Assignment: Literature Review

Name:

Class: Research Methods and Professional Practice April 2025 B

Professor:

The question I chose: #9 - The Impact of LLMs (Large Language Models) in a sector of your choice. Specifically, I focused on how LLMs are used in cybersecurity and the dual use aspect of both the “adversaries” and “defenders” making use of LLM’s, in a form of Cybersecurity AI Arms race.

—---

The Dual-Use Dilemma: A Literature Review on the Impact of Large Language Models in Cybersecurity

The rapid integration of Large Language Models (LLMs) represents a paradigm shift comparable to the advent of the internet itself. While their ability to generate sophisticated, human-like text has unlocked applications in numerous fields, their impact on cybersecurity is uniquely vast and contentious. The academic literature converges on a single point: LLMs are a quintessential dual-use technology (Polito and Pupillo, 2024). They have been described as a "double-edged sword" that has initiated a new technological "arms race" in the cybersecurity space (Tian, et al., 2025).

On one hand, LLMs offer defenders unprecedented tools for automating and enhancing security operations. On the other hand, they provide adversaries with powerful weapons to create and scale complex attacks. This review maps this new battleground by providing a holistic overview of the current research. It will first examine the offensive frontier, then evaluate the corresponding defensive posture, and finally, analyze the inherent flaws and research gaps that define the present state of LLMs in cybersecurity. Table 1 summarizes the dual-use capabilities of LLMs as identified across the literature referenced within this paper.

Table 1: The Duality of LLMs in Cybersecurity

Domain of Application	Offensive Applications	Defensive Applications	References

Social Engineering	Capable of generating realistic personalized phishing emails and deep fakes at scale.	Detection of phishing emails, malicious URLs, and anomalous language patterns.	(Gupta, et al., 2023; Tian, et al., 2025)
Malware	Automated generation of novel and obfuscated malware with polymorphic capabilities.	Malware analysis, classification of malware families, and reverse engineering of malicious code.	(Xu, H., et al., 2024; Hasanov, et al., 2024)
Vulnerability Management	Autonomous discovery and exploitation of software vulnerabilities, including zero-days.	Automated vulnerability scanning, code analysis, and patch generation.	(Moskal, et al., 2023; Zhang, J., et al., 2025; Infosecurity Magazine, 2024)
Autonomous Operations	Creation of autonomous agents for multi-stage interactive and adaptive attacks.	Automation of Security Operations Center (SOC) tasks like log analysis and incident response.	(Xu, J., et al., 2024; Binbeshr, et al., 2025)
Information Ecosystem	Dissemination of misinformation and content manipulation via indirect prompt injection.	Detection of misinformation campaigns and harmful content.	(Greshake, et al., 2023; Zhang, J., et al., 2025)

Starting with the offensive side, the literature strongly suggests that the most disruptive impact of LLMs has been the empowerment of offensive actors, shifting the accessibility, scale, and nature of cyberattacks. A primary theme is that LLMs significantly lower the technical barrier to entry for complex cyberattacks. Moskal, et al. (2023) argue that LLM-powered agents have effectively "killed the script kiddie," granting novice attackers capabilities once reserved for experts. This is echoed by Gupta, et al. (2023), who frame Generative AI as a technology that automates the creation of hyper-realistic phishing campaigns and polymorphic malware.

The most profound shift, however, is economic. LLM-driven automation drastically reduces the time and expertise required per attack, allowing threat actors to achieve both scale and sophistication simultaneously, a previously prohibitive combination. For instance, a highly targeted spear-phishing campaign that once required immense manual effort can now be fully automated, fundamentally altering the economics of cybercrime and forcing a strategic shift in defensive thinking (Hazell, 2023).

The literature documents a clear evolution from LLMs as simple code assistants to the core reasoning engines of autonomous agents. These agents can execute complex, multi-stage attacks with little human intervention. The AUTOATTACKER system, for example, automates "hands-on-keyboard" attacks across the MITRE ATT&CK framework, including reconnaissance and lateral movement (Xu, J., et al., 2024). This type of capability is validated by the CYBENCH framework, which benchmarked LLM agents on their ability to autonomously solve professional-level Capture the Flag (CTF) challenges (Zhang, A. K., et al., 2025).

