual scientific output
Source: output from Bibliometrix (2025).

The life-cycle indicators suggest a growth phase rather than maturation, while citation indicators (Figure 3) should be interpreted cautiously for recent years due to shorter accumulation windows. The decline in average citations per article in recent years reflects the temporal citation accumulation effect, not reduced topic relevance.

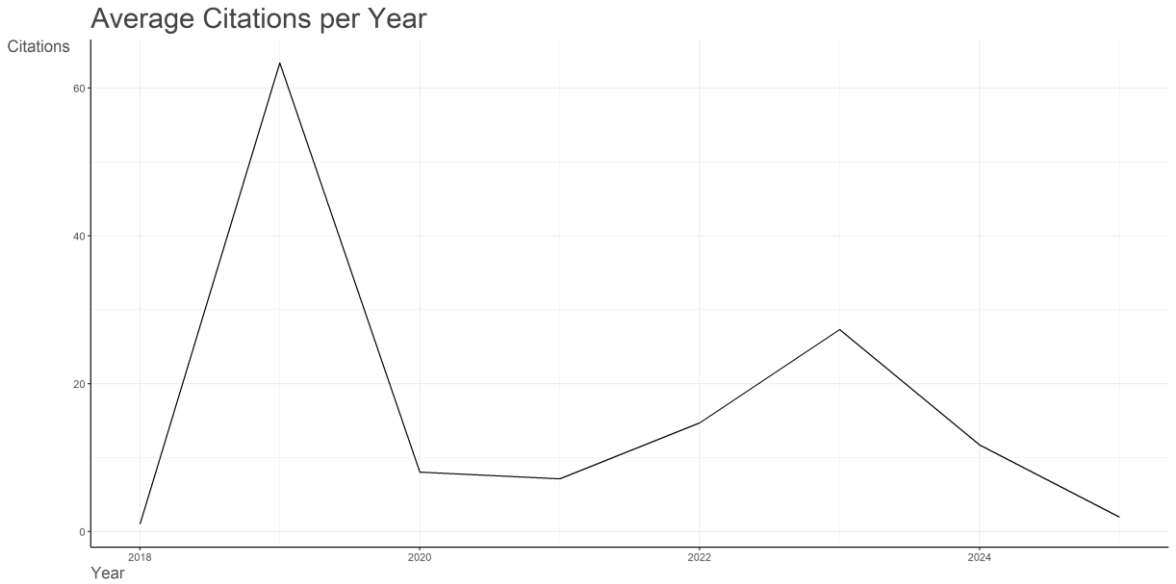

Fig 3. Graph showing average citations per article and per year
Source: output from Bibliometrix (2025).

4.2. Types of documents

The corpus is dominated by journal articles and conference proceedings, indicating both rapid dissemination and consolidation of findings (Figure 4).

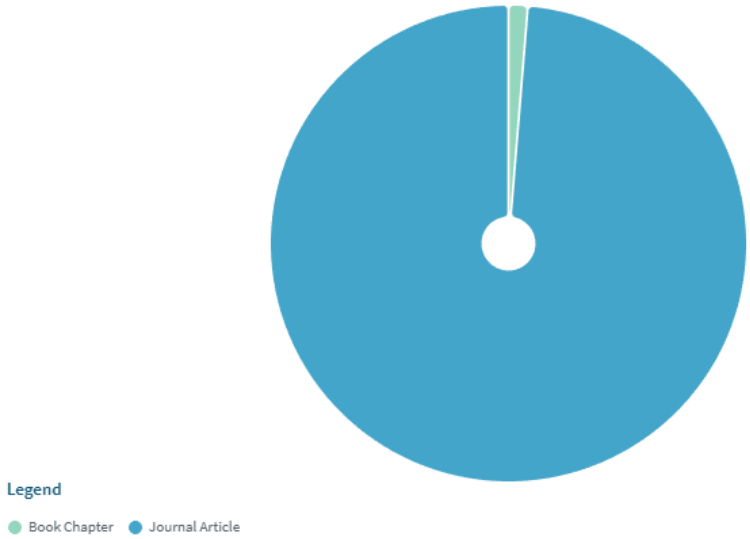

Fig 4. Distribution of publication types in the sample analyzed
Source: The Lens (2025).

