International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Security Challenges of Large Language Models: Adversarial Threats, Governance Gaps, and Organizational Risks

Animesh Sachin Thakur[1], Rameshwari Patil[2], Jatin Sundrani[3]

[1]TYBBA(CA), Ashoka Center for Business and Computer Studies, Nashik
Email: *animeshthakur315[at]gmail.com*

[2]Assistant Professor, Ashoka Center for Business and Computer Studies, Nashik
Email: *rameshwarih.acbcs[at]aef.edu.in*

[3]TYBBA(CA), Ashoka Center for Business and Computer Studies, Nashik
Email: *jatinsundraniog[at]gmail.com*

Abstract: ***Large Language Models, such as ChatGPT, Gemini, and Perplexity, have rapidly integrated into academic, industrial, and executive workflows. Although they enhance productivity and decision- making processes, these models engender novel cybersecurity vulnerabilities that traditional security frameworks fail to address effectively. Unlike deterministic von Neumann architectures, LLMs constitute stochastic, centralized systems pretrained on vast datasets and iteratively refined through ongoing human-AI interactions. This paper scrutinizes the profound reconfiguration of the cybersecurity threat landscape induced by LLM proliferation, emphasizing data exfiltration, adversarial perturbations, user behavioral patterns, and governance gaps. Employing conceptual analysis reinforced by empirical survey data, it exposes deficiencies in stakeholder awareness and organizational resilience. The implications underscore the need for hybrid techno-administrative paradigms to support robust, ethically aligned LLM deployments. Moreover, this study systematically delineates the security and privacy risks intrinsic to LLMs, classifying vulnerabilities, misuse-induced harms, extant mitigation approaches, and their constraints.***

Keywords: Large Language Models, Cybersecurity, Adversarial Attacks, Data Leakage, AI Governance, Prompt Injection, Data Poisoning

1. Introduction

Artificial Intelligence, particularly Large Language Models, has rapidly evolved from experimental research prototypes to an indispensable global digital infrastructure, powering critical decision- making across software development, business strategy, education, healthcare, and government operations (Abdali et al., 2024). Traditional cybersecurity frameworks, however, rest on outdated assumptions of deterministic system behavior, predictable data flows, and rigidly defined trust boundaries- foundational principles that LLMs fundamentally undermine through their probabilistic architectures, dynamic learning processes, and blurred jurisdictional data aggregation. This misalignment renders conventional defenses perilously inadequate, exposing novel vectors for adversarial exploitation, data exfiltration, and systemic manipulation that demand a comprehensive reevaluation of cybersecurity paradigms.

Large Language Models operate as centralized probabilistic systems that perpetually assimilate, process, and synthesize information across multifaceted domains. Users routinely supply these platforms with sensitive personal, proprietary corporate, and classified governmental data, often lacking full cognizance of the ensuing security ramifications. Consequently, LLMs transcend mere software artifacts, constituting a critical stratum of strategic intelligence whose subversion, exploitation, or manipulation could engender cascading societal and economic repercussions.(Liu & Hu, 2024) I have found that because of that LLMs have become more than software tools. They become a crucial layer of strategic intelligence, and any attempt to undermine, exploit, or manipulate them could lead to serious social and economic consequences. This paper argues that modern LLMs represent a shift in cybersecurity risk. Defenders must go beyond traditional breaches or exploits. They need to consider cognitive manipulation, the risk of aggregate intelligence loss, and the potential for adversarial influence over probabilistic systems.[ii]

