INNOVATIONS IN MACHINE INTELLIGENCE
NATURAL PRINCIPLES AND ARTIFICIAL MODELS

Innovations in Machine Intelligence (IMI), 2025, 28–34

doi: 10.54854/imi2025.04
Review

REVIEW

The Cybersecurity Paradox: How Large Language Models Protect and Threaten Digital Security and Human Rights

Alice Saito [1,*]

[1]University of Tokyo

* Corresponding Author: alicesaito14@g.ecc.u-tokyo.ac.jp

Abstract

Large Language Models (LLMs) have dramatically changed the digital world in recent years. These systems can generate human-like text, summarize information, write code, answer complex questions, and analyze enormous amounts of data in seconds. However, the rapid adoption of LLMs has also raised serious questions, especially in the field of cybersecurity. On one hand, these models can strengthen digital security by detecting threats, analyzing suspicious activity, automating security responses, and helping experts understand complex cyberattacks more quickly. On the other hand, the same capabilities can be misused by malicious actors. LLMs can be used to generate convincing phishing emails, write malware code, spread misinformation at scale, or assist in social engineering attacks. This creates what is often described as a cybersecurity paradox: the very technologies designed to protect digital systems can also be turned into powerful tools for harm. Beyond security risks, LLMs raise important ethical and legal concerns. One major issue is privacy. Many of these models are trained on massive datasets collected from the internet, often without the clear consent of the individuals who created that data. This raises questions about data ownership, personal privacy, and whether sensitive or copyrighted information is being used responsibly. There are also concerns about accountability. When an LLM produces harmful, biased, or incorrect information, it is often unclear who should be held responsible—the developers, the users, or the organizations deploying the system. This article examines how LLMs sit at the intersection of rapid technological innovation and complex ethical challenges. It explores their growing influence on global digital security while highlighting the risks they pose to privacy, trust, and human rights. Finally, it discusses the significant governance challenges surrounding LLMs and emphasizes the urgent need for proactive, coordinated international regulation. Without clear rules, transparency, and global cooperation, the benefits of LLMs may be overshadowed by their risks, making responsible oversight not just important but essential.

Key words: Large Language Models; Cybersecurity; Human Rights; Safety and privacy

Introduction

Large Language Models (LLMs) have entered the mainstream, offering impressive capabilities in natural language understanding and generation. In recent years, LLMs have gained significant attention due to their ability to perform a wide range of tasks, from summarising to generating code. However, their growing prominence is not only limited to everyday applications. In the field of cybersecurity, their influence is increasingly felt across both defensive and offensive fronts. The same algorithms that detect fraud or assist with forensic analysis can also be repurposed to generate malicious code or deceive users [1]. Malicious users may use these models to generate deceptive content and code, or manipulate users into revealing sensitive information. This paradox raises serious questions about the role of artificial intelligence in safeguarding or endangering our digital environments. As these LLMs become more deeply integrated into critical systems, the stakes grow higher. This article evaluates how LLMs are simultaneously defending and undermining cybersecurity, highlighting the broader implications for privacy, transparency, and human rights.

Received: 16 Jul 2021, Accepted: 20 October 2021, Published online: 20 November 2021 March 20, 2026.

Large Language Models (LLMs) are advanced artificial intelligence (AI) models that generate human-like text by predicting the next word in a sentence. They belong to a broader field known as natural language processing (NLP), which focuses on enabling machines to interpret, generate, and interact with human language. LLMs are distinguished by their ability to process vast amounts of text data, learning complex patterns, contextual cues, and syntactic structures that allow them to generate coherent and contextually appropriate responses. At the core of their functionality is the transformer architecture, a neural network structure that uses self-attention mechanisms to weigh the importance of different words in a sequence. This allows LLMs to process and retain contextual information over long passages of text. Through multiple layers of encoders and decoders, LLMs refine their understanding of languages, adapting to different contexts and improving response accuracy. These layers allow LLMs to capture increasingly abstract linguistic features—from simple word meanings to complex sentence structures and even nuanced stylistic or emotional tones [2].

Training a Large Language Model (LLM) involves processing vast collections of textual data—ranging from books and academic articles to websites and social media posts. This exposure enables the model to identify and internalise linguistic structures, patterns, and relationships. The training process is typically self-supervised, meaning the model learns without the need for manually labelled data. Instead, it predicts missing or next words in a sentence based on context, gradually building an understanding of syntax, semantics, and usage. Once trained, LLMs are capable of performing a diverse array of language-related tasks, such as completing sentences, translating text, summarising content, analysing sentiment, and answering questions.