4.3. Most active sources, and authors

Scientific production is concentrated in a small set of outlets and actors, representing a large share of publications, according to Bradford's Law (Figure 5).

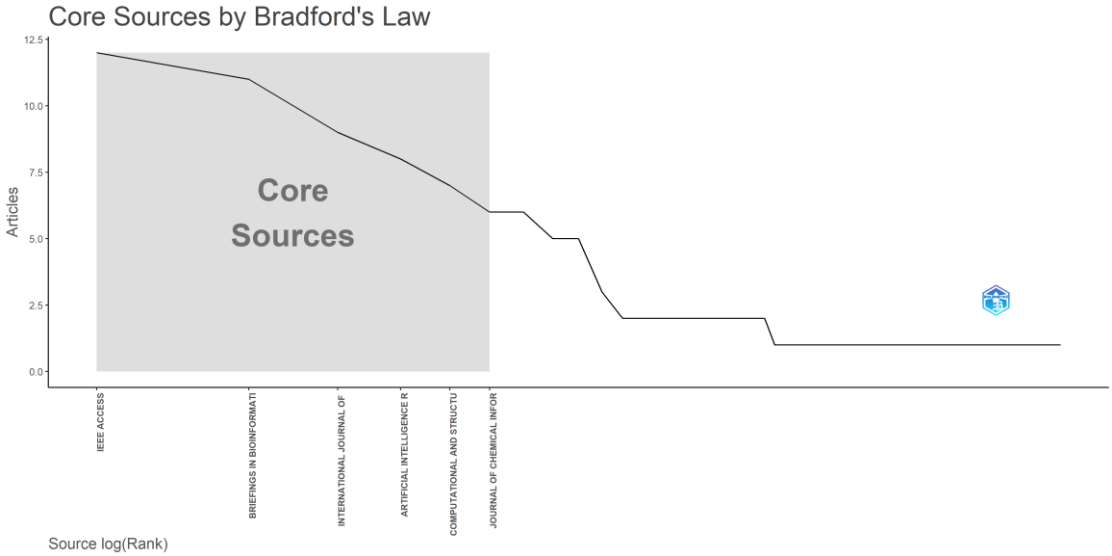

Fig 5. Core sources according to Bradford's Law

Source: output from Bibliometrix (2025).

IEEE Access emerges as the most influential journal, recording the highest metrics with an h-index of 8, a g-index of 12, and a total citation count of 1,333 (Table 1). These figures reflect a consolidated scientific impact maintained since 2019. Briefings in Bioinformatics follow as the second most relevant source; despite its more recent publication history, it demonstrates high impact (h-index = 7) and a m-index of 1.167, signaling rapid growth within the field.

Journal of Molecular Sciences, shows high m-indices (1.5), indicating significant initial impact notwithstanding their currently moderate total citation counts. Conversely, Future Internet shows a substantial volume of citations but a more gradual growth in impact (m-index = 0.75). The remaining sources in the analyzed corpus report comparatively lower levels of impact and productivity.

Table 1. Impact of sources by bibliometric index (Bibliometrix)

Source	h_index	g_index	m_index	TC	NP	PY_start
IEEE Access	8	12	1.143	1333	12	2019
Briefings in Bioinformatics	7	11	1.167	349	11	2020
Expert Systems	4	5	1.0	42	5	2022
Journal of Chemical Information and Modeling	4	6	1.0	215	6	2022
Journal of Cheminformatics	4	6	1.333	161	6	2023
Artificial Intelligence Review	3	8	1.5	109	8	2024
Future Internet	3	5	0.75	467	5	2022
International Journal of Molecular Sciences	3	9	1.5	150	9	2024
Archives of Computational Methods in Engineering: State of the Art Reviews	2	2	0.4	27	2	2021
Artnodes	2	2	0.333	14	2	2020