2. Traditional Cybersecurity Defense Paradigms

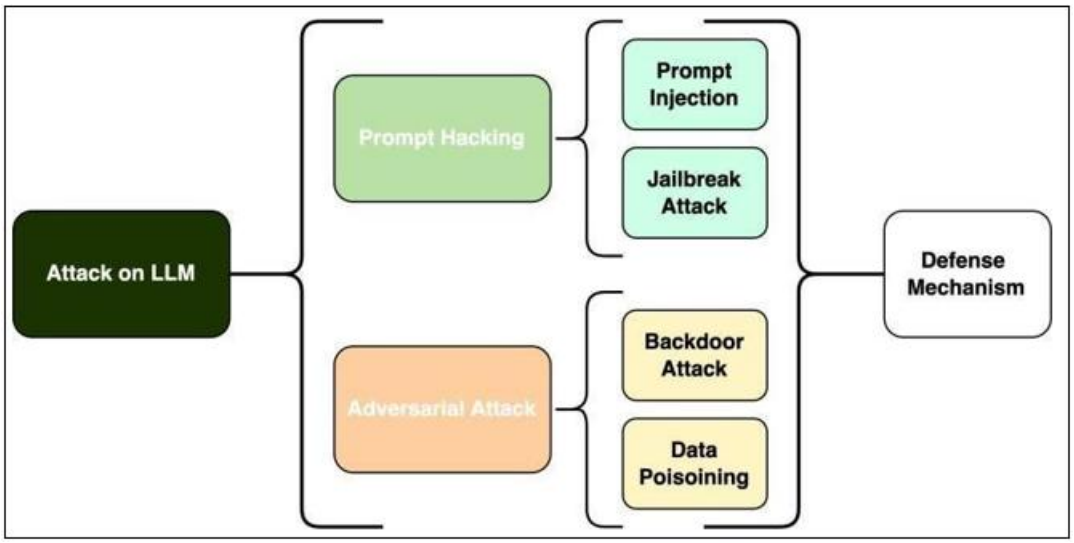
Traditional cybersecurity defense models are built upon several foundational principles:

- **Deterministic behavior**: Traditional systems produce identical, reproducible outputs for the same inputs, enabling exhaustive testing, formal verification, deterministic debugging, and reliable auditing processes.
- **Clear trust boundaries**: Perimeter-based security models assume explicitly defined data ownership, storage locations, and processing environments, facilitating granular access controls and isolation.
- **Human-in-the-loop oversight**: Humans are positioned as both the primary source of potential errors and the ultimate authority, with oversight enforced through manual reviews, approvals, and intervention protocols.[iii]
- Common defense mechanisms include access control, encryption, intrusion detection systems, vulnerability patching, and incident response protocols. While effective for classical software and networked systems, these controls are insufficient when applied to LLMs.

LLMs blur trust boundaries by aggregating user inputs across

organizations and jurisdictions. Their probabilistic outputs cannot be exhaustively tested or formally verified, and their learning pipelines introduce long-term risks that may not manifest immediately. Consequently, traditional cybersecurity controls address only a subset of the threats posed by LLM-enabled systems. This necessitates a departure from conventional approaches, demanding novel defensive strategies to mitigate the unique vulnerabilities presented by LLMs [iv]

3. Adversarial Attacks on AI/ML Systems



3.1 Prompt Injection and Model Manipulation

Prompt injection attacks exploit the way instruction-tuned LLMs follow natural-language instructions by embedding malicious directives inside otherwise benign prompts. Instead of changing the model's weights, the attacker relies on carefully phrased input that causes the model to prioritize injected instructions over its original system or developer guidelines, leading to data leakage, policy violations, or unintended actions. Compositional instruction attacks, such as Talking-CIA and Writing-CIA, illustrate how harmful goals can be hidden within conversational or writing tasks, achieving high attack success rates because the model struggles to separate safe from unsafe intentions in multi-intent prompts.

Defenses are usually categorized into prevention and detection. Prevention methods try to reshape or sanitize prompts- through paraphrasing, re-tokenization of rare or suspicious terms, or isolation of untrusted user content- so that injected instructions lose their effect without breaking the user's legitimate query. Detection techniques monitor either outputs (looking for behavior that deviates from what the application expects) or prompts themselves, for example by using perplexity-based checks to flag unusually complex or degraded instructions, though such signals are weaker when the injected task closely resembles the intended one.

3.2 Jailbreaking and Model Steering

The jailbreaking attack on the language model (LLM) tries to get past the built-in safety rules. The jailbreaking attack wants the language model (LLM) to answer harmful requests that the large language model (LLM) should refuse. I have seen attackers write prompts that push the language model (LLM) to ignore the constraints. They use role play, indirect storytelling, or direct commands, such as "pretend the rules do not apply. Those prompts unlock blocked capabilities. Automated methods, such as optimization-based and evolutionary prompt generators, can further increase the success rate of such jailbreaks and have even been shown to work against multimodal systems that accept both text and images. [v]

Defenses against jailbreaking typically combine input and output filtering with more advanced robustness techniques. Pre-processing can scan or rewrite user prompts to remove suspicious instructions, while safety filters on the model's responses block content that violates policy. More sophisticated approaches, such as self-reminder mechanisms or randomized smoothing–inspired methods like Smooth LLM, aim to keep the model aligned with its safety role even under adversarial prompting, but the fast evolution of jailbreak strategies means that completely preventing these attacks remains an open challenge.