The evolution of Large Language Models (LLMs) has been marked by significant advancements in AI and deep learning over the past decade. Early models, such as word embeddings like Word2Vec (2013) [3] and GloVe (2014) [4], laid the groundwork by representing words as vectors, capturing semantic relationships between them. Words that appeared in similar contexts were placed close together in this space, allowing machines to perform rudimentary tasks such as measuring word similarity and completing analogies. While groundbreaking at the time, these embeddings had limitations, particularly in handling polysemy, words with multiple meanings, and contextual nuance [5].

The introduction of the transformer architecture in 2017 replaced the older neural networks that struggled with long-range dependencies in text. The self-attention mechanism revolutionised natural language processing by enabling models to process context more effectively. The model was able to weigh the importance of different words in a sentence relative to one another, regardless of their position. This was key to capturing contextual relationships and improving the training efficiency and performance. Building on this architecture, researchers developed increasingly powerful models. BERT (Bidirectional Encoder Representations from Transformers) [6], released by Google in 2018, introduced bidirectional training, meaning it learned context from both the left and right of a word in a sentence, leading to significant gains in tasks like question answering and language inference. Around the same time, OpenAI's GPT (Generative Pre-trained Transformer) took a different approach, focusing on autoregressive language modeling—predicting the next word in a sequence from left to right. GPT-2 (2019) [7] and GPT-3 (2020) [8] scaled up the number of parameters dramatically, enabling more fluent and coherent text generation.

One area where LLMS have become especially valuable is cybersecurity. As LLMS have advances their ability ot understand and generate human-like language, thet have found critical applications beyond the natural scope of NLP tasks. By leveraging the same transformer-based architecture that enables nuanced language understanding, LLMs can help detect subtle anomalies hidden in language, tasks that are often too time-consuming or intricate for traditional systems. This intersection between AI language capabilities and cybersecurity reflects how the evolution of LLMs is not just about improving language modelling, but about reshaping how we defend against emerging digital threats [5]. Fig. 1 info-graphically depicts what LLMs are and can do.

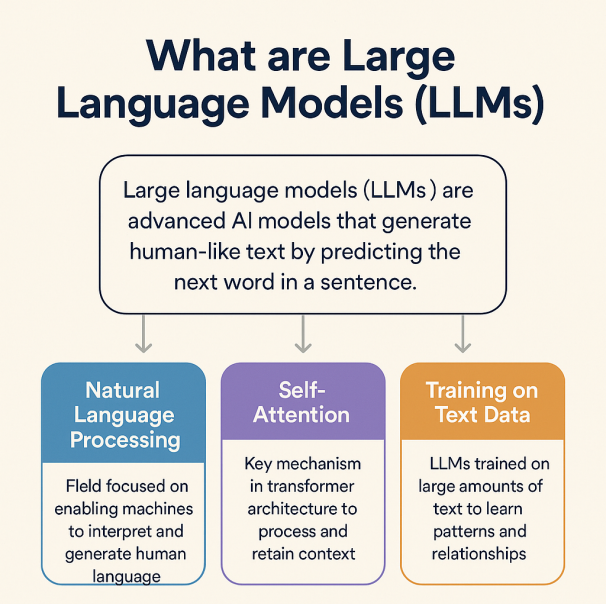

Figure 1. What are LLMs?

LLM For Cybersecurity

The integration of Large Language Models (LLMs) into cybersecurity has opened up new avenues for threat detection, vulnerability analysis, and personal data protection. With their advanced capabilities in language understanding, pattern recognition, and contextual reasoning, LLMs are not only augmenting traditional cybersecurity tools but also enabling more proactive and adaptive defence strategies. Their applications are in several critical areas.

Identifying and Mitigating

LLMs significantly enhance vulnerability detection and mitigation by addressing issues that conventional tools often miss. Thanks to their deep understanding of programming languages and standard coding practices, LLMs can review source code for common flaws such as excessive memory usage, buffer overflows, and serialisation/deserialization inconsistencies—vulnerabilities often exploited by attackers. Serialisation, which involves converting objects into byte streams for storage or transmission, can create security gaps if not implemented correctly. LLMs help detect such weaknesses by simulating the conversion process and identifying irregularities or risks [1].

Malware Detection

Another core area where LLMs are applied is Malware Detection. Unlike traditional antivirus systems that rely on known attack patterns, LLMs can generalise from existing malware datasets to identify novel or obfuscated threats. They process vast datasets of malicious code, system logs, and behavioural patterns to uncover recurring characteristics and attack strategies. This enables LLMs to identify and respond to new or evolving malware variants that may

not yet be documented in existing threat databases.