Source: Authors' elaboration based on data from Impact of sources by bibliometric (Bibliometrix) (2025)

An analysis of the most relevant authors reveals a high concentration of research among a small group of researchers, characterized by significant co-authorship (Table 2). Zhang Y is the most prolific contributor with 8 articles and the highest fractional authorship (1.54), signifying a substantial individual. Wang X follows with 5 publications (0.93 fractional), while

Li C, Li Y, Wang Z, and Zhang J demonstrate moderate productivity (4 articles each) with fractional values ranging from 0.51 to 0.82.

Table 2. Most relevant authors by number of documents (Bibliometrix)

Author	Articles	Articles Fractionalized
Zhang Y	8	1.54
Wang X	5	0.93
Li C	4	0.51
Li Y	4	0.53
Wang Z	4	0.59
Zhang J	4	0.82
Chen L	3	0.41
Li X	3	0.44
Ali S	2	0.29
Bao T	2	0.40

Source: Authors' elaboration based on data from most relevant authors by number of documents by bibliometric (Bibliometrix) (2025)

4.4. Collaboration networks

The analysis of the author collaboration network (Figure 6) reveals a concentration of scientific influence among a few strongly interconnected key actors, alongside several isolated peripheral groups. Within the central core (red cluster), a high density of cooperation is observed, where Zhang Y serves as the primary central node (hub), directly linking to prolific researchers such as Wang X, Li Y, and Li C.

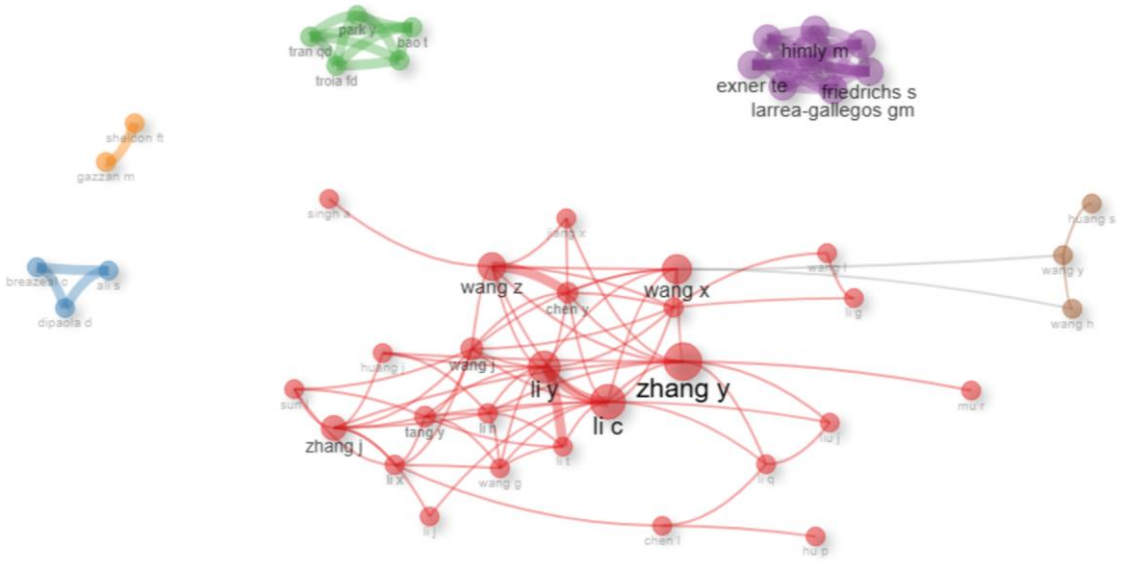

Fig 6. Collaboration network between authors
Source: output from Bibliometrix (2025).

