SECURITY & DEFENCE QUARTERLY
https://securityanddefence.pl/

# The dual-use dilemma of generative artificial intelligence in cybersecurity: Navigating the explosive growth in offensive and defensive applications

**Karol Chlasta**

kchlasta@kozminski.edu.pl

https://orcid.org/0000-0002-6539-566X

Department of Management in Networked and Digital Societies, Kozminski University, 57/59 Jagiellońska Street, 03-301, Warsaw, Poland; and Department of Artificial Intelligence, WarsawIQ, 43A/37B Jana Pawła II Avenue, 01-001, Warsaw, Poland

## Abstract

*This article systematically reviews the academic literature on the applications of artificial intelligence (AI) in cybersecurity, with a specific focus on generative techniques (generative AI or GenAI), such as large language models (LLMs). The analysis covers publications from 1993 to 2025 extracted from the IEEE Xplore Digital Library (1,647 publications) and SpringerLink (1,742 publications) databases, resulting in a collection of 3,389 documents. The litstudy tool was utilised for thematic mapping, automatic topic modelling, and n-gram analysis. The analysis shows an exponentially upward interest since 2022, particularly between 2023 and 2024, indicating rapidly growing interest in GenAI methods. Through topic modelling, the study identified the following key thematic areas: LLMs (734 publications; 21.88%), identified as the most dominant topic, blockchain and cyber attacks (394 publications; 11.74% each topic), generative coding for software (296 publications; 8.82%), smart energy and internet of things (278 publications; 8.29%), and malware (70 publications; 2.09%). The study revealed that GenAI and LLMs present a significant dual-use dilemma. Malicious actors increasingly leverage these methods for offensive purposes. Conversely, the same methods are actively developed to enhance cybersecurity defences. GenAI and LLMs are fundamentally reshaping the cybersecurity landscape, as evidenced by the visible growth in research interest. In light of the dual-use dilemma posed by GenAI, organisations should urgently consider investing in securing their sensitive data, and enhance staff capabilities in threat detection and response using GenAI-/LLM-driven methods.*

**Keywords:**

generative AI, large language models (LLMs), AI-driven threats, offensive security, defensive security

**Article info**
Received: 01 June 2025
Revised: 29 December 2025
Accepted: 25 January 2026
Available online: 12 March 2026

Citation: Chlasta, K. (2026) 'The dual-use dilemma of generative artificial intelligence in cybersecurity: Navigating the explosive growth in offensive and defensive applications', *Security and Defence Quarterly*, 53(1). doi: 10.35467/sdq/217364

# Introduction

Security in cyberspace is one of the key challenges of today's world, particularly in the context of the dynamic development of information and communication technologies and the increasing scale and complexity of digital threats. In this era of digital transformation, the importance of modern and integrated mechanisms to protect data and information and communication technology (ICT) infrastructure is growing—especially when considering regulatory requirements and international security standards, such as information security management systems International Organization for Standardization(ISO)/International Electrotechnical Commission (IEC) 27001 and the National Institute of Standards and Technology (NIST, 2018) Cybersecurity Framework.

A significant factor shaping today's cybersecurity ecosystem is the development of artificial intelligence (AI), which, on the one hand, serves as an effective tool for supporting the protection of information systems and, on the other, can be exploited as an attack vector by malicious actors. Generative models, including Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014), exemplify this dual-use nature. GANs operate based on a "game" between two neural networks: a generator and a discriminator. The generator creates data that mimic the training set, while the discriminator learns to distinguish between synthetic and real data. Transformer-based architecture introduced in 2017 (Vaswani *et al.*, 2017), which leverage a self-attention mechanism, enable the parallel processing of data sequences. This makes them highly effective in tasks, such as natural language processing (NLP) tasks, anomaly detection, and threat classification. These two architectures not only enable the generation of realistic synthetic data (e.g. images, text, or audio) but also facilitate complex attacks, such as deepfakes, spear-phishing using NLP techniques, manipulation of training data in machine learning (ML) pipelines (Brundage, 2018), and the creation of disinformation content (Creswell *et al.*, 2018).

Progress in the field of generative AI (GenAI) is exemplified by the rapid evolution of architectures since the introduction of the Transformer in 2017 (Vaswani *et al.*, 2017). This progress includes models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019), generative pre-trained transformer 2 (GPT-2) (Solaiman *et al.*, 2019), and more recent large language models (LLMs) such as Claude[1], Gemini (Team Google, 2024), and GPT-4 family (OpenAI, 2023). These models are characterised by an increasing number of parameters, enabling them to learn more complex patterns and generate increasingly realistic outputs. Moreover, they are now accessible not only via the cloud but also for edge computing applications (Zheng *et al.*, 2025).

Although the literature already contains multiple reviews and surveys on AI in cybersecurity, most of the existing works remain fragmented in scope and limited in methodological depth. Recent reviews tend to focus either on traditional machine learning approaches or on narrowly defined applications of GenAI, such as vulnerability detection, cyber operations automation, or adversarial attacks against AI systems (Afolabi and Akinola, 2024; Barrett *et al.*, 2023; Fu *et al.*, 2023). Even broader surveys of GenAI in cybersecurity (Hasanov *et al.*, 2024; Yigit *et al.*, 2024) primarily adopt qualitative or manually curated taxonomies, offering descriptive overviews, rather than a quantitative assessment of how research priorities and thematic structures have evolved. As a result, the existing literature does not provide a consolidated data-driven view of the simultaneous expansion of GenAI across both offensive and defensive cybersecurity domains, particularly in the post-2022 period marked by the rapid proliferation of LLMs.

[1] *Anthropic's all models' overview: https://docs.anthropic.com/en/docs/about-claude/models/all-models*

It can be anticipated that GenAI and LLMs will play a pivotal role in cybersecurity, in both defensive operations—for instance, by automating threat detection—and offensive capabilities, such as the generation of malware code.

## Materials and Methods

This article is a review of the academic literature on AI applications in the area of cyber security, with a particular focus on new generative techniques, such as LLMs. To support the systematic analysis of the collected literature, the *litstudy* tool was used as an automated framework for large-scale bibliometric and thematic analysis. *Litstudy* is a Python-based tool that processes bibliographic metadata and textual content (titles, abstracts, and keywords) to construct a corpus, extract n-grams, and perform frequency analysis and unsupervised topic modelling methods. The tool also applies dimensionality reduction techniques to visualise semantic relationships between documents in a two-dimensional (2D) topic landscape. This approach enables reproducible data-driven exploration of thematic trends across large datasets while reducing the subjectivity of manual literature reviews (Heldens *et al.*, 2022). The tool was used to develop a thematic map of publications as well as automatic topic modelling, analysis of n-grams in article abstracts, and visualisation of results.

The aim of this literature analysis is to assess the extent to which GenAI—particularly LLMs, following the introduction of the Transformer architecture in 2017—is applicable within the field of cybersecurity. The analysis covers publications from the period 1993 to 2025 and is based on sources retrieved from two well-established academic databases: the IEEE Xplore Digital Library and SpringerLink databases. The search was conducted using the keywords "cybersecurity" and "AI." A total of 1,647 publications were retrieved from the IEEE Xplore database and exported in comma separated values (CSV) format, while an additional 1,742 documents were obtained from SpringerLink database. After merging and deduplicating the datasets, a final corpus of 3,389 articles was assembled.

These databases were selected based on the author's prior experience and their complementary coverage. Publications indexed in IEEE Xplore predominantly emphasise technical- and engineering-oriented aspects of cybersecurity, such as threat detection and data classification. In contrast, SpringerLink database provides broader interdisciplinary perspectives, frequently addressing socio-ethical considerations, including privacy, regulatory compliance, and societal perceptions of AI technologies, often within the edited volumes and book series.