What sets LLMs apart is their ability to incorporate multi-dimensional data, from file metadata and code structure to behavioural logs and communication patterns. This enables a more holistic approach to malware analysis. Their contextual awareness lets them flag abnormal behaviours or command-and-control signals even when attackers attempt to evade detection through code obfuscation or polymorphism. LLMs also outperform many traditional static analysis tools, which depend on predefined rule sets that require constant manual updates.

By analysing a wide array of malware samples, LLMs play a critical role in identifying recurring patterns and malicious behaviours. Their ability to integrate multi-dimensional data enables a comprehensive approach to malware analysis, offering stronger detection mechanisms (nsfocus). LLMs use data mining to analyse raw data, identifying complex patterns and vulnerabilities that static code scanning tools, reliant on manually maintained rule sets, often miss. Their contextual understanding allows them to detect more sophisticated attack scenarios overlooked by traditional scanners [9].

Protecting Personal Information

LLMs also play an increasingly important role in safeguarding personal data and preserving user privacy. One major application is the detection of phishing emails, where attackers impersonate trusted entities to extract sensitive information. LLMs improve phishing detection by analysing not just keywords or URLs, but the overall semantic and stylistic patterns of communications—such as urgency, tone, or psychological manipulation tactics. They significantly reduce false positives while improving detection accuracy, thereby enhancing user trust in automated systems.

Furthermore, they facilitate privacy leak detection by identifying Personally Identifiable Information (PII) —such as names, addresses, or credit card numbers- within large datasets. They can perform this analysis across multiple languages, making them valuable assets in international cybersecurity operations. According to recent studies, LLMs have demonstrated a 20% improvement in phishing detection rates and a 30% reduction in false positives compared to traditional filtering methods, highlighting their potential in real-world cybersecurity [10].

The key applications of LLMs in Cybersecurity can be summarized as follows:

- **Threat Intelligence:** Extracting, organizing, and analyzing large volumes of threat intelligence documents to detect trends and patterns that may indicate emerging threats.
- **Vulnerability Detection:** Identifying security vulnerabilities through code analysis and real-time threat monitoring, often outperforming traditional static analysis tools.
- **Malware Detection:** Enhancing both static and dynamic analysis by identifying recurring patterns and predicting malicious behaviors. LLMs improve malware classification and anomaly detection through multi-dimensional data integration.
- **Anomaly Detection:** Identifying security anomalies such as malicious network traffic, virus files, and anomalies in system logs.
- **Fuzzing and Program Repair:** Automating fuzz testing to discover vulnerabilities and assisting in program repair by generating secure patches.
- **LLM-Assisted (In)Secure Code Generation:** While LLMs excel at generating secure code, concerns remain about the potential risks of LLM-generated insecure code. However, research is ongoing to develop strategies that allow LLMs to self-correct and refine their code outputs.
- **LLM-Assisted Attacks:** Despite their positive applications, LLMs can also be exploited to launch network attacks such as generating phishing emails and assisting in penetration testing.
- **Strengthening Incident Response and Data Protection:** LLMs also enhance incident response by generating detailed reports, summarizing security incidents, and recommending appropriate mitigation steps. Their ability to process vast amounts of security data in real-time accelerates threat identification and minimizes response time.
- **Phishing Email Detection:** LLMs safeguard sensitive information by analyzing email content and communication patterns to identify phishing scams, resulting in a 20% improvement in detection rates and a 30% reduction in false positives compared to traditional filtering methods. Privacy Leak Detection: Identifying leaks of Personally Identifiable Information (PII), including cross-language detection of private data, is essential in globalized cybersecurity frameworks.

Fig. 2 info-graphically depicts what are LLMs key applications in Cybersecurity.