4.5. Thematic Clusters and Architectures (QI2)

The keyword co-occurrence network (Figure 7) reveals a cohesive yet diverse research landscape, organized into two primary interconnected clusters. This structure indicates a mature field where methodological frameworks and technical implementations are deeply integrated.

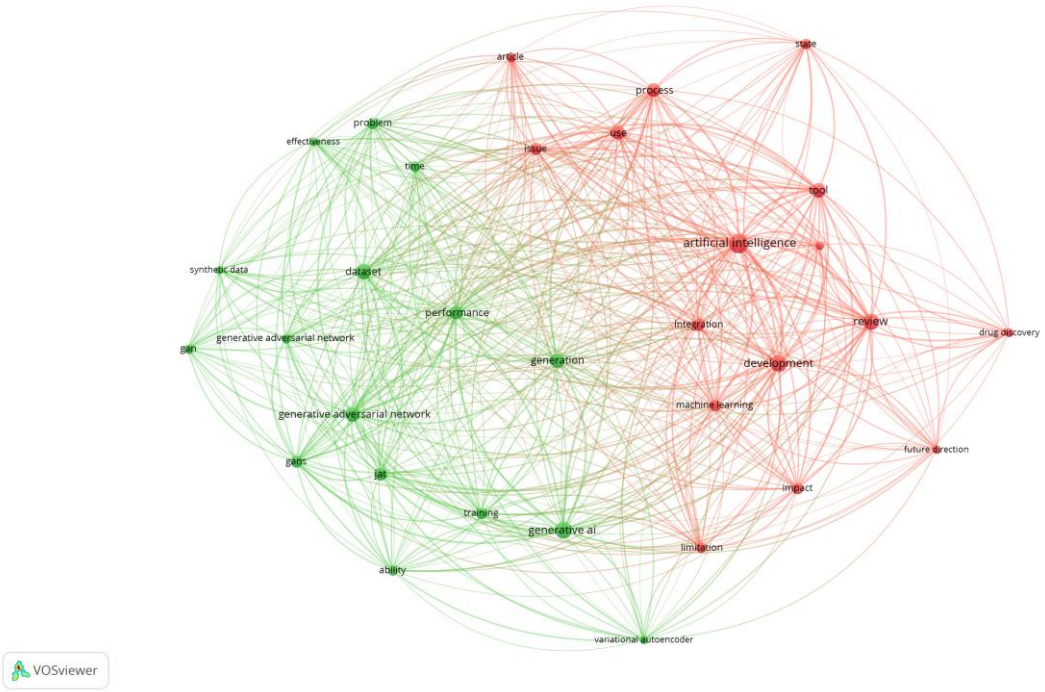

Fig 7. Thematic clustering of literature
Source: output from VOSviewer (2025).

Cluster red - Methodological and prospective: This grouping centers on foundational and transversal themes, such as “artificial intelligence”, “machine learning”, and “systematic reviews”.

Cluster green - Technical and generative: Focused on operational research, this cluster aggregates keywords related to generative architectures, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and synthetic data. The prevalence of terms like “training”, “datasets”, and “performance” reflects an experimental emphasis on model optimization and data synthesis.

The high link density between these clusters, bridged by pivotal nodes such as generation and integration, demonstrates that advancements in generative AI are intrinsically linked to broader machine learning developments. Overall, the network depicts a rapidly expanding domain where generative models serve as a central axis for recent scientific progress.

The thematic map in Figure 8 categorizes the conceptual structure of the studied domain into four quadrants, correlating relevance (centrality) with the degree of development (density) of the various topics.

- Motor Themes: These emerge as the current pillars of development, with significant prominence given to “Generative AI”, “Large Language Models (LLMs)”, and “Reinforcement Learning”. These themes exhibit high centrality and density, representing the leading edge of current research.
- Basic Themes: These represent transversal foundations, such as “Machine Learning”, “Artificial Intelligence”, and “Deep Learning”. Although fundamental to sustaining the field, they exhibit lower density, as they are generalist concepts that have become broadly integrated across literature.
- Niche Themes: Highly specialized areas, such as “diffusion models” and “drug design”, demonstrate isolated technical development. While technically sophisticated, these specific applications have not yet achieved global centrality within the broader research network.
- Emerging or Declining Themes: Topics such as “creativity” and “generative algorithms” suggest new research frontiers. Although these are currently in their nascent stages, they possess the potential to migrate toward quadrants of higher relevance as the field evolves.

