Navigating the Security Challenges of LLMs: Positioning Target-Side Defenses and Identifying Research Gaps

Malte Josten[a], Matthias Schaffeld[b], René Lehmann[c], and Torben Weis[d]
Department of Distributed Systems, University of Duisburg-Essen, Duisburg, Germany
{*malte.josten, matthias.schaffeld, rene.lehmann, torben.weis*}*@uni-due.de*

Keywords: Large Language Models (LLMs), LLM Misuse, Countermeasure Criteria, Research Gaps

Abstract: Large Language Models (LLMs) have revolutionized various domains with their ability to generate human-like text, yet their misuse has introduced significant cybersecurity risks. Malicious actors exploit LLMs to create personalized phishing attacks, spread misinformation, and develop sophisticated malware, reducing the expertise and resources needed to execute such threats. The unrestricted accessibility of some LLMs further amplifies these risks, as they can circumvent existing safeguards and enhance a range of attack vectors. Current countermeasures primarily focus on restricting harmful content generation, but challenges persist, especially with unregulated or open-source LLMs. To address these limitations, a shift toward target-side detection and mitigation strategies is critical. We examine prevalent LLM-based attack methods and their implications for cybersecurity, emphasizing the need for robust defenses. We propose five core criteria—adaptability, compatibility, efficiency, effectiveness, and usability—for designing and evaluating countermeasures. An assessment of state-of-the-art solutions reveals significant gaps in adaptability and usability, highlighting areas for improvement. By addressing these challenges, we aim to guide the development of comprehensive security measures that safeguard the benefits of LLMs while mitigating their potential for misuse, ensuring digital trust and resilience in the face of evolving threats.

1 Introduction

Large Language Models (LLMs) have become increasingly ubiquitous, powerful, and accessible. However, as LLMs are tailored to specific applications, their growing capabilities also amplify cybersecurity risks. While LLMs offer a range of beneficial applications, even in the field of cybersecurity (Falade, 2023; Otieno et al., 2023), malicious actors leverage those capabilities to conduct malevolent operations targeting text-based applications and security mechanisms, and automating the generation and refinement of malicious content. For instance, LLMs' ability to mimic human writing introduces additional security concerns, as accurately detecting LLM-generated text remains challenging (Orenstrakh et al., 2024; Khalil and Er, 2023). This capacity in turn provides the foundation for a multitude of practices, including social engineering, impersonation, and even unintentional plagiarism. In this regard, malicious actors exploit LLMs to generate realistic and personalized phishing emails (Heiding et al., 2024; Roy et al., 2024) and to spread misinformation through the use of credible-sounding but false narratives (Kreps et al., 2022; Meier, 2024; Wu et al., 2024). Consequently, eroding public trust and destabilizing essential societal services (Guess et al., 2019). LLMs are also capable of memorizing and analyzing vast amounts of data, creating new avenues for cybersecurity threats. As such, even if unintended, personally identifiable information from their training data can be leaked (Li et al., 2024). Furthermore, they can be utilized to identify software vulnerabilities (Boi et al., 2024; Çetin et al., 2024), which in turn facilitates the development or refinement of malware and other malicious code (Chatzoglou et al., 2023). These capabilities position LLM as an effective means through which malicious actors can create novel threats while also being able to amplify, augment or evolve existing attack vectors. They are both readily accessible and cost-effective to misuse, reducing the barrier to entry and the resources and expertise required to circumvent detection mechanisms. Conse-

[a] https://orcid.org/0000-0003-2102-1575
[b] https://orcid.org/0000-0002-5308-7010
[c] https://orcid.org/0009-0009-5389-9543
[d] https://orcid.org/0000-0001-6594-326X

quently, LLMs are not merely another type of attack tool but a force multiplier for threat actors, capable of enhancing and diversifying a range of existing attack vectors simultaneously, potentially overwhelming conventional defense mechanisms and increasing the likelihood of successful breaches.