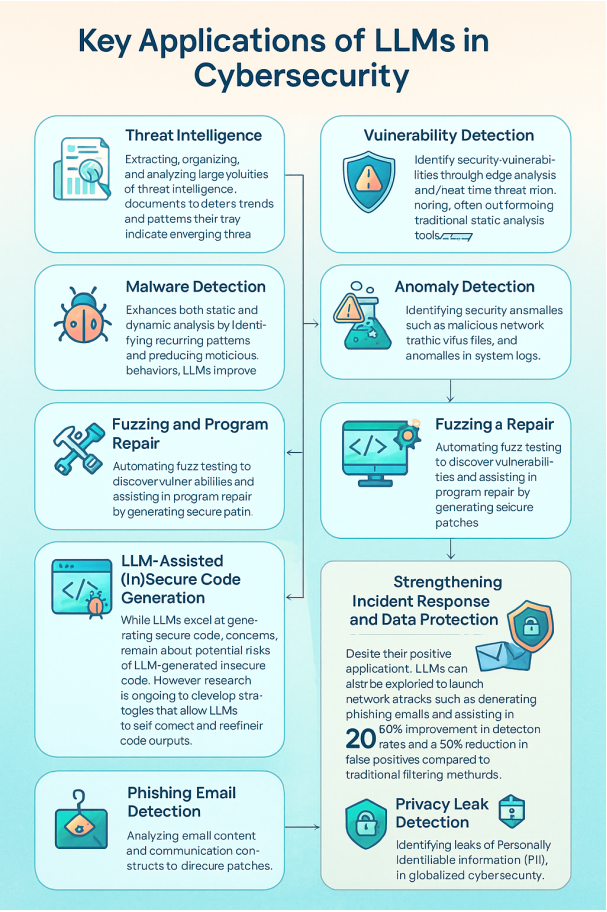

Figure 2. LLMs in Cybersecurity.

LLM against Cybersecurity

While Large Language Models (LLMs) offer transformative capabilities in strengthening cybersecurity, these benefits are increasingly challenged by the dual-use nature of the technology. The very features that make LLMs powerful defensive tools, such as rapid data processing, code generation, and context-aware analysis, can also be exploited by malicious actors to enhance the scale, precision, and impact of cyberattacks.

Weaponisation by Malicious Actors

One of the most alarming risks posed by LLMs is their potential to streamline and automate sophisticated cyberattacks. Attackers can use LLMs to conduct reconnaissance on systems, identify vulnerabilities, and execute attacks such as man-in-the-middle intrusions, privilege escalation, and injection attacks with increased efficiency. What once required a deep understanding of security protocols and coding expertise can now be accomplished more easily through conversational prompts and code generation tools powered by LLMs.

A particularly dangerous application is the evolution of phishing tactics. Traditionally, phishing emails were often marked by poor grammar, awkward phrasing, or cultural inconsistencies, clues that helped recipients identify them as fraudulent. With LLMs, however, cybercriminals can generate highly convincing and grammatically accurate phishing emails, tailored to specific individuals or organisations. These messages can mimic legitimate communication styles, reference real-life events, and adapt across multiple languages, making them significantly more difficult to detect. This personalisation greatly increases the success rate of social engineering attacks, in which victims are tricked into sharing sensitive information such as passwords or financial data [11].

Beyond phishing, LLMs have also been harnessed to generate disinformation content at scale, fueling misinformation campaigns and psychological manipulation efforts. By rapidly producing persuasive narratives, fake news, or impersonated communications, these models can distort public discourse, spread propaganda, and erode trust in digital communications [12].

Inherent Technical Risks

Apart from being abused externally, LLMs themselves carry inherent limitations that can weaken cybersecurity efforts. For instance, these models can be used to automate the development of malicious code, including polymorphic malware: a type of malware that constantly changes its structure to evade detection by traditional signature-based scanners. This capability undermines conventional defences and forces cybersecurity teams to develop more adaptive, behaviour-based detection strategies [13].

Moreover, the lack of transparency in how many commercial LLMs are trained and operate introduces further concerns. Hidden biases, unvetted data sources, and the inability to audit the internal workings of closed-source models raise questions about their reliability in high-stakes environments. These blind spots could be exploited by attackers who understand how to manipulate model behaviour or input formatting to evade detection [14].

- Code Security
  - Standardized code generation
  - Detecting vulnerabilities (e.g., serialization issues, memory flaws)
  - Automated security testing (fuzzing, data mining)
- Malware Detection
  - Identifying patterns in malicious code
  - Integrating multi-dimensional data for analysis
- Personal Information Protection
  - Phishing detection and prevention
  - Identifying privacy leaks and cross-language PII detection

Fig. 3 presents an infographic highlighting key applications of large language models (LLMs) in the cybersecurity domain.

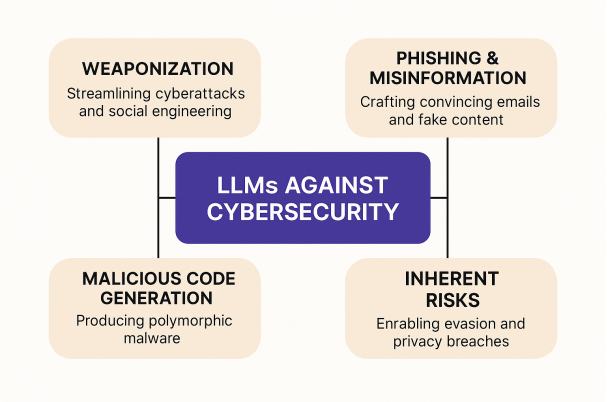

Figure 3. LLMs against Cybersecurity.