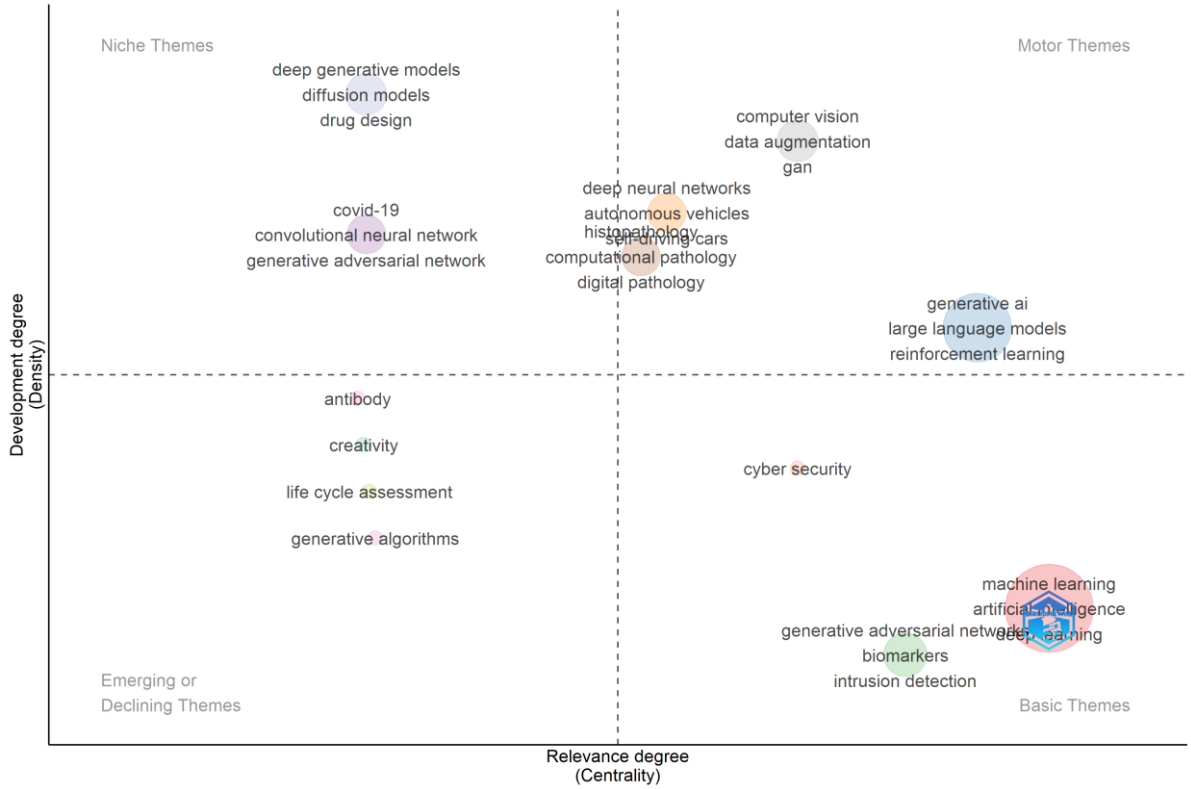

Fig 8. Thematic map – centrality versus density of themes
Source: output from Bibliometrix (2025).

In summary, the analysis reveals a research field in a state of advanced maturity and accelerated specialization. This is characterized by a definitive transition from general predictive models toward complex generative architectures, following a well-defined evolutionary trajectory.