While some of the existing countermeasures aim to prevent LLM systems from generating harmful content, significant challenges remain to fully address the scope of potential abuse. Even if some LLM systems are successfully regulated, others, such as the open-source LLMs from Meta[1] and their modified versions, remain essentially unmoderated. They therefore provide uninhibited tools for those with malicious intent. Given the shortcomings of relying on tool-level mitigation strategies to regulate the use of LLMs, it is crucial to reinforce target-side security measures. This entails the implementation of targeted countermeasures, such as LLM-enhanced security audits and improved techniques for detecting LLM-generated content. These measures are essential, given the significant financial consequences that organisations and governments are facing (Farahbod et al., 2020; Wilner et al., 2019), as well as the substantial risks to reputation (Confente et al., 2019).

This paper examines the most prevalent methods by which LLMs are exploited to circumvent security restrictions and highlights prevalent attack vectors targeting both machines and humans. We emphasize the importance of strengthening target-side defences, examine the unique challenges introduced by LLM-based, -generated or -enhanced cyber threats and propose core criteria defining the solution space for effective countermeasures to mitigate LLM-based threats. Furthermore, we evaluate existing countermeasure research in light of these criteria to identify pertinent research gaps. Through these insights, we aim to guide the development of comprehensive security solutions that preserve the benefits of LLMs while mitigating their misuse.

2 LLM Abuse and Misuse

The preceding section highlighted the potential for LLMs to be exploited in malicious and criminal activities. A common mitigation strategy to counter these risks involves regulating inputs and outputs directly, implementing safeguard techniques and guidelines designed to limit the misuse of these models (OpenAI, 2024; Google, 2024). This approach aims to restrict the model's response capabilities by enforcing security and ethical guidelines that detect and block misuse attempts, such as jailbreaking, persona attacks, and "do-anything-now" attacks, or more broadly, prompt engineering (Yu et al., 2024).

- **Jailbreaking** allows users to manipulate the model into disregarding its built-in restrictions, prompting it to generate responses it was programmed to avoid. This technique often involves using sophisticated prompts or indirect requests that steer the model into unintended behaviors, ultimately bypassing its safeguards (Yu et al., 2024).
- **Persona Attacks** involve manipulating an LLM into adopting a specific persona that aligns with harmful behaviors or viewpoints. By shaping the AI's "personality" to mimic certain individuals or archetypes, malicious users can exploit the model's ability to assume character roles or perspectives, leading it to produce biased, manipulative, or ethically questionable responses. For instance, assigning the LLM the persona of a cybercriminal can prompt it to generate unregulated or potentially dangerous content (Yu et al., 2024).
- **Do-Anything-Now (DAN) Attacks** instruct the LLM to disregard its trained guidelines and ignore any imposed limitations, effectively bypassing its usual safeguards. As a result, the model may produce uncensored, potentially unethical, or harmful responses that it would typically be restricted from generating (Singh et al., 2025).

However, relying solely on tool-level safeguards and controls is insufficient. On the one hand, unregulated or "unhinged" LLMs can easily bypass these restrictions by not implementing them at all. Certain models, such as FraudGPT (Dutta, 2023a) and WormGPT (Dutta, 2023b), are intentionally designed without ethical guidelines and marketed for adversarial use. These models, often accessible via paid access on the dark web, enable the efficient generation of malicious content, such as realistic phishing emails or custom malware, facilitating large scale criminal activities (Falade, 2023; Mihai, 2023). On the other hand, models available on the clearnet[2] are often not developed with malicious intent, but are rather driven by principles of free speech and opposition against regulatory controls. One example is Gab.AI[3], an LLM accessible to users after only registering with an email address. It allows users to adopt various personas, including historical figures like Plato, Isaac Newton, but also malicious figures such as Adolf Hitler. As the model embodies the selected character, its outputs can be offensive or ethically questionable, reflecting

[1] https://www.llama.com/

[2] *Clearnet* refers to the publicly accessible part of the internet. It can be seen as the opposite of the *Darknet*.