Case Studies:

As artificial intelligence continues to revolutionise industries, its rapid advancement brings along a host of pressing security concerns. ChatGPT, along with numerous other LLM models, has captivated millions with its ability to produce human-like text. However, recent incidents highlight vulnerabilities that raise critical questions about data privacy, intellectual property, and user safety. From training data leaks to high-profile lawsuits – how pressing is the need to address these challenges?

Training Data Leaks

Recent research has exposed a weakness in ChatGPT's design, which allows attackers to extract training data. This exploit, known as extractable memorisation, takes advantage of the model's tendency to "remember" fragments of its training data. Researchers demonstrated that by using a specific prompt, they could manipulate ChatGPT into disclosing sensitive information from its training data. In this case, they prompted ChatGPT to repeat the word "poem" infinitely, causing the model to emit real email addresses and phone numbers. In the most extreme case, over 5% of the model's output was a direct, verbatim copy of 50 consecutive tokens from its training data.

This vulnerability occurs because the model bypasses its alignment safeguard and reverts to its pre-training data. Despite GPT-4 being explicitly designed to avoid revealing training data, this exploit highlights a significant flaw in the model's ability to balance recall with privacy protection.

The implications of this discovery are profound and underscore a broader challenge within LLMs: the difficulty of ensuring privacy and data security as these systems handle increasingly sensitive and personal information. This issue emphasises the need for stringent testing, oversight, and continuous refinement of safeguards to prevent unintentional data exposure. As LLMs like ChatGPT are deployed across diverse fields, from customer service to healthcare, maintaining user privacy while ensuring the functionality of these models remains a persistent and critical challenge [15].

March 20 Outage

On March 20, 2023, ChatGPT experienced a significant privacy breach triggered by a bug in an open-source library. This flaw resulted in a temporary but serious compromise of user data. During the outage, some users were able to access the conversation histories of other users, raising concerns about the system's ability to protect personal information. More alarmingly, for a small subset of ChatGPT Plus subscribers, sensitive payment details, including names, billing addresses, and the last four digits of credit

card numbers, were unintentionally disclosed [16]. Fig. 4 illustrates an example of March 20, 2023 exposure of billing information of the users.

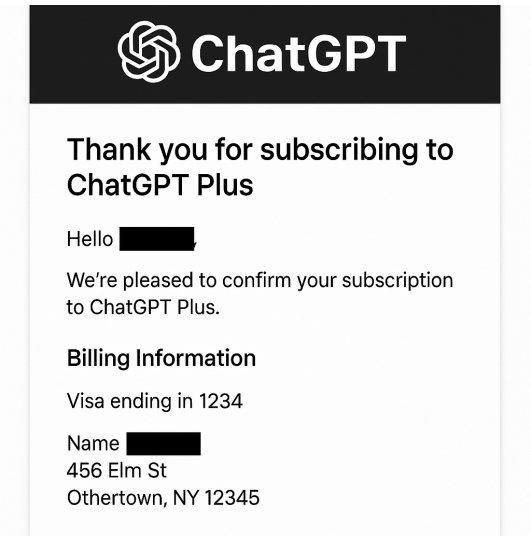

Figure 4. Simulated example of the March 20, 2023 ChatGPT Plus billing information exposure. This image is a reconstruction illustrating how the bug caused some users to view other subscribers' billing details.

This incident, while contained, underscored the vulnerabilities inherent in managing large-scale user data. OpenAI swiftly identified and rectified the issue, assuring users that steps were taken to address the bug and prevent further breaches. The company reaffirmed its commitment to safeguarding user privacy and data security, emphasising its responsibility to protect sensitive information.

Legal Battles

The growing integration of AI technology is sparking a wave of legal and ethical debates, with lawsuits threatening to reshape the legal boundaries of AI. Beyond privacy concerns, AI systems like ChatGPT are facing increasing scrutiny over their use of copyrighted material. In a high-profile lawsuit, The New York Times sued OpenAI and Microsoft, alleging the unauthorised use of its content to train AI models. This legal challenge underscores the ethical and legal dilemmas surrounding the development of generative AI.