5. Discussion

Bibliometric evidence suggests that GenAI in cybersecurity has moved from exploratory work (2018–2021) to rapid expansion after 2022, with a pronounced concentration of attention in 2024–2025 around LLM-centered automation. This acceleration is accompanied by an asymmetry between attack and defense: offensive capabilities (e.g., mutability, obfuscation, and social engineering scale) mature faster than defensive evaluation frameworks, particularly for adversarial robustness and functional validation of synthetic malware. Collaboration networks indicate an increase in international co-authorship, but also a concentration of production in a limited set of countries and institutions, in line with unequal access to

computing, data, and security test beds. In practice, the field would benefit from shared references and audit-ready evaluation protocols that integrate threat modeling, red teaming, and governance requirements (e.g., traceability, controllability, and accountability) for GenAI systems deployed in high-risk cyber contexts.

6. Conclusions

This review provides a PRISMA-guided bibliometric synthesis of research on General Artificial Intelligence (GenAI) in cybersecurity, concluding that the research field is growing rapidly, with the number of publications increasing sharply after 2022 and focusing on a limited set of stakeholders.

Based on the formulated research questions, the analysis of the results allows a set of conclusions to be drawn. Regarding RQ1, scientific production in the field of GenAI applied to cybersecurity shows a marked increase from 2022 onward, concentrated in a limited number of publication venues (notably IEEE Access) and among a small group of highly productive authors (with prominence of Zhang Y.).

Regarding RQ2, the literature reveals the predominance of well-defined thematic clusters and specific GenAI architectures. Dominant themes cluster around (i) defensive applications such as intrusion detection systems (IDS), cyber threat intelligence (CTI), and synthetic data generation; (ii) mutability, offensive evasion, and adversarial attacks; and (iii) IoT and edge security under resource-constrained environments.

Concerning RQ3, the distribution of the literature between defensive applications and research oriented towards attack or evasion indicates a predominance of defensive approaches in terms of the volume of empirical studies. Nevertheless, offensive research remains persistent and dynamic, particularly in contexts where large language models (LLMs) enable rapid and automated code transformations, thereby facilitating novel attack and evasion techniques.

Finally, the analysis associated with RQ4 highlights a set of significant limitations and research gaps that open avenues for future work. Key gaps include auditability and governance mechanisms for GenAI systems, the adversarial robustness of defensive models, and the lack of standardized protocols to validate the functional fidelity of synthetic malware.

Despite its contributions, this study presents certain limitations. Primarily, the data collection was restricted to a single database (The Lens), and the dataset for 2025 remains incomplete due to the timing of the analysis. Additionally, as the bibliometric mapping is strictly quantitative, further qualitative research is required to provide a more nuanced understanding of this field of study. It is also suggested that future research develop: i) Standardized references for mutable malware and GenAI-assisted attacks; ii) Adversely robust IDS pipelines, validated under realistic attacker adaptation; iii) Governance and audit mechanisms that enable traceable, controllable, and accountable deployment of GenAI in security operations; iv) Plans for hybrid defensive architectures. These architectures should integrate generative AI alongside formal verification and symbolic validation techniques, which would help reinforce systems to combat complex and constantly changing attacks. These architectural approaches really need to integrate generative AI in a continuous manner.

The study demonstrates that the results of this bibliometric study show that generative AI is indeed transforming cybersecurity, even if it is not yet fully complete. Some important audit frameworks are indeed necessary, in addition to risk assessment measures that are attentive to model changes and governance structures that integrate clarity and explainability with accountability. This analysis demonstrates that any future research focused on the technical effectiveness and reliability, resilience, and overall sustainability of AI-driven cybersecurity infrastructure is indeed vital and important.