As part of the data preparation for this analysis, the three CSV files were exported from each database and loaded to *litstudy*, which facilitated the generation of descriptive statistics (including n-gram frequency histograms) and thematic exploration of the documents shown in Figures 2–5.

To build the corpus, the function *build_corpus* was used with the parameter *ngram_threshold=0.8*, which allowed popular phrases to be included as consistent lexical units. Then, using the *compute_word_distribution* function, a frequency analysis of n-grams in the corpus was carried out. This allowed the identification of key concepts present in the literature, such as "artificial_intelligence," and "machine_learning."

The results of the automatic topic modelling are presented in the form of a visualisation of a topic landscape, based on a word cloud, using Nonnegative Matrix Factorisation (NMF), a factorisation algorithm commonly used for dimensionality reduction in unsupervised machine learning (Xu *et al.*, 2003) and a landscape plot, which uses non-linear dimensionality reduction *algorithm*s, such as **t-**distributed stochastic neighbour embedding (t-SNE) (Van der Maaten and Hinton, 2008) and unform manifold approximation and projection (UMAP) (McInnes *et al.*, 2018), to arrange documents on a 2D plane. Such maps allow intuitive identification of thematic groups in which closely located documents represent similar semantic content. However, it should be emphasised here that for the landscape plot the distances between points in the visualisation are illustrative and should not be interpreted in a metric sense.

Figure 1 presents a concise overview of the methodological workflow, illustrating how the dataset was systematically processed and transformed throughout the analysis. The diagram summarises the sequential stages of the study, beginning with data acquisition and deduplication, followed by corpus construction, n-gram extraction, and frequency analysis, and continuing through topic modelling and semantic mapping. By visualising these steps in a unified pipeline, the figure clarifies how raw bibliographic records were converted into interpretable thematic structures.

This framing is particularly important, given the rapid post-2022 expansion of AI-related publications. Rather than treating this growth as a standalone observation, the methodological design enables a deeper examination of how GenAI has reshaped research priorities, redistributed thematic clusters, and influenced the balance between offensive and defensive cybersecurity applications.

**Figure 1. Methodology pipeline for automated bibliometric and thematic analysis.**

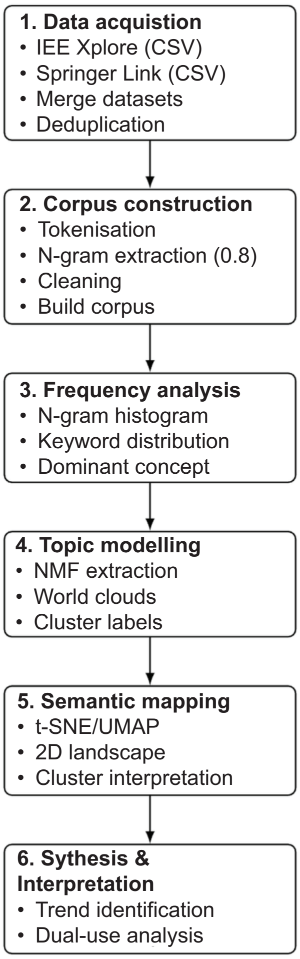
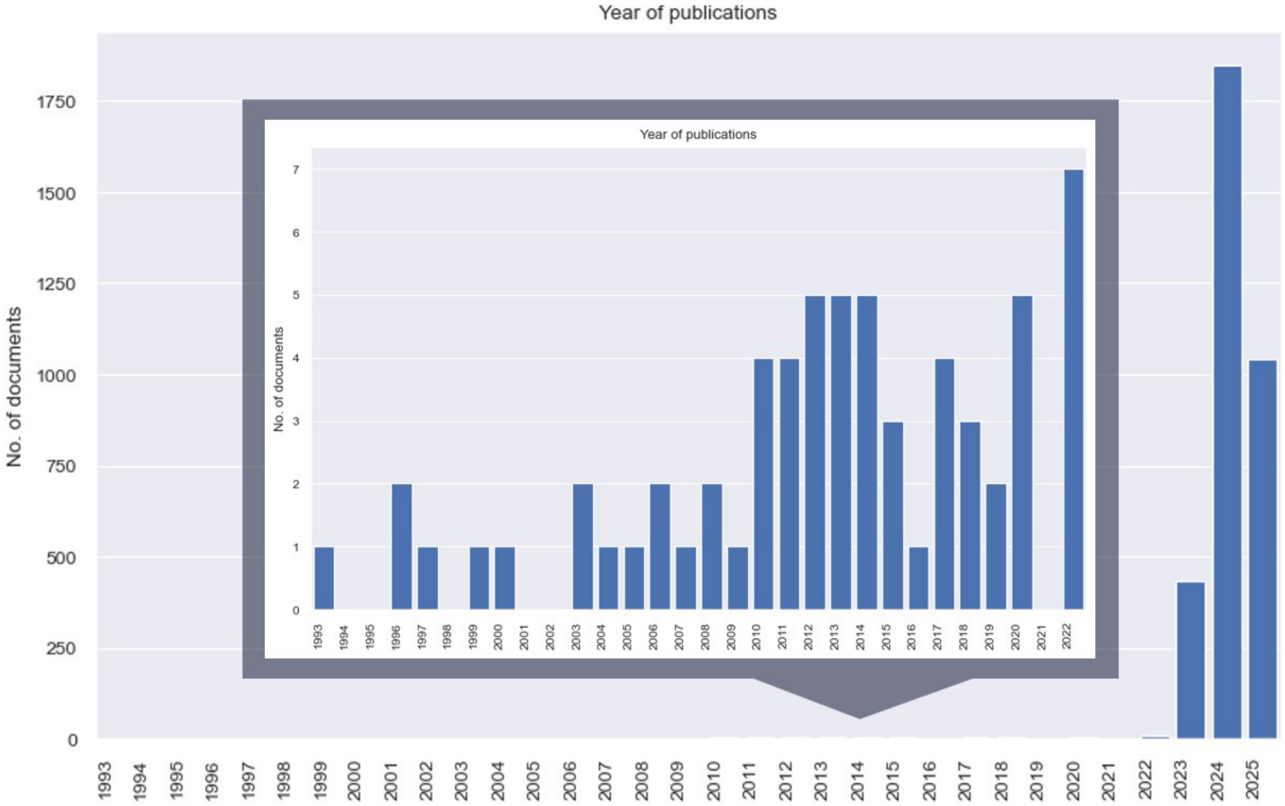


**Figure 2. Publication counts for AI and cybersecurity research (1993–2025) included to provide a longitudinal baseline for the bibliometric and thematic analyses.**



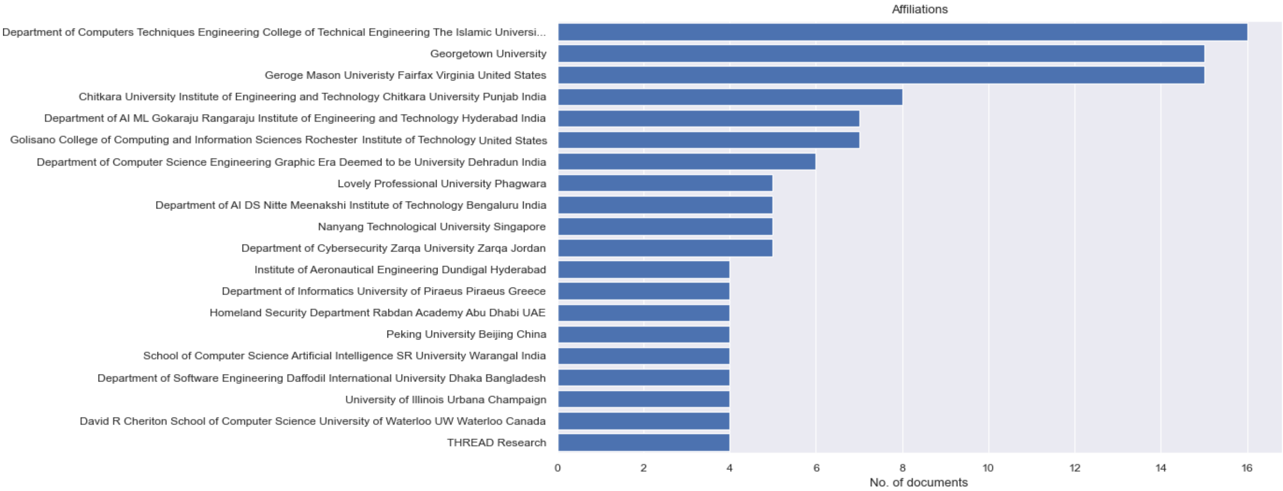
# Results

The collected results present a comprehensive picture of the current state of research at the intersection of AI and cyber security, highlighting the growing role of GenAI techniques—in terms of threat detection, vulnerability analysis, simulation of attack scenarios, and automation of incident response.