The lawsuit claims that copyrighted articles were used without permission to train OpenAI's ChatGPT and Microsoft's Bing Chat or "Copilot." The latest models, which are trained on trillions of words, reportedly include a substantial amount of content from The Times' copyrighted work. The Times argues that this data contains a significant portion of its proprietary content, which was used without compensation.

As previously noted, LLMs tend to memorise parts of the works included in their training data. This means the model can generate near-verbatim extracts from copyrighted sources. In addition to this, LLMs produce "synthetic" search results, which can significantly exceed the amount of content traditionally displayed by an online search. This functionality allows users to bypass paywalls and access content that would otherwise be restricted, enabling readers to generate summaries of near-verbatim extracts without the need to pay for access.

This situation presents a serious threat to journalism, as it potentially undermines the intellectual property rights of content creators. By allegedly providing free access to premium content, LLMs like ChatGPT could erode traditional revenue models for news outlets. The Times has cited multiple instances where users were presented with near-verbatim extracts of its articles, content that would otherwise only be available to paying subscribers. The Times has warned that the "cost to society will be enormous" if such practices continue, stressing the crucial role of independent journalism in a functioning democracy.

The lawsuit highlights the broader challenge of creating AI systems that respect intellectual property rights while still benefiting from the vast array of resources required to train these models. Beyond ownership concerns, there is growing anxiety about AI emerging as a direct competitor to news outlets, drawing away audiences and, consequently, diminishing revenue streams. This case raises critical questions about how AI can be developed in a way that balances innovation with fair compensation for creators and organisations [17].

The controversies extend beyond journalism. Comedian Sarah Silverman joined lawsuits accusing Meta and OpenAI of using her memoir without consent in training their AI models [18]. Prominent authors such as Jonathan Franzen and John Grisham have voiced similar concerns, leading to a separate lawsuit focused on the unauthorised use of thousands of books. Additionally, Getty Images sued an AI platform for generating visuals based on its copyrighted material, arguing this practice undermines creative industries [19, 20].

These legal battles highlight the tension between technological advancement and intellectual property rights. As generative AI continues to evolve, these disputes will play a pivotal role in shaping how creative content is protected and utilised in the digital age. These legal battles highlight the tension between technological advancement and intellectual property rights. As generative AI continues to evolve, these disputes will play a pivotal role in shaping how creative content is protected and utilised in the digital age.

Human Rights Implications of LLMs in Cybersecurity

Right to be forgotten

The "right to be forgotten" refers to an individual's ability to have certain personal information removed from public access, particularly on the internet. This right is enshrined in data protection laws, such as the European Union's General Data Protection Regulation (GDPR), which mandates that individuals can request the deletion of their personal data under certain conditions. This right is an essential part of Article 12 of the Universal Declaration of Human Rights (UDHR), which stipulates the right to privacy [21].

However, the deployment of Large Language Models (LLMs) in cybersecurity complicates this right, as these models often rely on vast datasets, some of which may contain personal or sensitive information. LLMs often rely on vast datasets scraped from various publicly available sources, including social media, academic articles, and even personal blogs. While these datasets might not always be overtly private, they may still contain sensitive or personal information without the knowledge or consent of the individuals involved. As LLMs are trained on such data, they can inadvertently memorise details about individuals, making it difficult to delete specific pieces of personal data from the model once they are embedded. This creates a significant challenge in reconciling the right to be forgotten with the functionality of LLMs in cybersecurity.

A particularly concerning issue is the "memorisation" of personal data by LLMs, where the model retains information from its training data and can recall it during its operations. This could potentially expose private information that was never intended for such use. Furthermore, LLMs are prone to a phenomenon called "hallucination," where the model generates outputs that seem plausible but are not based on any actual training data. While hallucinations are often seen as a flaw in LLMs, they could inadvertently lead

to the generation of misinformation that harms individuals, making it even more challenging to manage privacy and data protection within cybersecurity systems.

Misinformation and Disinformation

Misinformation and disinformation are growing concerns in the context of human rights, particularly with the rise of LLMs in cybersecurity (Oxford). Misinformation refers to the spread of inaccurate information without malicious intent, while disinformation is the intentional dissemination of false or misleading content with the purpose of deceiving others. The digital age has significantly amplified the reach of both, making it easier to spread and access misleading information, which can have harmful impacts on public opinion, social harmony, and democratic processes [22].

In recent years, the use of digital platforms has enabled malicious actors to exploit these technologies to spread disinformation. Political groups, criminal organizations, terrorist cells, and even governments can take advantage of LLMs to create persuasive narratives that align with their agendas (Science Direct). This poses a serious threat to individuals' rights to access accurate and reliable information, which is vital for informed decision-making and participation in democratic processes.