References

Acosta-Bermejo, R., Terrazas-Chavez, J. A., & Aguirre-Anaya, E. (2025). Automated Malware Source Code Generation via Uncensored LLMs and Adversarial Evasion of Censored Model. *Applied Sciences (Switzerland)*, *15*(17). https://doi.org/10.3390/app15179252

Balasubramanian, P., Liyana, S., Sankaran, H., Sivaramakrishnan, S., Pusuluri, S., Pirttikangas, S., & Peltonen, E. (2025). Generative AI for cyber threat intelligence: applications, challenges, and analysis of real-world case studies. *Artificial Intelligence Review*, *58*(11). https://doi.org/10.1007/s10462-025-11338-z

Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access*, *11*, 76071–76094. https://doi.org/10.1109/ACCESS.2023.3296707

Gaber, M. G., Ahmed, M., & Janicke, H. (2024). Malware Detection with Artificial Intelligence: A Systematic Literature Review. *ACM Computing Surveys*, *56*(6). https://doi.org/10.1145/3638552

Gazzan, M., Alobaywi, B., Almutairi, M., & Sheldon, F. T. (2025). A Deep Learning Framework for Enhanced Detection of Polymorphic Ransomware. In *Future Internet* (Vol. 17, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/fi17070311

Guo, Y. (2023). A Survey of Machine Learning-Based Zero-Day Attack Detection: Challenges and Future Directions. In *Computer Communications* (Vol. 198, pp. 175–185). Elsevier B.V. https://doi.org/10.1016/j.comcom.2022.11.001

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. In *IEEE Access* (Vol. 11, pp. 80218–80245). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2023.3300381

Ibrar, W., Mahmood, D., Al-Shamayleh, A. S., Ahmed, G., Alharthi, S. Z., & Akhunzada, A. (2025). Generative AI: a double-edged sword in the cyber threat landscape. *Artificial Intelligence Review*, *58*(9). https://doi.org/10.1007/s10462-025-11285-9

Javaheri, D., Lalbakhsh, P., & Hosseinzadeh, M. (2021). A Novel Method for Detecting Future Generations of Targeted and Metamorphic Malware Based on Genetic Algorithm. *IEEE Access*, *9*, 69951–69970. https://doi.org/10.1109/ACCESS.2021.3077295

Kumar, D., Pawar, P. P., Addula, S. R., Meesala, M. K., Oni, O., Cheema, Q. N., Haq, A. U., & Sajja, G. S. (2025). AI-Powered Security for IoT Ecosystems: A Hybrid Deep Learning Approach to Anomaly Detection. *Journal of Cybersecurity and Privacy*, *5*(4), 90. https://doi.org/10.3390/jcp5040090

Labaca-Castro, R. (2023). Machine Learning under Malware Attack. In *Machine Learning under Malware Attack*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-40442-0

Liu, Y., Huang, J., Li, Y., Wang, D., & Xiao, B. (2025). Generative AI model privacy: a survey. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11024-6

Liu, Y., Huang, J., Li, Y., Wang, D., & Xiao, B. (2025). Generative AI model privacy: a survey. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11024-6

Mahmoudi, I., Boubiche, D. E., Athmani, S., Toral-Cruz, H., & Chan-Puc, F. I. (2025). Toward Generative AI-Based Intrusion Detection Systems for the Internet of Vehicles (IoV). In *Future Internet* (Vol. 17, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/fi17070310

Pei, G., Ma, K., Eece, ucasaccn, Beijing, U., Dongpeng Zhang, C., Xu, Q., Huang, Q., Zhang, D., & Sun, C. (n.d.). A Unified Framework for Stealthy Adversarial Generation via Latent Optimization and Transferability Enhancement. In *Proceedings of the 33nd ACM*

International Conference on Multimedia (MM '25), October 27â•fiOctober 31, 2025, Dublin, lreland (Vol. 1). ACM.

Radanliev, P. (2025). Collaborative penetration testing suite for emerging generative AI algorithms. *Applied Intelligence*, *55*(16). https://doi.org/10.1007/s10489-025-06908-1

Yazdani, S., Singh, A., Saxena, N., Wang, Z., Palikhe, A., Pan, D., Pal, U., Yang, J., & Zhang, W. (2025). Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01247-x