Analysis of the number of publications between 1993 and 2025 (Figure 2) provides a historical baseline for the dataset, illustrating how publication volume evolved over time, enabling the post-2022 surge to be contextualised within a longer trajectory. Starting with one publication in 1993, the number increased slowly until 2022 (seven publications), after which the rate of growth increased significantly. A particularly significant jump was observed between 2022 and 2023, where the number of publications reached 435 per year, only to expand to 1,848 in 2024. Such a growth might indicate not only a increasing interest in the topic of AI in the context of cyber security but also an increased awareness among researchers of the potential defensive and offensive applications of these technologies.

Figure 3 presents the institutional affiliations for publications related to AI and cybersecurity and reveals a diverse set of contributors across academic and research institutions worldwide. Surprisingly, the most frequently occurring affiliation is the Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq, contributing sixteen documents to the dataset. This is followed by Georgetown University and George Mason University (Fairfax, VA, USA), each contributing fourteen documents.

**Figure 3. The twenty most frequent affiliations in the collected dataset.**



Other significant contributors include Chitkara University Institute of Engineering and Technology (Punjab, India) with eight documents, and several Indian institutions, such as Gokaraju Rangaraju Institute of Engineering and Technology (Hyderabad, India) and Graphic Era Deemed to be University (Dehradun, India), each contributing between six to seven documents. International representation is evident, with institutions from the United States, India, Singapore (Nanyang Technological University), Jordan (Zarqa University), Greece (University of Piraeus), United Arab Emirates (Rabdan Academy), China (Peking University), Bangladesh, and Canada (University of Waterloo).

This diverse academic participation highlights a strong global interest in the intersection of AI and cybersecurity, with a noticeable concentration of contributions from institutions in Asia and North America.

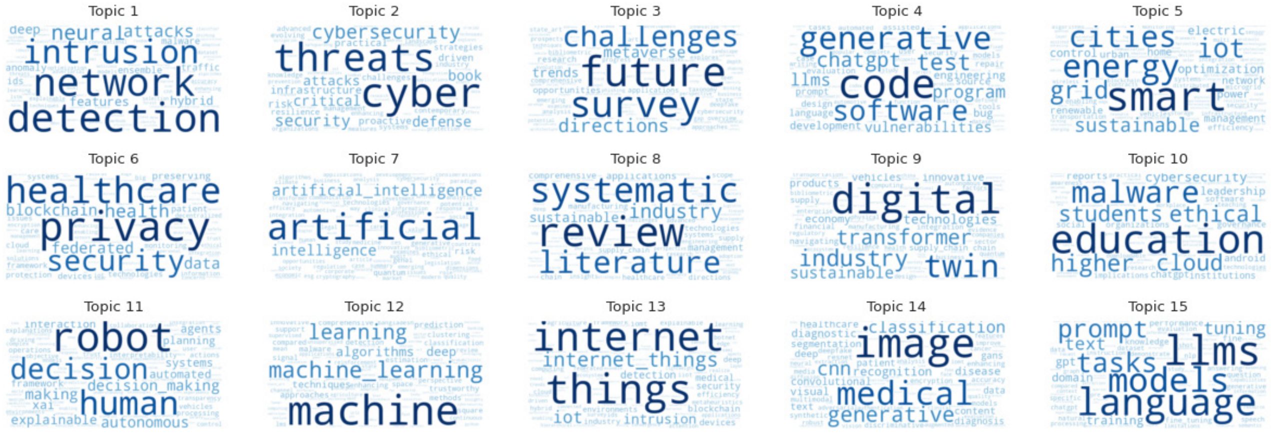
Through topic modelling analysis, six clusters (Topics 15, 4, 10, 1, 5, and 9) emerge as foundational to understanding how GenAI is reshaping the cybersecurity landscape. These topics emphasise the importance of standards, systematic reviews, deep learning-based network protection, industrial integration, AI foundations, and malware in the changing threat landscape.

Figure 4 presents a visualisation of the key topics within the dataset generated using the NMF method.

- Topic 15 (LLMs, language, and models): This topic shows the highest prevalence. This confirms that LLMs and associated generative technologies are a central theme in the corpus. This aligns perfectly with the main aspect of this article, pertinent to understanding the capabilities, applications, and implications of LLMs within the wider "AI" space.

- Topic 9 (digital, twin, and industry) and Topic 1 (detection, network, and intrusion): Both these topics show a notable presence. The "detection, network, intrusion" theme provides a strong grounding in core cybersecurity concerns, forming the backdrop against which the role of GenAI in both emerging threats and defensive strategies can be analysed. The "digital" theme relates to secure data handling, industrial systems, or even novel applications in conjunction with AI, relevant to robust defensive strategies or new standards.

**Figure 4. Word cloud presenting automatic topic modelling results in the dataset.**



- Topic 4 (code, generative, and software): This relates to new "generative" aspects of coding. It points to discussions on AI-driven software development, automated code generation, and potentially the exploitation of such tools for malicious code creation, a key area for analysing emerging threats and defensive strategies (e.g. AI for vulnerability detection in generated code).

- Topic 5 (smart, energy, and cities): This content related to "smart" technologies, including internet of things (IoT). This highlights critical application domains where GenAI could introduce new vulnerabilities or, conversely, enhance security. The security of IoT systems is a well-established concern, and the intersection with GenAI is an important area for exploring threats and defences.