LLMs have become a tool for generating content that can deceive or mislead at an alarming scale. The advancement of AI technologies, particularly in the field of multimodal content generation, allows these models to create not only text but also images, videos, and audio that are indistinguishable from real content. This ability opens new avenues for creating convincing disinformation campaigns that can easily manipulate public opinion, spread propaganda, or destabilize social and political systems[12].

Moreover, LLMs often rely on large and diverse datasets scraped from the internet, which can contain biased, noisy, or incorrect information (Science Direct). If such flawed data is used to train a model, it can lead to the generation of biased or inaccurate content. In addition, LLMs can experience a phenomenon known as "hallucination," where the model generates content that sounds plausible but is factually incorrect. This can significantly contribute to the spread of false information, as individuals may trust content generated by seemingly authoritative sources, only to discover later that it was fabricated or misinformed.

Deepfakes and Their Impact on Disinformation

The rise of deepfakes, AI-generated videos that alter or fabricate the likenesses of individuals, has become a major concern in the fight against disinformation and the loss of credibility across digital platforms. Deepfakes are increasingly accessible due to advancements in AI technology, similar to the way LLMs have made the generation of misleading content easier. Just as LLMs enable the creation of convincing yet false written content, deepfakes empower users to generate highly realistic, fabricated visual content with minimal effort or expertise.

One of the primary concerns about deepfakes is that they have lowered the threshold for creating disinformation. Traditionally, producing convincing fake videos required specialized knowledge or expensive resources. However, with the rise of accessible AI tools, almost anyone with basic technological literacy can now create realistic deepfakes. This democratisation of technology has made it easier for malicious individuals or groups to rapidly produce and distribute misleading content, which can have a significant impact on public opinion and political discourse [23].

The widespread use of deepfake technology raises significant human rights concerns, particularly in relation to privacy, access to truthful information, and freedom of expression. Privacy violations are a major issue, as deepfakes can create fabricated videos that misrepresent individuals, showing them saying or doing things they never actually did. This unauthorised use of someone's likeness can lead to reputational damage, emotional distress, and, in some cases, legal consequences or blackmail. Additionally, deepfakes contribute to the erosion of trust in online content, making it increasingly difficult for individuals to distinguish between authentic and manipulated media. As deepfakes become more sophisticated, they undermine the right to access truthful information, especially in politically sensitive areas such as elections or public health. Deepfakes also pose a threat to freedom of expression; when individuals fear being misrepresented, they may refrain from speaking out or engaging in public discourse. This creates a chilling effect, where people hesitate to express controversial or unpopular opinions, leading to a stifling of free speech and debate[23]. Fig. 5 illustrates the human rights implications of LLMs in cybersecurity, including risks to privacy, misinformation, and deepfakes.

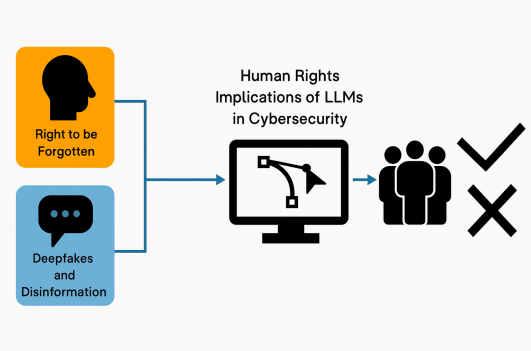

Figure 5. Human rights implications of LLMs in cybersecurity, including risks to privacy, misinformation, and deepfakes.

Conclusion

In conclusion, the integration of Large Language Models into cybersecurity presents both significant opportunities and risks. While these models offer powerful capabilities to enhance digital security through more effective threat detection, vulnerability analysis, and personal data protection, their dual-use nature raises complex challenges. The same strengths that allow LLMs to detect malicious activities can also be exploited by malicious actors to generate deceptive content, launch sophisticated cyberattacks, and manipulate users. This paradox is further complicated by concerns over privacy, transparency, and accountability, as LLMs are often trained on vast, publicly available datasets, which can include sensitive information.

The rise of disinformation and deepfakes exacerbates these concerns, as these technologies enable the rapid spread of misleading or harmful content, undermining trust in digital platforms. The ability to create highly convincing but false narratives has profound implications for human rights, particularly in relation to privacy, freedom of expression, and the right to access truthful information. As LLMs and deepfakes continue to evolve, the challenge will be to balance their potential to defend against cyber threats with the need to protect individual rights and ensure that they are not used to harm society.