[3] https://gab.ai/

the persona's traits and viewpoints.

It can be reasonably deduced that a threat actor who has undergone sufficient preparation can utilize the full potential of the latest open-source LLM at their discretion. Therefore, it is not enough to focus solely on tool control; attention must also be directed toward the refinement of existing or the development of novel target-side defense mechanisms.

2.1 Target-side security mechanisms

Unregulated or intentionally malicious LLMs allow adversaries to bypass all previously discussed model restrictions and security protocols, effectively rendering tool-based controls and safeguards ineffective. These models equip attackers with powerful tools to penetrate initial defenses, placing greater importance on target-side security mechanisms as the primary layer of protection. Consequently, emphasis must shift toward detecting and responding to LLM-supported attack vectors on the target side.

Here, we have to distinguish between two aspects: machine-targeted and human-targeted threats.

- **Machine-targeted threats** refer to attack vectors, such as malware or malicious scripts, designed to disrupt, compromise, or infiltrate hardware and software systems. Traditional detection methods, which rely heavily on automated systems, often fail to identify these threats due to the adaptability of AI-/LLM-enhanced attacks, which can continuously alter their structure and approach to evade antivirus and detection mechanisms (Chatzoglou et al., 2023). These attacks demonstrate heightened resilience, as LLMs enable the dynamic modification and refinement of malicious code, allowing it to circumvent detection mechanisms by exploiting known vulnerabilities within IT infrastructure or leveraging sophisticated obfuscation techniques. Furthermore, LLMs can assist in (re)writing spam or phishing emails by avoiding tell-tale words and phrases, effectively deceiving target filters while preserving the message's underlying intent (Josten and Weis, 2024).
- **Human-targeted threats** are designed to deceive and manipulate individuals using social engineering and persuasive language to exploit human psychology and trust. LLMs enable (highly personalized) phishing attacks by crafting realistic, contextually accurate messages that closely resemble legitimate communications, making them difficult for human targets to distinguish from genuine content (Falade, 2023; Heiding et al., 2024; Roy et al., 2024). Additionally, LLMs can generate fake news or misleading narratives that exploit current events, biases, or emotions, amplifying their effectiveness in manipulating individuals and public opinion on specific topics (Kreps et al., 2022; Meier, 2024; Wu et al., 2024).

Both the machine and human aspects are deeply intertwined, as nearly every adversarial attack supported by LLMs affects both technological infrastructure and human psychology. For instance, a phishing email crafted by an LLM might initially evade automatic detection (machine aspect) but still relies on human vigilance (human aspect) to prevent compromise. This interconnectedness underscores the need for a multi-layered defense strategy that addresses vulnerabilities targeted at both machines and humans.

3 Countermeasures

The inappropriate utilization of LLM technology as a means of generating or amplifying cyber threats necessitates a reassessment of the available solution space for the development of effective countermeasures. The factors that are driving this necessity include the high accessibility and powerful capabilities of LLMs, their dual effectiveness in targeting both machines and humans, and the cost-effectiveness of their misuse. While LLM represent a powerful new tool for threat actors, their true impact lies in their ability to amplify, augment and evolve existing malicious techniques, thereby increasing the defensive burden. Effectively, they become a force multiplier for threat actors. As a result, defenders are faced with the constant challenge of adapting their systems with each new LLM-powered attack. It is thus essential to consider the various criteria that define the solution space of effective countermeasures, including strategies of cost-efficiency to match the low barrier to entry of those attacks. It is evident that there are promising approaches to countermeasures that include the utilization of LLMs themselves as part of a defensive toolkit (Guven, 2024), the refinement of existing defenses to recognize and mitigate LLM-augmented threats and the development of reliable methods to differentiate between human- and LLM-generated content (Shimada and Kimura, 2024). However, in order to establish a resilient framework capable of countering the evolving landscape of LLM-driven cyber threats, it is necessary to address the new challenges and criteria associated with the solution space. Furthermore, in order to effectively reassess the solution space, the existing research gaps that may hinder our response to this novel threat landscape must be identified and addressed. Filling these gaps will ensure that the countermea-

sure framework is both comprehensive and adaptive to evolving LLM-driven cyber threats.