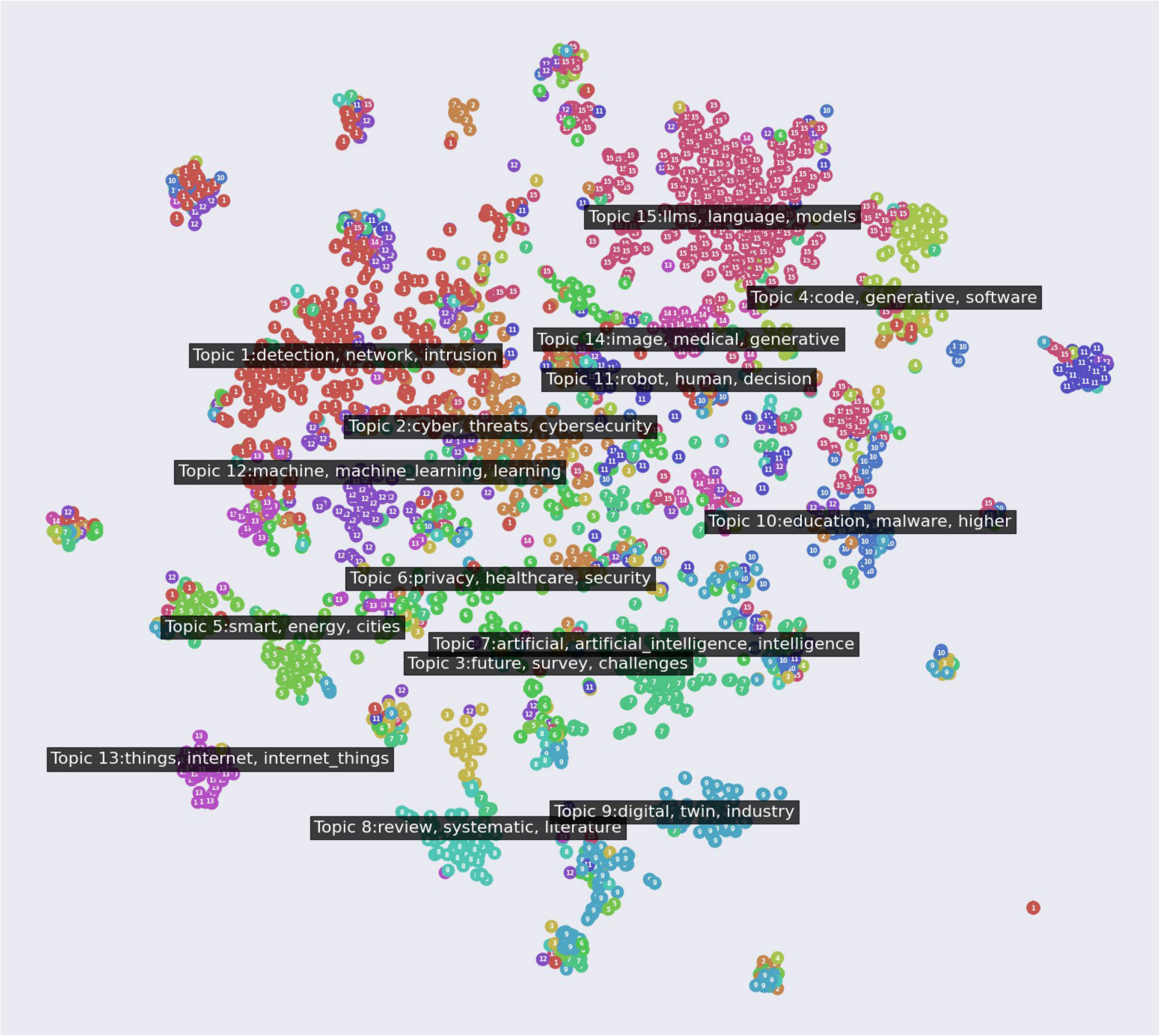
- Topic 10 (education, malware, and higher): While this topic's presence indicates discussions on specific, well-known cyber threats, GenAI could potentially be used to create more sophisticated and evasive malware (an emerging threat), or to improve malware awareness or detection techniques (a defensive strategy).

Figure 5 illustrates the distribution of topic clusters in the 2D space reduced using t-SNE and UMAP techniques. This landscape map provides insights into how discussions around GenAI are situated within the broader cybersecurity and AI research landscape. The map visually confirms that it is an increasingly integrated part of the cybersecurity discourse. The clear thematic links between generative methods, attack vectors, defensive machine learning, and fundamental security principles provide a strong argument for considering generative methods as both emerging threats and new defensive strategies. This is evidenced through the interconnectedness of these themes, with closely located documents representing similar semantic content.

The key observations are as follows:

- *Prominence and connectivity of GenAI themes*: Topic 15 (LLMs, language, and models) and Topic 4 (code, software, and generative) are clearly visible as distinct clusters. Their position on the map shows that they are not isolated but are connected to the core cybersecurity detection, network, and intrusion areas. For instance, Topic 15 is situated near Topic 14 (image, medical, and network, often involving artificial neural networks) and is part of a larger super-cluster that includes Topic 4. This indicates a strong established link between generative techniques and broader AI applications.

**Figure 5. Landscape of the topics included in the dataset.**



- *Intersection with core cybersecurity and AI/ML*: These GenAI clusters (Topics 15 and 4) are adjacent to, and show pathways towards, foundational cybersecurity topics such as Topic 1 (detection, network, and attacks in red colour), Topic 2 (cyber, threats, and cybersecurity in orange colour), and general AI/ML topics, such as Topic 12 (machine, machine_learning, and learning in purple colour), that are concentrated in proximity. It visually supports the premise of emerging threats, as generative capabilities are discussed in the context of the existing attack vectors and threat landscapes.

- *Centrality of security fundamentals*: Topics such as Topic 2 (cyber, threats, and cybersecurity in orange colour) and Topic 7 (artificial, artificial intelligence, and intelligence in green colour) are located within the main dense area of the map, suggesting their cross-cutting importance across various cybersecurity and AI discussions. The need for explainable AI (XAI) is particularly relevant for standards and building trust in AI-driven security solutions, as it refers to the techniques that make model decisions transparent and understandable. In cybersecurity, this is essential because security-critical alerts and automated responses must be interpretable, auditable, and aligned with regulatory standards. XAI helps analysts trust and verify AI-driven security systems.

- *Specialised application/threat areas*: More specific topics such as Topic 10 (education, malware, and higher in blue colour) and Topic 5/13 (related to IoT, smart, and energy in green and purple colours) form their own sub-clusters. Their relationship to the AI cluster can indicate specific areas where GenAI might introduce novel threats (e.g. sophisticated malware variants) or offer new defensive tools (e.g. AI-driven IoT security analytics). Topic 9 (digital, twin, and industry in light blue colour) also represents a new distinct technological area with its own security considerations (e.g. Industrial Digital Twin Security, a specific area within industrial cybersecurity focused on securing digital representations of physical industrial processes).

The data presented in Table 1 highlights the key trends in cybersecurity as of May 2025, with LLMs emerging as the most dominant topic, comprising 21.88% of all surveyed publications. This reflects the growing integration of these GenAI technologies into cybersecurity domains. Topics related to blockchain and cyberattacks are also prominent, each representing 11.74% of the research, indicating sustained interest in both emerging technologies and persistent threats. Generative software and coding, as well as smart energy and IoT infrastructure, follow with 8.82% and 8.29%, respectively, showing a balanced focus on innovation and infrastructure security. In contrast, malware—although still relevant—commands a smaller share of attention at just 2.09%, possibly suggesting a shift in research priorities towards either broader or more focused technological cluster.

## Exploring Dual-Use Dilemma of Generative AI and LLMs

In recent years, GenAI has begun to play an increasingly important role in both defensive and offensive operations in cyberspace (Yigit *et al.*, 2024). New platforms hosting LLMs, such as GPT-4, are widely impacting threat analysis, penetration test automation, and vulnerability detection in source code (Fu *et al.*, 2023). Authors observed (Hasanov *et al.*, 2024) that LLMs have been proved highly effective in phishing attack simulations and in managing cybersecurity administrative aspects, including defending against advanced exploits.

At the same time, the same mechanisms are used by malicious actors to generate real phishing attacks, automatically break through security or create complex malware (Gupta *et al.,* 2023). This impact is therefore multifaceted, presenting both opportunities for enhancing defences automation (Sultana *et al.*, 2023) and risks through their potential misuse by malicious actors. The landscape is rapidly evolving, necessitating ongoing research and evaluation. This forces organisations to rethink their risk models and security strategies, considering not only technical effectiveness but also resistance to human manipulation and new process-related challenges in the data platforms they use (Khoje, 2024).

**Table 1. Results of automated topic analysis (top 6) in cyber security and AI.**

| Topic | Number of publications | Percentage share |
|---|---|---|
| LLMs | 734 | 21.88% |
| Cyberattacks and blockchain | 394 | 11.74% (each) |
| Generative (software and code) | 296 | 8.82% |
| Smart (energy and IoT) | 278 | 8.29% |
| Malware | 70 | 2.09% |

There is an emerging need for upskilling dynamic regulatory frameworks and "cyber resilience" solutions that can adapt to the changing nature of threats.