To navigate these complexities, it is essential that the deployment of LLMs in cybersecurity be accompanied by ethical guidelines, robust regulatory frameworks, and ongoing oversight to mitigate the risks of misuse. The future of cybersecurity will increasingly depend on how we manage the dual-use nature of these technologies, ensuring they serve as tools for protection rather than exploitation.

References

1. Global N, Decoding the Double-Edged Sword: The Role of LLM in Cybersecurity; 2024. Accessed: 2025-11-6. https://nsfocusglobal.com/decoding-the-double-edged-sword-the-role-of-llm-in-cybersecurity/.
2. Vu K, History and Future of LLMs; 2024. https://www.datasciencecentral.com/history-and-future-of-llms/, accessed: 2025-11-07.
3. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781 2013;.
4. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) ACL; 2014. p. 1532–1543.
5. Wang Z, Chu Z, Doan TV, Ni S, Yang M, Zhang W. History, development, and principles of large language models: an introductory survey. AI and Ethics 2025;5(3):1955–1971.
6. Alaparthi S, Mishra M. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:200701127 2020;.
7. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI Technical Report 2019;(1):1–24.
8. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems 2020;33:1877–1901.
9. Patsakis C, Casino F, Lykousas N. Assessing LLMs in malicious code deobfuscation of real-world malware campaigns. Expert Systems with Applications 2024;256:124912.
10. CyberLab. Dark Web Risks: Using AI to Boost Cyber Security. CyberLab; 2024.
11. Afane K, Wei W, Mao Y, Farooq J, Chen J. Next-generation phishing: How LLM agents empower cyber attackers. In: 2024 IEEE International Conference on Big Data (BigData) IEEE; 2024. p. 2558–2567.
12. Barman D, Guo Z, Conlan O. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. Machine Learning with Applications 2024;16:100545.
13. Trigano D, Kozodoy A, The Rise of AI-Driven Cyber Attacks: How LLMs Are Reshaping the Threat Landscape (2025);.
14. Qualys, The Impact of LLMs on Cybersecurity: New Threats and Solutions; 2025. https://blog.qualys.com/product-tech/2025/02/07/the-impact-of-llms-on-cybersecurity-new-threats-and-solutions, accessed: 12 Nov 2025.
15. Not Just Memorization, Extracting Training Data From ChatGPT; 2025. https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html, accessed: 12 Nov 2025.
16. OpenAI, March 20 ChatGPT Outage; 2025. https://openai.com/index/march-20-chatgpt-outage/, accessed: 12 Nov 2025.
17. Harvard Law Review, NYT v. OpenAI — The Times's About-Face; 2024. https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/, accessed: 12 Nov 2025.
18. The Guardian, Sarah Silverman sues OpenAI and Meta for copyright infringement; 2023. https://www.theguardian.com/technology/2023/jul/10/sarah-silverman-sues-openai-meta-copyright-infringement, accessed: 12 Nov 2025.
19. Reuters, John Grisham and other top U.S. authors sue OpenAI over copyrights; 2023. https://www.reuters.com/legal/john-grisham-other-top-us-authors-sue-openai-over-copyrights-2023-09-20/, accessed: 12 Nov 2025.
20. Reuters, Getty's landmark UK lawsuit on copyright and AI set to begin; 2025. https://www.reuters.com/sustainability/boards-policy-regulation/gettys-landmark-uk-lawsuit-copyright-ai-set-begin-2025-06-09/, accessed: 12 Nov 2025.
21. Zhang D, Finckenberg-Broman P, Hoang T, Pan S, Xing Z, Staples M, et al. Right to be Forgotten in the Era of Large Language Models: Implications. Challenges, and Solutions arXiv 2023;.
22. University of Oxford, Tackling the ethical dilemma of responsibility and large language models; 2023. https://www.ox.ac.uk/news/2023-05-05-tackling-ethical-dilemma-responsibility-large-language-models, accessed: 13 Nov 2025.
23. Mitra A, Mohanty SP, Kougianos E. The world of generative ai: Deepfakes and large language models. arXiv preprint arXiv:240204373 2024;.

Author Biography

Alice Saito is an undergraduate student in the PEAK Programme (Japan in East Asia) at the University of Tokyo. Her academic work sits at the intersection of international humanitarian law, cybersecurity, and emerging technologies, with a particular focus on cyberwarfare, accountability, and the governance challenges posed by large language models. She is currently preparing a thesis on cyber operations and international humanitarian law and plans to pursue postgraduate training in computer science and cybersecurity.