This progression signifies the emergence of agent-based threats. The danger is no longer just that an LLM can generate an exploit; it is that an agent can reason, plan, and react to its environment to achieve a malicious objective. The architecture of agents like AUTOATTACKER mirrors the human OODA loop (observe, orient, decide, act), with the LLM serving as the core decision-making engine, not merely a script executor (Xu, J., et al., 2024). This creates a dynamic and adaptive threat. While a static script fails against an unexpected configuration, an LLM agent can self-correct after a failed command, making it far more resilient and dangerous (Xu, J., et al., 2024).

The integration of LLMs into applications has created entirely new attack surfaces. The work of Greshake, et al. (2023) on Indirect Prompt Injection (IPI) is a seminal example. IPI attacks exploit the fact that LLM-integrated applications often retrieve and process data from external, untrusted sources. An attacker can poison these data sources with hidden instructions that the LLM processes as legitimate commands, blurring the line between data and code. This can be used to exfiltrate a user's private data, manipulate the application's output to spread misinformation, or even create a computer worm that propagates the malicious prompt to other users (Greshake, et al., 2023).

The potency of IPI lies in its exploitation of user trust. Users are not conditioned to distrust the output of a seemingly benign application from a major technology company. The malicious payload is delivered by an authoritative source—the AI assistant itself. This attack vector circumvents the user's mental "firewall", representing a fundamental shift from attacking technical vulnerabilities to attacking cognitive and trust-based ones (Greshake, et al., 2023).

In response to these offensive capabilities, the literature details a growing field of defensive LLM applications designed to augment human analysts and automate tedious security tasks. A strong consensus exists on the value of LLMs in automating key functions within Security Operations Centers (SOCs). Binbeshr, et al. (2025) conceptualize this as the rise of the "Cognitive SOC," where AI addresses chronic problems like "data deluge" and "analyst burnout." LLMs excel at parsing, summarizing, and correlating the vast, unstructured datasets from logs and alerts that overwhelm human teams. By automating Level 1 triage, LLMs dramatically improve the signal-to-noise ratio, freeing expert personnel to focus on high-level strategic tasks like proactive threat hunting. The defensive utility of LLMs is therefore best understood as a force multiplier in a human-machine team, not a replacement for human expertise (Binbeshr, et al., 2025).

Beyond reactive defense, LLMs are being leveraged for proactive security measures like secure code generation, automated vulnerability detection, and software patch generation (Zhang, J., et al., 2025). However, the literature reveals a critical issue. While LLMs can be guided to produce secure code, by default they frequently generate insecure code by replicating unsafe patterns found in their training data (Xu, H., et al., 2024). This contradiction highlights a significant defensive gap. The promise of LLMs for proactive security is entirely conditional upon expert human oversight, high-quality training data, and meticulous prompting, making its practical application currently more aspirational than realized.

The arms race narrative is built upon a foundation with fundamental limitations that temper the technology's potential and represent the most significant areas for future research. A unifying theme across the literature is that LLMs are fundamentally untrustworthy. Surveys consistently identify a core set of vulnerabilities, including Prompt Injection, Data Poisoning, Privacy Leakage, and Hallucinations (Al Siam, et al., 2025; Zhang, J., et al., 2025). These are not merely software bugs but emergent properties of the LLM architecture itself. This creates a paradox where the features that make LLMs powerful, their ability to follow complex instructions, are also their greatest weaknesses. By its current design, a model cannot reliably distinguish between a developer's system prompt and a malicious instruction embedded in user data (Greshake, et al., 2023). This suggests that defenses must focus on building a secure perimeter around the model through input sanitization, output validation, and sandboxing, rather than attempting to "fix" the core model itself.