## Offensive Applications of Generative AI and LLMs

Generative AI and LLMs present a significant dual-use dilemma in the realm of cybersecurity, possessing the potential to enable malicious activities. Cybercriminals are increasingly leveraging these advanced capabilities to develop more sophisticated, evasive, and automated attacks, contributing to an "AI arms race" (Barrett *et al.*, 2023). The integration of AI into cyberattacks enhances their effectiveness and can potentially disrupt legitimate AI defence mechanisms (Samonte *et al.*, 2024).

The dataset collected highlights several key ways in which malicious actors utilise GenAI and LLMs for offensive purposes:

- *Facilitating fraud and generating malicious content or code*: LLMs are misused for fraudulent activities, impersonation, and the generation of malicious software. These revelations underscore the security-related challenges posed by such models (Imtiaz *et al.*, 2023). Tools such as WormGPT, described as a BlackHat GPT for cybercriminals, are capable of launching various types of attacks (Firdhous *et al.*, 2023) or creation of metamorphic malware (Madani, 2023), which dynamically modifies its structure to avoid signature and behavioural detection. Researchers (Gupta *et al.*, 2023) demonstrate that GenAI can be used to create automated ransomware tools that not only locate and encrypt data but also modify their code in real time to evade detection. Of particular concern is the embedding of a Python interpreter in the malicious code, enabling dynamic communication with LLM (e.g. ChatGPT) to generate new pieces of malware ("on-the-fly malware generation").

- *Enabling sophisticated social engineering and phishing attacks*: GenAI plays an amplifying role in social engineering attacks. Attackers can leverage LLMs for strategic advantage, including the orchestration of social engineering attacks using sophisticated language-based techniques. AI is a tool for social engineering attackers, enhancing the likelihood of successful attacks. The existing GenAI-based chatbot services, such as OpenAI's ChatGPT, Google's BARD, can be exploited to create smishing (SMS phishing) texts and eventually leading to craftier smishing campaigns.[2] Research (Shibli *et al.*, 2024) provides strong empirical evidence of this exploitation, often by crafting prompt injection attacks to circumvent ethical standards in these services.

- *Generating deceptive content and deepfakes*: LLMs can be used by malicious actors for strategic advantage, including the manipulation of public opinion through the generation of deceptive content. The rapid development of GenAI technology has led to the misuse of deepfake technology, resulting in issues such as telecommunication fraud and the manipulation of public opinion. Deepfakes (Mi and Zhang, 2025), often built using GAN models, can be used for malicious purposes, such as creating forged speeches and defaming individuals, particularly politicians.

- *Fueling various cybercrimes and crafting tailored attack*: Beyond social engineering, phishing, and deepfakes, LLMs facilitate illegal activities, such as password cracking, typosquatting, and malware propagation. AI-assisted cyberattacks are becoming

[2] *AbuseGPT: https://github.com/ashfakshibli/AbuseGPT*

increasingly successful across the cyber-defence lifecycle. Tools such as MalGAN can be deployed in the reconnaissance phase to automatically exploit vulnerabilities in cyber-defence systems (Klopper and Eloff, 2024). Specialised frameworks, such as ThreatGPT (Gupta *et al.*, 2023), based on jailbroken versions of models, are developed as well to generate code, attack scenarios, and even simulate socio-technical interactions.

Development of open-source language models (e.g. Chinese 01.ai[3] and availability of repositories such as ChatGPT_DAN[4]) increases the risks related to "Do Anything Now" (DAN) jailbreaks. These factors, combined with the general, public knowledge of the existing hardware vulnerabilities, and well-described ransomware attacks[5] (encrypting data, e.g. WannaCry, Ryuk, and REvil), or erasers (NotPetya), are creating a new class of threats.

A customised LLM named HackerGPT has been specifically tailored for generation of cyberattack and has been demonstrated to attack virtual campus area network (CAN) networks, Bluetooth systems, and Key Fobs, showing successful compromises in certain vehicle models (Usman *et al.*, 2024).

- *Targeting AI systems themselves*: The increasing prevalence of AI introduces new vulnerabilities, making AI-driven systems susceptible to cybercriminal activities (Afolabi and Akinola, 2024). AI models themselves can be the targets of adversarial attacks, including data poisoning, model manipulation, or evasion attacks. Adversarial attacks trick AI systems with malicious data to cause misclassification. Vulnerabilities in LLMs also include prompt injection and denial of service.

In summary, the sources indicate that GenAI and LLMs are actively being weaponised by cybercriminals to enhance the scale and sophistication of attacks, generate malicious content, create convincing deceptive materials, and even target the AI systems designed to protect against them.

## Defensive Applications of Generative AI and LLMs

The sophisticated capabilities of AI, including its ability to process vast datasets, recognise complex patterns, and learned over time, enable advancements beyond traditional security measures, offering a powerful arsenal to prevent and combat the evolving threats (Ankalaki *et al.*, 2025). The dataset identifies several key areas where GenAI and LLMs are being applied to enhance cybersecurity defences:

- *Enhanced threat intelligence and information processing*: LLMs are proving effective in improving the analysis of vast amounts of security data, which is critical for cybersecurity threat intelligence (CTI). They can be used for Named Entity Recognition (NER) (Qiao *et al.*, 2023) on security-related data to build knowledge graphs, enabling fine-grained analysis of multi-source threat intelligence. LLMs can parse and categorise cyber-related information from sources such as news articles, enhancing real-time vigilance and cyber threat modelling (Patel *et al.*, 2024). This capability assists security teams in understanding prospective dangers and

[3] *Yi Foundation models: https://www.01.ai/#yi-foundation-models*
[4] *ChatGPT DAN: https://github.com/0xk1h0/ChatGPT_DAN*
[5] *Cloudflare "What are Petya and NotPetya?": https://www.cloudflare.com/learning/security/ransomware/petya-notpetya-ransomware/*

developing defence measures, for example, by facilitating log analysis with LogGPT (Qi *et al.*, 2023). LLMs can also automate report generation from Intrusion Detection System (IDS) alerts (Amin *et al.*, 2024), improving the efficiency of threat detection analysis. AI-driven threat intelligence platforms can analyse extensive datasets to detect patterns indicative of cyber threats. Organisations such as CaixaBank pilot AI-powered intrusion detection systems using federated learning to enhance data privacy across financial institutions (AI4FIDS), alongside TRUST4AI.xAI, an explainability module that ensures transparency and accountability in AI decision-making. Experimental results presented by Karampasi *et al.* (2024) from the project demonstrate that the integrated system effectively detects network intrusions while preserving user privacy and improving trust through interpretability.

- *Automated threat detection and analysis*: GenAI and LLMs contribute to automating and improving the accuracy and speed of threat detection (Ghazal *et al.*, 2024). GANs and Variational Autoencoders (VAEs) can be used not only to simulate cyber-attacks for testing defences but also for detecting anomalies and generating adaptive countermeasures (Vadisetty and Polamarasetti, 2024). Deepfake and phishing content detection mechanisms contribute to reducing the effectiveness of social engineering attacks (Enathe VP *et al.*, 2024). GANs, by creating synthetic data instances that mimic real-world threats, can enhance the resilience and flexibility of Network Intrusion Detection Systems (NIDS) in dynamic environments (Sarika *et al.*, 2024). LLMs assist in interpreting IDS rules and predicting attacker tactics, techniques, and procedures (TTPs) within frameworks such as MITRE ATT&CK, aiding in defensive cyber operations and suggesting proactive/reactive strategies, classify sensitive data for adaptive security frameworks, or assist with code analysis (Ferrag *et al.*, 2024). Large-scale language models and AI-powered code clone search models are being applied to firmware vulnerability detection and asset management in military contexts (Beninger *et al.*, 2024).