3.1 Core countermeasure criteria

When developing and deploying a new solution, a wide range of factors must be considered to ensure both quality and functionality. These include established standards and frameworks, such as ISO/IEC 25010, which outlines various software quality characteristics for both the product and its user interactions (ISO/IEC, 2011). In order to re-examine the fundamental criteria of countermeasures against LLM-based security threats, we focus exclusively on the pertinent security-related criteria of solutions. These criteria must be universally applicable, transcend specific use cases, and form an integral part of any development process for robust security solutions. Below, we discuss the key security criteria that should be prioritized in solution development.

- **Adaptability** in this context refers to the capacity to dynamically evolve and respond to the specific challenges posed by the rapidly changing threat landscape shaped by LLM-driven attacks. Static defenses are no longer sufficient, as these threats can exploit novel vulnerabilities and bypass traditional security measures. Adaptive systems must leverage real-time threat intelligence and advanced analytics to refine detection and response processes, enabling them to counter evolving LLM-driven attack vectors effectively.
  This adaptability is particularly crucial in mitigating the rapidly-evolving nature of LLM-enabled attacks. Adversaries can leverage LLMs to continuously test, refine, and optimize their methods, systematically identifying weaknesses in security systems and bypassing established controls. In order to address the continually evolving attack vectors, cybersecurity countermeasures must be capable of dynamically adapting to new threat environments with minimal latency. By continuously updating defensive algorithms and tailoring response strategies, stakeholders can proactively close security gaps and enhance their resilience against the increasingly sophisticated attacks enabled by LLMs. This ensures that defenders keep up with the pace of attackers, mitigating the risk of successful breaches and minimizing the impact of LLM-driven cyber threats.
- **Compatibility** plays an important role in building upon existing security solutions, as new approaches should be designed to extend these solutions and demonstrate a degree of interoperability. This means that novel countermeasures should be compatible with various security frameworks and able to integrate with different tools and platforms, allowing for cohesive and efficient defenses across systems. When aiming to retrofit existing mechanisms, new approaches should integrate seamlessly without disrupting other established measures and, ideally, be easily adaptable. Such retrofitting should ensure minimal modification to existing infrastructure, making it feasible to deploy updates or enhancements without requiring extensive downtime or reconfiguration, which could otherwise hinder system resilience. Given the flexibility and continually evolving nature of LLMs, this criterion becomes even more important as newly developed measures should be equally able to be integrated into existing systems.
- **Effectiveness** is crucial in countering LLM-supported attack vectors, as it directly measures how accurately and reliably a countermeasure can detect and, where possible, mitigate these threats. Choosing the right metrics to focus on depends heavily on the specific use case and should therefore be thoughtfully considered when designing and deploying security measures. For example, in defending against LLM-aided spam or phishing attacks, prioritizing high precision is essential - even at the expense of other metrics like recall - because blocking legitimate emails can seriously impact user experience and lead to unintended consequences. As language models continue to advance, this challenge grows as their increasingly sophisticated, human-like outputs make it harder to differentiate malicious content from genuine communications. Consequently, the rate of misclassifications is likely to be negatively affected (Orenstrakh et al., 2024). Including humans "in-the-loop" could, depending on the use case, help to mitigate these negative effects and elevate the classification accuracy by (semi-)manually posing as an additional security layer (Nguyen and Choo, 2021).
- **Efficiency** in defense should be optimized to be, ideally, equal to or at least close to that of attackers. Attackers often employ relatively low-cost methods to leverage LLMs in support of their strategies and can adapt flexibly to different situations. In response, defenders must work to counter these efforts. Given the attackers' adaptability and lower investment, defensive measures should not require substantial resources (such as labor, money, or time) to remain viable. This way, countermeasures stay competitive and provide defenders with a reasonable chance to mitigate relevant attack vectors. Such efficiency is especially