The literature also reveals a profound gap in the ability to reliably evaluate the security capabilities of LLMs. Frameworks like CYBENCH represent a significant step forward by using real-world CTF challenges and objective metrics (Zhang, A. K., et al., 2025). However, even this approach has limitations, as CTFs are often artificial in nature. This points to a deeper "observer effect" in LLM security: the act of publishing a benchmark influences the models it aims to measure. To mitigate this, CYBENCH used recent challenges, but as new LLMs are continuously trained on data scraped from the public internet, any static benchmark will inevitably be absorbed into future training sets. This means any static benchmark has a built-in expiration date. This dynamic renders one-off evaluations insufficient and underscores the call across the literature for continuous, adaptive testing methodologies, such as ongoing red-teaming exercises.

The proliferation of LLMs has ignited a technological arms race in cybersecurity. LLMs have democratized sophisticated offensive capabilities while simultaneously offering defenders powerful tools for automation and analysis. However, this competition rests on a flawed foundation. LLMs are inherently untrustworthy, susceptible to attacks that exploit their core design, and the community lacks reliable methods to evaluate these evolving risks.

Summarizing the future research directions observed across the literature, a consistent strategy is observed. The technical trajectory points toward secure-by-design LLMs that can fundamentally distinguish between trusted instructions and untrusted data (Greshake, et al., 2023). The analytical trajectory emphasizes Explainable AI (XAI) to mitigate the "black box" problem and allow human analysts to trust and verify AI-driven decisions (Binbeshr, et al., 2025). Finally, the operational trajectory converges on enhancing human-AI collaboration and developing sophisticated autonomous agents for both offense and defense (Zhang, J., et al., 2025).

Ultimately, the literature suggests the strategic advantage in the AI-driven cybersecurity landscape will not belong to those who simply develop more powerful models. Rather, it will be secured by those who can master their inherent vulnerabilities, develop robust methods for continuous evaluation, and effectively integrate these powerful yet flawed tools into resilient, human-centric, and ethically-grounded security frameworks.

References:

Al Siam, A., et al. (2025) ‘A Comprehensive Review of AI’s Current Impact and Future Prospects in Cybersecurity.’

Binbeshr, F., et al. (2025) ‘The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches’, IEEE Open Journal of the Computer Society, 6, pp. 360-375.

Greshake, K., et al. (2023) ‘Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection’, in Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23). Copenhagen, Denmark: ACM, pp. 79-91.

Gupta, M., et al. (2023) ‘From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy’, IEEE Access, 11, pp. 80218-80245.

Hasanov, I., et al. (2024) ‘Application of Large Language Models in Cybersecurity: a Systematic Literature Review’, IEEE Access, 12, pp. 176751-176771.

Hazell, J. (2023) Spear phishing with large language models. Available at: arXiv:2305.06972.

Infosecurity Magazine (2024) Google Researchers Claim First Vulnerability Found Using AI. Available at: https://www.infosecurity-magazine.com/news/google-first-vulnerability-found/ (Accessed: 14 June 2025).

Moskal, S., et al. (2023) LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing. Available at: arXiv:2310.06936.

Polito, C. and Pupillo, L. (2024) ‘Artificial Intelligence and Cybersecurity.’

Tian, S., et al. (2025) ‘Exploring the Role of Large Language Models in Cybersecurity: A Systematic Survey’, IEEE Transactions on Network Science and Engineering, XX(XX).

Xu, H., et al. (2024) ‘Large language models for cyber security: A systematic literature review’, Cybersecurity, 8(55).

Xu, J., et al. (2024) AUTOATTACKER: A Large Language Model Guided System to Implement Automatic Cyber-attacks. Available at: arXiv:2403.01038.

Zhang, A. K., et al. (2025) ‘CYBENCH: A FRAMEWORK FOR EVALUATING CYBERSECURITY CAPABILITIES AND RISKS OF LANGUAGE MODELS’, in

International Conference on Learning Representations (ICLR 2025).

Zhang, J., et al. (2025) ‘When LLMs meet cybersecurity: a systematic literature review’, Cybersecurity, 8(55).