- *Cyber deception and honeypots*: GenAI and LLMs can be used for cyber deception by simulating adaptive honeypots (Childs and Wood, 2024). These AI-enhanced systems can generate realistic, human-like responses in real-time to engage attackers, making decoys harder to identify and improving intelligence gathering on attacker TTPs (Ragsdale and Boppana, 2023). Structured prompt engineering with GenAI can automate the creation of scalable deception ploys (Ahmed *et al.*, 2025).

- *Automated response*: AI systems can automate responses to detected threats, reducing incident response period. This includes automating security operations and addressing incidents in real time. Automated response actions are enabled through dynamic risk assessment, adaptive access controls, and incident remediation. LLMs can act as conversational agents to help administrators with cybersecurity information, log analysis, event detection, and instructions. Hybrid AI models incorporating LLMs can support autonomous cyber defence agents. Automating tasks such as threat identification and response increases operational efficiency (Ferrag *et al.*, 2024).

- *Vulnerability management*: AI can assist in managing system vulnerabilities. LLMs can be valuable assistants for developers and security professionals, helping with secure coding, identifying vulnerabilities in code (Shestov *et al.*, 2025), and providing guidance and remediation strategies. This aids in developing more secure applications. Tools such as SecurityLLM and FalconLLM facilitate the identification of malicious code patterns in vast data sets (Ferrag *et al.*, 2023). Other tools such as DeepExploit (Sychugov and Grekov, 2024) and PentestGPT (Deng *et al.*, 2024) allow automated security assessments of applications and IT infrastructure. Automatic detection of malware code also has been documented (Veprytska and Kharchenko, 2022).

- *Securing AI systems*: A vital defence is securing GenAI and LLM systems themselves against adversarial attacks such as data poisoning, model manipulation, and evasion. This requires robust defences, including adversarial machine learning defences, secure-by-design principles, continuous learning, and governance. Protecting the privacy and security of the large datasets used by these models is also critical, employing techniques such as differential privacy (Behnia *et al.*, 2022; Khoje, 2024) and federated learning (Xi *et al.*, 2024; Zhang *et al.*, 2024). Security testing frameworks for LLMs are also a defensive measure; defences against specific LLM vulnerabilities, such as prompt injection, are being developed (Ferrag *et al.*, 2024).

In essence, GenAI and LLMs are being developed actively and applied across multiple layers of defensive cybersecurity, from improving intelligence and detection to enabling dynamic response and deception, while also necessitating the development of defences for AI systems themselves.

## Discussion

This article explores selected aspects of the development of AI and its impact on the security of information systems, with particular emphasis on GenAI. It presents a literature review covering the period from 1993 to 2025, incorporating automated topic modelling and trend analysis of scientific publications related to AI and cybersecurity. In all, 3,389 scholarly publications were analysed from the IEEE Xplore Digital Library and SpringerLink databases. The findings revealed an explosive increase in research activity in this area, underscoring the growing significance of GenAI and LLMs in cybersecurity contexts. Turning to the results, the main research topics identified were: (1) LLMs with 734 publications (21.88%), (2) blockchain and cyberattacks, each with 394 publications (11.74%), (3) generative coding and software generation with 296 publications (8.82%), (4) smart energy infrastructure and IoT with 278 publications (8.29%), and (5) malware with 70 publications (2.09%).

However, it should be noted that these two databases may not comprehensively represent the full spectrum of global research on AI and cybersecurity. Future work will involve expanding the corpus by incorporating additional sources, such as Scopus, Digital Bibliography & Library Project (DBLP), and preprint repositories such as arXiv.

Based on the comprehensive analysis of the provided literature, a clear conclusion emerges regarding the transformative impact of GenAI and LLMs on the domain of cybersecurity. These technologies are not merely theoretical concepts but are being actively developed and deployed across both offensive and defensive strategies within the cyberspace landscape. The dual-use nature of GenAI and LLMs presents a significant dilemma, necessitating a proactive and adaptable approach to cybersecurity risk management and the development of "cyber resilience" solutions.

The case of Poland illustrates some of the new challenges. Poland is among the fastest-growing adopters of cloud computing in the European Union (EU); according to a recent Eurostat (2023) study on ICT services and e-commerce, it currently ranks tenth in the EU in terms of enterprise cloud usage. The rapid evolution of Poland's information technology ecosystem—driven by the deployment of global cloud infrastructures—has significantly enabled digital innovation. At the same time, it has increased exposure to

AI-augmented cyber threats, particularly those related to the processing and protection of sensitive organisational data by AI.

This development challenges traditional notions of "organisational boundaries" that require protection. In a rapidly changing environment shaped by the growing momentum of GenAI and LLM technologies, an increasing proportion of organisational data in the EU is likely to be generated, processed, and analysed in cloud environments—often outside traditional and on-premises data centres. Consequently, attack surfaces are shifting towards cloud infrastructures themselves, new AI models, and data availability through application programming interfaces (APIs) as well as edge computing components.

Edge computing, which distributes computation and data storage closer to the data source, further complicates this landscape. While it offers performance and scalability benefits, it also introduces new security challenges and opportunities for cybersecurity teams, particularly in the environments where LLMs are deployed across both cloud and edge layers (Zheng *et al.*, 2025).

## Conclusions

On the offensive front, cybercriminals are rapidly weaponising GenAI and LLMs to enhance the scale, sophistication, and automation of attacks. Key offensive applications include the facilitation of fraud and generation of malicious content or code, enabling the creation of metamorphic malware and automated ransomware tools capable of evading detection. Furthermore, these models enable sophisticated social engineering and phishing attacks, crafting persuasive language-based techniques and fuelling various illegal activities such as password cracking and typo squatting. The generation of deceptive content and deepfakes using models such as GANs poses risks for manipulating public opinion and enabling telecommunication fraud. Crucially, the increasing prevalence of AI means that AI systems themselves are becoming direct targets of adversarial attacks, such as data poisoning, model manipulation, and prompt injection.

In parallel, GenAI and LLMs are being harnessed to significantly enhance defensive cybersecurity capabilities across multiple layers. They are proving effective in improving threat intelligence and information processing through advanced analysis of security data, NER, knowledge graph construction, and automated report generation. For automated threat detection and analysis, GANs and VAEs simulate attacks and detect anomalies, while LLMs assist in interpreting IDS rules, predicting attacker TTPs, and analysing code for vulnerabilities. GenAI and LLMs also contribute to cyber deception by simulating adaptive honeypots and automating deception ploys, thus improving intelligence gathering. Furthermore, AI systems are automating real-time incident response and supporting vulnerability management by assisting secure coding and identifying code vulnerabilities. A critical component of the defensive strategy involves securing the GenAI and LLM systems themselves, requiring robust adversarial machine learning defences, secure-by-design principles, and protecting training datasets using techniques such as differential privacy and federated learning.

Therefore, in the context of dual-use dilemma associated with GenAI methods, organisations should urgently invest in technologies to secure their systems and strengthen employee competencies in threat analysis and incident response using similar GenAI-/LLM-driven techniques.

In conclusion, the evidence strongly indicates that GenAI and LLMs are deeply integrated into the contemporary cybersecurity landscape, serving as powerful tools for both attackers and defenders. Their application spans diverse areas from intelligence gathering and threat detection to automated response and deception. The rapid evolution and dual-use potential of these technologies necessitate continuous research, adaptation of security strategies, and the crucial development of defences specifically aimed at protecting AI systems themselves from malicious exploitation.

**Acknowledgement**

The author would like to thank Piotr Struzik for proofreading this manuscript as well as for his valuable remarks.

**Funding**

The research received no external funding.

**Disclosure Statement**

No potential conflict of interest was reported by the author. The author read and agreed to the published version of the manuscript.