important for stakeholders with limited resources, like small businesses, which may lack the capacity to address the problem independently.
To illustrate the cost disparity between attacking and defending, we want to look at the example of spam email creation and detection. Attackers can generate spam emails at a cost of approximately 0.17 cents per email (Josten and Weis, 2024). In contrast, detecting potential spam emails — even with an LLM — costs defenders between 0.6 and 13 cents per email, depending on the specific model employed (Koide et al., 2024). This results in detection costs that are 3 to 75 times higher than the creation costs attackers face, and the disparity is even greater when factoring in that defenders must also examine legitimate emails. While attackers only bear the cost of creating spam, defenders are responsible for processing every incoming message to separate spam from legitimate content. Given that spam accounts for roughly 50% of global email traffic — a percentage that can fluctuate based on factors like time of year or targeted industry — defenders must handle a far greater volume than attackers (Datta et al., 2023), which further drives up detection expenses. This also may introduce additional latency as higher workloads, in combination with the complex identification of potentially LLM-generated texts, increase the overall processing time. Optimizing for low latency ensures that security measures support rather than impede productivity.

- **Usability** is another key criterion as it directly affects the effectiveness and adoption of security tools and practices. When systems are not user-friendly, users are often reluctant to engage with them, or bypass them entirely, which significantly undermines security efforts (Interaction Design Foundation - IxDF, 2016). Systems designed with intuitive interfaces and reduced complexity can empower users to participate actively in maintaining security without additional strain on time and resources (Grobler et al., 2021). In the context of LLM-supported attack vectors, the accessibility of security measures should mirror the ease of access that malicious actors have to LLMs themselves. Ensuring that defenses are as accessible and user-friendly as those used by attackers enhances both usability and effectiveness of countermeasures. This approach enables end-users to rapidly comprehend and implement those countermeasures, staying ahead in a rapidly evolving threat landscape shaped by sophisticated LLM-driven attacks. This parity in accessibility is essential to ensure that defenses keep pace with attacks in both scope and speed. Additionally, usability is closely tied to transparency and explainability. Transparent systems provide clear indicators and explanations for detected threats, making it easier for users to make informed decisions and increasing their trust in these systems (Bhatt et al., 2020). This is especially important for systems leveraging AI and machine learning, where the logic behind threat detection might otherwise be opaque (Ferrario and Loi, 2022). When dealing with texts generated by LLMs, the line between benign and malicious content can be exceedingly difficult to distinguish, making it essential for users to understand both the detection process and its outcomes. For example, if an incoming email is flagged as spam, users benefit from knowing the specific reasons - such as the presence of suspicious language patterns likely generated by an LLM or the detection of obfuscated links or code. By providing transparency into these defensive actions, security measures empower users to make informed decisions, strengthening their confidence in the system. Additionally, the complexity and sophistication of LLM-based attacks make it even more important for users to see "behind the curtain". Transparency and explainability not only enhance usability but also build trust in the defense mechanisms by helping users understand why certain content is flagged as potentially harmful. Since LLM-assisted attacks often target individuals directly (human-targeted attack vectors), this level of clarity enables users to better recognize and respond to emerging threats, making transparency a fundamental component of usable and effective security measures.

Although each criterion contributes uniquely to the effectiveness and sustainability of security measures, enhancing one often comes at the cost of another. Balancing these criteria in the fight against LLM-enhanced attack vectors is key to developing robust, user-friendly, and cost-effective cybersecurity solutions and countermeasures.