**Data Availability Statement**

Not applicable.

## References

**Afolabi, A.S. and Akinola, O.A.** (2024) 'Vulnerable AI: A survey', in *2024 IEEE international symposium on technology and society (ISTAS)*. Puebla, Mexico, pp. 1–7. doi: 10.1109/ISTAS61960.2024.10732647.

**Ahmed, S. Mohaimenur Rahman, A.B.M., Alam, Md M. and Sajid, Md S.I.** (2025) 'Spade: Enhancing adaptive cyber deception strategies with generative AI and structured prompt engineering', in *2025 IEEE 15th annual computing and communication workshop and conference (CCWC)*. Piscataway, NJ: IEEE, pp. 01007–01013. doi: 10.1109/CCWC62904.2025.10903748.

**Amin, Q.K., Gillani, S.H.A.S., Shah, S.N.M. and Hussain, A.** (2024) 'A generative AI-driven CTI framework for IDs using machine learning and knowledge graph', in *2024 26th International multi-topic conference (INMIC)*. Piscataway, NJ: IEEE, pp. 1–6. doi: 10.1109/INMIC64792.2024.11004337.

**Ankalaki, S., Rajesh, A.A., Pallavi, M., Hukkeri, G.S., Jan, T. and Naik, G.R.** (2025) 'Cyber attack prediction: From traditional machine learning to generative artificial intelligence', *IEEE Access*, 13, 3547433. doi: 10.1109/ACCESS.2025.3547433.

**Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., *et al.*** (2023) 'Identifying and mitigating the security risks of generative AI', *Foundations and Trends® in Privacy and Security*, 6(1), pp. 1–52. doi: 10.1561/3300000041.

**Behnia, R., Ebrahimi, M.R., Pacheco, J. and Padmanabhan, B.** (2022) 'EW-tune: A framework for privately fine-tuning large language models with differential privacy', in *2022 IEEE international conference on data mining workshops (ICDMW)*. Piscataway, NJ: IEEE, pp. 560–566. doi: 10.1109/ICDMW58026.2022.00078.

**Beninger, M., Charland, P., Ding, S.H.H. and Fung, Benjamin C.M.** (2024) 'Ers0: Enhancing military cybersecurity with AI-driven SBOM for firmware vulnerability detection and asset management', in *2024 16th International conference on cyber conflict: over the horizon (CyCon)*. Tallinn, Estonia, pp. 141–160. doi: 10.23919/CyCon62501.2024.10685598.

**Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., *et al.*** (2018) *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Technical report, University of Oxford. Available at: https://maliciousaireport.com/ (Accessed: 15.05.2025).

**Childs, L. O.P. and Wood, M. A.** (2024) 'Hypnotic honey: Adaptive honeypots managed by local large language models', in *2024 Cyber research conference-Ireland (Cyber-RCI)*. Carlow, Ireland, pp. 1–8. doi: 10.1109/Cyber-RCI60769.2024.10941449.

**Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, Anil A.** (2018) 'Generative adversarial networks: An overview', *IEEE Signal Processing Magazine*, 35(1), pp. 53–65. doi: 10.1109/MSP.2017.2765202.

**Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., *et al.*** (2024) '{PentestGPT}: Evaluating and harnessing large language models for automated penetration testing', in *33rd USENIX security symposium (USENIX Security 24).* pp. 847–864. USENIX Association. Berkeley, CA, United States.

**Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.** (2019) 'Bert: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

**Enathe V.P.S., Mouli S.C. and Dheepthi, R.** (2024) 'Llm-enhanced deepfake detection: Dense CNN and multi-modal fusion framework for precise multimedia authentication', in *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*. Chennai, India, pp. 1–6. doi: 10.1109/ADICS58448.2024.10533511.

**Eurostat** (2023) *Cloud computing—Statistics on the use by enterprises from 2023 EU survey on ICT usage and E-commerce in enterprises*. Technical report. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php (Accessed: 15.05.2025).

**Ferrag, M.A., Ndhlovu, M., Tihanyi, N., Cordeiro, Lucas C., Debbah, M. and Lestable, T.** (2023) 'Revolutionizing cyber threat detection with large language models', *arXiv preprint*, arXiv:2306.14263, pp. 195–202. doi: 10.48550/arXiv.2306.14263

**Ferrag, M.A., Ndhlovu, M., Tihanyi, N., Cordeiro, Lucas C., Debbah, M., Lestable. T.** (2024) 'Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for IoT/IIoT devices', *IEEE Access*, 12, pp. 23733–23750. doi: 10.1109/ACCESS.2024.3363469.

**Firdhous, M.F.M., Elbreiki, W., Abdullahi, I., Sudantha, B.H. and Budiarto, R.** (2023) 'Wormgpt: A large language model chatbot for criminals', in *2023 24th International Arab conference on information technology (ACIT)*. Ajman, United Arab Emirates, pp. 1–6. doi: 10.1109/ACIT58888.2023.10453752.

**Fu, M., Tantithamthavorn, C.K., Nguyen, V. and Le, T.** (2023) 'Chatgpt for vulnerability detection, classification, and repair: How far are we?', in *2023 30th Asia-Pacific software engineering conference (APSEC)*. Seoul, Korea, Republic of, pp. 632–636. doi: 10.1109/APSEC60848.2023.00085.

**Ghazal, T.M., Janjua, J.I., Abushiba, W. Ahmad, M., Ihsan, A. and Al-Dmour, N.A.** (2024) 'Cybersecurity revolution via large language models and explainable AI', in *2024 17th International conference on security of information and networks (SIN)*. Sydney, Australia, pp. 1–6. doi: 10.1109/SIN63213.2024.10871324.

**Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Farley, D.W. and Ozair, S.** (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, 27, 2672–2680. doi: 10.48550/arXiv.1406.2661

**Gupta, M., Akiri, C.K., Aryal, K., Parker, E. and Praharaj, L.** (2023) 'From chatgpt to threatgpt: Impact of generative AI in cybersecurity and privacy', *IEEE Access*, 11, pp. 80218–80245. doi: 10.1109/ACCESS.2023.3300381.

**Hasanov, I., Virtanen, S., Hakkala, A. and Isoaho, J.** (2024) 'Application of large language models in cybersecurity: A systematic literature review', *IEEE Access*. 12, pp. 176751–176778. doi: 10.1109/ACCESS.2024.3505983.

**Heldens, S., Sclocco, A., Dreuning, H., Werkhoven, B.V., Hijma, P., Maassen, J., *et al.*** (2022) *Litstudy*: A python package for literature reviews', *SoftwareX*, 20, p. 101207. doi: 10.1016/j.softx.2022.101207.

**Imtiaz, A., Shehzad, D., Nasim, F., Afzaal, M., Rehman, M. and Imran, A.** (2023) 'Analysis of cybersecurity measures for detection, prevention, and misbehaviour of social systems', in *2023 Tenth international conference on social networks analysis, management and security (SNAMS)*. Abu Dhabi, United Arab Emirates, pp. 1–7. doi: 10.1109/SNAMS60348.2023.10375405.

**Karampasi, A., Radoglou-Grammatikis, P., Pawlicki, M., Choraś, R., Pozuelo, R.M. De, Sarigiannidis, P., Puchalski, D., Pawlicka, A., Kozik, R. and Choraś, M.** (2024) 'Towards transparent AI-powered cybersecurity in financial systems: The deployment of federated learning and explainable AI in the caixabank pilot', in *2024 IEEE International conference on data mining workshops (ICDMW)*. Abu Dhabi, United Arab Emirates, pp. 270–277. doi: 10.1109/ICDMW65004.2024.00041.