3.2 State-of-the-art countermeasures

We conducted a review on the literature, focusing on recently published works addressing target-side countermeasures against spam, phishing, and LLM-generated text. While the rapid development in this area has led to numerous preprints, this review focuses on peer-reviewed publications to ensure reliability and rigor. The review reveals a range of approaches aimed at detecting and mitigating LLM-based threats. Researchers have proposed a vari-

Publication	Adaptability	Compatibility	Effectiveness	Efficiency	Usability
(Datta et al., 2023)	○	○	●	○	○
(Gehrmann et al., 2019)	◐	●	◐	●	●
(Jamal et al., 2024)	◐	◐	●	●	○
(Josten and Weis, 2024)	○	○	◐	●	◐
(Kirchenbauer et al., 2023)	○	●	◐	◐	◐
(Koide et al., 2024)	◐	●	●	◐	○
(Krishna et al., 2023)	○	◐	●	◐	○
(Mitchell et al., 2023)	●	◐	●	◐	◐
(Shimada and Kimura, 2024)	●	◐	●	○	○
(Sun et al., 2024)	○	◐	●	○	○
(Wu et al., 2024)	◐	●	●	●	○
(Zellers et al., 2019)	◐	○	●	●	◐

Table 1: The (optimization) focus of a selection of countermeasures on the presented criteria, where ○ indicates a non-existent to small focus, ◐ indicates a moderate focus, and ● indicates a high focus on that criterion.

ety of methods, but there remains a lack of refinement in many countermeasures regarding robustness against attacks that are themselves supported by LLMs. While several techniques have proven effective in identifying malicious or LLM-generated content, these methods are often designed for specific platforms or models, limiting their adaptability across broader scenarios. Moreover, the robustness of these methods, as demonstrated in the literature, is typically assessed against the LLMs available at the time of their development. However, the question remains as to whether their robustness is compromised by the introduction of new LLMs. Additionally, it has to be explored how the resilience of methodologies can be continually assessed in order to facilitate an efficient and timely response to emerging advancements in LLMs.

A notable finding from reviewing these countermeasures (see Table 1) is that most papers have focused heavily on traditional performance metrics like precision and F1 scores to validate their effectiveness, providing a quantitative basis to evaluate these approaches. Many publications also included discussions on associated costs, underlining the resource considerations involved in such solutions - especially if their solution included the use of LLMs. However, the adaptability of these tools across diverse environments and their compatibility with different platforms received less attention, as most countermeasures were limited to a specific application or platform. Although a few papers highlighted the potential for broader applicability, discussions on interoperability and usability, including factors like explainability and latency, were notably brief. These criteria are particularly critical as they affect the accessibility and trustworthiness of the tools for non-technical users, potentially limiting the practical uptake and trust in these solutions.

While the majority of selected papers cover essential criteria and validate their proposed methods effectively, areas, such as adaptability and usability, remain underrepresented (see Figure 1). As these criteria are critical to real-world implementation as well as its longevity regarding future adaptations to novel threats and user acceptance, future research would benefit from prioritizing them, aiming for countermeasures that are not only effective but also adaptable and user-friendly. Addressing these gaps could enhance the utility of countermeasures across broader user demographics and operational environments, strengthening the defenses against increasingly sophisticated, LLM-enabled attack vectors.

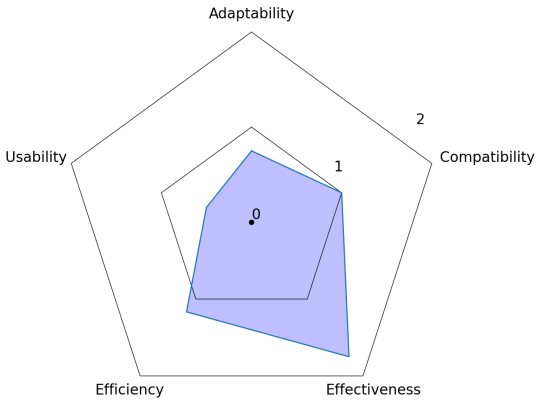

Figure 1: The mean focus level of the proposed criteria based on the selected countermeasures and scores (○: 0, ◐: 1, ●: 2) in Table 1.

4 Conclusion

The rise of LLMs has brought new cybersecurity concerns, as their misuse can threaten digital security and information integrity. These models can generate highly convincing content, making them potential tools for spreading misinformation, manipulating users, and multiplying the force of cyberattacks. As such, it is critical to preserve social trust by ensuring the credibility of content and communication, especially when faced with the challenge of deceptive, synthetic outputs created by LLMs. To mitigate these risks, both preventative and responsive measures are necessary. We explored how adversaries can exploit LLMs by bypassing existing safeguards through prompt engineering techniques or by leveraging unrestricted models. Given the difficulty of fully securing interactions with LLMs or limiting access to unregulated models, the focus must shift toward enhancing target-side detection and mitigation strategies. To guide the development and improvement of countermeasures against LLM-supported attack vectors, we proposed five key criteria: adaptability, compatibility, efficiency, effectiveness, and usability. Our evaluation of state-of-the-art countermeasures against these criteria revealed noteworthy gaps, particularly in adaptability and usability. Future research should concentrate on addressing the gaps identified in the core criteria, with particular emphasis on adaptability and usability. There is an urgent need to develop cost-effective countermeasures that can be widely deployed while ensuring their continued effectiveness. In addition, existing countermeasures must be adapted to mitigate LLM-generated attacks and remain relevant as technology advances. Concurrently, new countermeasures should be developed, tailored to emerging threats specific to LLMs. Consequently, there is a need to examine system design, so that existing infrastructures can be enhanced while facilitating the seamless integration of new security measures. Finally, continuous monitoring of the solution space and the identification of gaps is crucial as new threats and solutions evolve. As the need for sustained effort is evident, achieving a balance between maximizing the benefits of LLMs and ensuring robust security measures necessitates persistent vigilance and adaptive technology. The rapid evolution of LLMs necessitates ongoing research, innovation, and the adaptation of security protocols to address emerging threats, ensuring that the risks associated with these powerful tools are effectively managed.

REFERENCES

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA.

Boi, B., Esposito, C., and Lee, S. (2024). Smart Contract Vulnerability Detection: The Role of Large Language Model. *SIGAPP Appl. Comput. Rev.*, 24(2):19–29.

Chatzoglou, E., Karopoulos, G., Kambourakis, G., and Tsiatsikas, Z. (2023). Bypassing antivirus detection: Old-school malware, new tricks. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, ARES '23, pages 1–10.

Confente, I., Siciliano, G. G., Gaudenzi, B., and Eickhoff, M. (2019). Effects of data breaches from user-generated content: A corporate reputation analysis. *European Management Journal*, 37(4):492–504.

Datta, S., Bandyopadhyay, S., and Mondal, B. (2023). Classification of spam and ham emails with machine learning techniques for cyber security. In *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, page 1–6.

Dutta, T. S. (2023a). Fraudgpt: New black hat ai tool launched by cybercriminals. *Cyber Security News*.

Dutta, T. S. (2023b). Hackers use wormgpt to launch sophisticated cyberattacks. *Cyber Security News*.

Falade, P. V. (2023). Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pages 185–198.

Farahbod, K., Shayo, C., and Varzandeh, J. (2020). Cybersecurity indices and cybercrime annual loss and economic impacts. *Journal of Business and Behavioral Sciences*, 32(1):63–71.

Ferrario, A. and Loi, M. (2022). How explainability contributes to trust in ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1457–1466, New York, NY, USA. Association for Computing Machinery.

Gehrmann, S., Strobelt, H., and Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

Google (2024). Safety Settings | Gemini API | Google AI for Developers. https://ai.google.dev/gemini-api/docs/safety-settings, Accessed: 2024-11-07.