**Khoje, M.** (2024) 'Navigating data privacy and analytics: The role of large language models in masking conversational data in data platforms', in *2024 IEEE 3rd International conference on AI in cybersecurity (ICAIC)*. Houston, TX, USA, pp. 1–5. doi: 10.1109/ICAIC60265.2024.10433801.

**Klopper C. and Eloff, Jan H.P.** (2024) 'Data fingerprinting and visualization for AI-enhanced cyber-defence systems', *IEEE Access. 12, 154054–154065.* doi: 10.1109/ACCESS.2024.3482728.

**Madani, P.** (2023) 'Metamorphic malware evolution: The potential and peril of large language models', in *2023 5th IEEE international conference on trust, privacy and security in intelligent systems and applications (TPSISA)*. Atlanta, GA, USA, pp. 74–81. doi: 10.1109/TPS-ISA58951.2023.00019.

**McInnes, L., Healy, J. and Melville, J.** (2018) 'UMAP: Uniform Manifold Approximation and Projection'. *Journal of Open Source Software,* 3(29), 861. doi: 10.21105/joss.00861.

**Mi X. and Zhang, B.** (2025) 'Digital communication strategies for coping with deepfake content distribution', in *2025 Communication strategies in digital society seminar (ComSDS).* Saint Petersburg, Russian Federation, pp. 82–86. doi: 10.1109/ComSDS65569.2025.10971333.

**National Institute of Standards and Technology (NIST)** (2018) *Framework for improving critical infrastructure cybersecurity, version 1.1, 2018*. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf (Accessed: 15.05.2025).

**OpenAI** (2023) *Gpt-4 technical report, 2023*. Available at: https://openai.com/research/gpt-4 (Accessed: 15.05.2025).

**Patel, U.K., Yeh, F.-C. and Gondhalekar, C.** (2024) 'Canal-cyber activity news alerting language model: Empirical approach vs. expensive LLMS', in *2024 IEEE 3rd international conference on AI in cybersecurity (ICAIC)*. Houston, TX, USA, pp. 1–12. doi: 10.1109/ICAIC60265.2024.10433839.

**Qi, J., Huang, S., Luan, Z., Fung, C.J., Yang, H. and Qian, D.** (2023) 'LogGPT: Exploring ChatGPT for Log-Based Anomaly Detection', in *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys).* Melbourne, Australia, pp. 273–280, doi: 10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00045.

**Qiao, Z., Zhang, C. and Du, G.** (2023) 'Improving cybersecurity named entity recognition with large language models', in *2023 6th International conference on software engineering and computer science (CSECS)*. Chengdu, China, pp. 1–6. doi: 10.1109/CSECS60003.2023.10428218.

**Ragsdale, J. and Boppana, R.V.** (2023) 'On designing low-risk honeypots using generative pre-trained transformer models with curated inputs', *IEEE Access*, 11, pp. 117528–117545. doi: 10.1109/ACCESS.2023.3326104.

**Samonte, M.J.C., Goc-ong, A.E., Matoza, R.B.F. and Viernes, R.G.A.** (2024) 'Evaluating the effectiveness of artificial intelligence in integrated system architectures to combat cybersecurity threats', in *2024 IEEE 7th international conference on computer and communication engineering technology (CCET)*. Beijing, China, pp. 222–226. doi: 10.1109/CCET62233.2024.10838195.

**Sarika, P.S., Paul, A., Kumar, A., George, A.** (2024) 'AI meets cyber defense: Enhancing network security with GAN-driven NIDS', in *2024 11th International conference on advances in computing and communications (ICACC)*. Kochi, India, pp. 1–6. doi: 10.1109/ICACC63692.2024.10845511.

**Shestov, A., Levichev, R., Mussabayev, R., Maslov, E., Zadorozhny, P., Cheshkov, A., Mussabayev, R., Toleu, A., Tolegen, G. and Krassovitskiy, A.** (2025) 'Finetuning large language models for vulnerability detection', *IEEE Access*. doi: 10.1109/ACCESS.2025.3546700.

**Shibli, A Md., Pritom, M.M.A. and Gupta, M.** (2024) 'AbuseGPT: Abuse of generative AI chatbots to create smishing campaigns', in *2024 12th International symposium on digital forensics and security (ISDFS)*. San Antonio, TX, USA, pp. 1–6. doi: 10.1109/ISDFS60797.2024.10527300.

**Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K. and Wang, J.** (2019) 'Release strategies and the social impacts of language models', *arXiv preprint, arXiv:1908.09203*. doi: 10.48550/arXiv.1908.09203.

**Sultana, M., Taylor, A., Li, L. and Majumdar, S.** (2023) 'Towards evaluation and understanding of large language models for cyber operation automation', in *2023 IEEE conference on communications and network security (CNS)*. Orlando, FL, USA, pp. 1–6. doi: 10.1109/CNS59707.2023.10288677.

**Sychugov, A. and Grekov, M.** (2024) 'Automated penetration testing based on adversarial inverse reinforcement learning', in *2024 International Russian smart industry conference (SmartIndustryCon)*. Sochi, Russian Federation, pp. 373–377. doi: 10.1109/SmartIndustryCon61328.2024.10515504.

**Team Google** (2024) *Gemini: A family of highly capable multimodal models*. Available at: https://arxiv.org/abs/2312.11805 (Accessed: 16.05.2025).

**Usman, Y., Gyawali, P.K., Gyawali, S. and Chataut R.** (2024) 'The dark side of AI: Large language models as tools for cyber attacks on vehicle systems', in *2024 IEEE 15th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*. Yorktown Heights, NY, USA, pp. 169–175. doi: 10.1109/UEMCON62879.2024.10754676.

**Vadisetty, R. and Polamarasetti, A.** (2024) 'Generative AI-driven distributed cybersecurity frameworks for AI-integrated global big data systems', in *2024 International conference on emerging technologies and innovation for sustainability (EmergIN)*. Greater Noida, India, pp. 595–600. doi: 10.1109/EmergIN63207.2024.10961616.

**Van der Maaten, L. and Hinton, G.** (2008) 'Visualizing data using t-sne', *Journal of Machine Learning Research*, 9(11), pp. 2579–2605.

**Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I.** (2017) 'Attention is all you need', in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, p. 30. doi: 10.48550/arXiv.1706.03762.

**Veprytska, O. and Kharchenko, V.** (2022) 'AI-powered attacks against AI-powered protection: Classification, scenarios and risk analysis', in *2022 12th International conference on dependable systems, services and technologies (DESSERT)*. Athens, Greece, pp. 1–7. doi: 10.1109/DESSERT58054.2022.10018770.

**Xi, Y., Sheng, Z. and Huan, Y.** (2024) 'Spatio-temporal modeling methods for point cloud video based on federated learning', in *2024 5th International conference on computer, big data and artificial intelligence (ICCBD+AI)*. Jingdezhen, China, pp. 501–506. doi: 10.1109/ICCBD-AI65562.2024.00089.

**Xu, W., Liu, X. and Gong, Y.** (2003) 'Document clustering based on nonnegative matrix factorization', in *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 267–273. doi: 10.1145/860435.860485.

**Yigit, Y., Buchanan, W.J., Tehrani, M.G. and Maglaras, L.** (2024) 'Review of generative AI methods in cybersecurity', *arXiv preprint*, arXiv:2403.08701. doi: 10.48550/arXiv.2403.08701.

**Zhang, J., Vahidian, S., Kuo, M., Li, C., Zhang, R., Yu, T., Wang, G. and Chen, Y.** (2024) 'Towards building the federatedgpt: Federated instruction tuning', in *ICASSP 2024—2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Seoul, Korea, Republic of, pp. 6915–6919. doi: 10.1109/ICASSP48485.2024.10447454.

**Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y. and Chen, J.** (2025) 'A review on edge large language models: Design, execution, and applications', *ACM Computing Surveys*, 57(8), pp. 1–35. doi: 10.1145/3719664.