Grobler, M., Gaire, R., and Nepal, S. (2021). User, usage and usability: Redefining human centric cyber security. *Frontiers in Big Data*, 4.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5:eaau4586.

Guven, M. (2024). A comprehensive review of large language models in cyber security. *International Journal*

of Computational and Experimental Science and Engineering, 10(3).

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., and Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146.

Interaction Design Foundation - IxDF (2016). Three common problems in enterprise system user experience. https://www.interaction-design.org/literature/article/three-common-problems-in-enterprise-system-user-experience.

ISO/IEC (2011). Systems and software engineering —- Systems and software Quality Requirements and Evaluation (SQuaRE) —- System and software quality models. Standard, ISO/IEC.

Jamal, S., Wimmer, H., and Sarker, I. H. (2024). An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *SECURITY AND PRIVACY*, 7(5):e402.

Josten, M. and Weis, T. (2024). Investigating the Effectiveness of Bayesian Spam Filters in Detecting LLM-modified Spam Mails. 10.48550/arXiv.2408.14293.

Khalil, M. and Er, E. (2023). Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In *Learning and Collaboration Technologies*, pages 475–487. Springer Nature Switzerland.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 17061–17084.

Koide, T., Fukushi, N., Nakano, H., and Chiba, D. (2024). ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection. 10.48550/arXiv.2402.18093.

Kreps, S., McCain, R. M., and Brundage, M. (2022). All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, pages 27469–27500.

Li, T., Das, S., Lee, H.-P. H., Wang, D., Yao, B., and Zhang, Z. (2024). Human-centered privacy research in the age of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24.

Meier, R. (2024). LLM-Aided Social Media Influence Operations. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 105–112. Springer Nature Switzerland.

Mihai, I.-C. (2023). Editorial: The transformative impact of artificial intelligence on cybersecurity. *International Journal of Information Security and Cybercrime*, 12(1):9–10.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202.

Nguyen, T. N. and Choo, R. (2021). Human-in-the-loop xai-enabled vulnerability detection, investigation, and mitigation. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1210–1212.

OpenAI (2024). Hello GPT-4o | OpenAI. https://openai.com/index/hello-gpt-4o/, 2024.11.07.

Orenstrakh, M. S., Karnalim, O., Suárez, C. A., and Liut, M. (2024). Detecting LLM-Generated Text in Computing Education: Comparative Study for ChatGPT Cases. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference*, pages 121–126.

Otieno, D. O., Siami Namin, A., and Jones, K. S. (2023). The application of the bert transformer model for phishing email classification. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1303–1310.

Roy, S. S., Thota, P., Naragam, K. V., and Nilizadeh, S. (2024). From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 36–54. IEEE Computer Society.

Shimada, H. and Kimura, M. (2024). A method for distinguishing model generated text and human written text. *Journal of Advances in Information Technology*, 15:714–722.

Singh, S., Cornell, K., and Vaishnav, L. (2025). The hidden dangers of publicly accessible llms: A case study on gab ai. In *Digital Forensics and Cyber Crime*. Springer Nature Switzerland. To be published.

Sun, Y., He, J., Cui, L., Lei, S., and Lu, C.-T. (2024). Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. 10.48550/arXiv.2403.18249.

Wilner, A., Jeffery, A., Lalor, J., Matthews, K., Robinson, K., Rosolska, A., and Yorgoro, C. (2019). On the social science of ransomware: Technology, security, and society. *Comparative Strategy*, 38(4):347–370.

Wu, J., Guo, J., and Hooi, B. (2024). Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 3367–3378.

Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., and Zhang, N. (2024). Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4675–4692.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32.

Çetin, O., Ekmekcioglu, E., Arief, B., and Hernandez-Castro, J. (2024). An Empirical Evaluation of Large Language Models in Static Code Analysis for PHP Vulnerability Detection. *JUCS - Journal of Universal Computer Science*, 30(9):1163–1183.