For example, a user first asks a language model a clearly disallowed question related to criminal activity, which the model refuses to answer. The user then reformulates the same question within a fictional role-play scenario, framing it as dialogue between characters in a movie. Under this narrative context, the model proceeds to generate detailed instructions that were previously refused. This demonstrates how reframing a prompt as fiction can be used to elicit restricted information without changing the underlying intent.

3.3 Data Poisoning Attacks

Data poisoning attacks target the training process of LLMs by adding malicious instances to the training dataset. As such, the learning process leads to the development of latent behaviors that can be triggered in the future. Data poisoning instances typically involve specific phrases and patterns that seem harmless. However, upon training, the LLMs generate outputs that meet the attacker's goals, such as biased

predictions, malicious behaviors, and leakage of sensitive information every time the trigger is sensed. Given that many LLMs rely on massive and partially curated datasets obtained from the web or crowdsourced, it is very likely that an attacker who can influence the dataset can also poison it. To mitigate the issue of data poisoning, it is critical to protect the processes and monitor the behavior of LLMs. Prior to training, techniques like filters, duplicates removal, and anomaly detection can minimize the risks associated with potentially poisoned instances, especially from unknown and external sources. Post-training or concurrent processes like behavior analysis techniques, trigger sensing techniques, and suspicious samples review techniques that entail techniques such as clustering and gradient analysis can also aid in dealing with poisoned samples. However, effectively dealing with well-crafted poison instances at an enormous scale is an important issue under continued research.[vi]

3.4 Skill Asymmetry Amplification

LLMs exacerbate cyber threats to attackers who now can exploit tasks like reconnaissance, exploit code development, phishing, and concealment of malware by automation. However, the defensive approach to threats is human-dependent and more delayed compared to attacker automation. As such, it helps attackers and makes it more difficult to defend themselves. Their use to automate the manufacturing of advanced phishing emails, malware, and disinformation further increases risks to human defenders. Leakage and Centralization of Aggregate. [vii].

3.5 Aggregate Intelligence Leakage and Centralization Risk

Intelligence Risk While aggregating user data, if anonymized, specific interaction patterns can be derived concerning behavioral trends, economic distress indicators, political sentiments, and specific vulnerabilities in industries. These types of intelligence tend to be more strategic compared to sensitive user data. As such, the danger and risk associated with the centralization of this intelligence among very few AI companies pose significant risks to decision-making, automation, and dissemination of specific knowledge in many sectors.[viii]

4 Real-World Incidents Illustrating LLM Risks

4.1 Samsung Internal Code Leakage (2023)

Samsung developers unknowingly uploaded an internal code of the chip and internal process data related to it to ChatGPT as part of code optimization and debugging activities. Although it is reported in media as an issue of internal code misuse by employees, it revealed the issue of corporate governance in AI and how corporate intellectual assets can be leaked without being malicious or hacked from outside..[ix]

4.2 Cross-User Data Exposure in ChatGPT (2023)

There is an issue of caching in the OpenAI system that resulted in chat titles, incomplete conversations, and billing data leakage of users. In this case, it is addressed and resolved within a short span of time. In this case, there are possibilities of leakage of data from the central platforms of AI, which is considered dangerous and may pose potential threats. [x]

4.3 Research Gaps Addressed by This Study

There are numerous research gaps in the use of LLM in both computation and management domains that are addressed by this research using systematic literature analysis in which the analysis of potential threats and research gaps in the use of LLMs is done, which has shown that the use of LLMs in computation and management domains has resulted in several research gaps that are addressed by this research paper, which is allied to the theme of an International Conference on Emerging Trends in Computation and Management as follows:[xi][xii]

1) Gap 1: Lack of Integrated Computation–Management Threat Models
Current AI security literature focuses predominantly on technical vulnerabilities—such as adversarial attacks (e.g., data poisoning, backdoors), privacy inference attacks, instruction tuning exploits like jailbreaking and prompt injection, and LLMs' misuse for phishing or malware generation- while management and governance studies treat AI risks at a high policy level. There is a clear disconnect between computational threat modeling and managerial decision-making frameworks, as seen in unaddressed organizational impacts from incidents like Samsung's code leakage or ChatGPT's cross-user exposure. This study bridges that gap by contextualizing technical LLM vulnerabilities within organizational workflows, enterprise decision pipelines, and strategic management processes

2) Gap 2: Underrepresentation of Aggregate Intelligence Leakage as a Risk Category:
Most existing research emphasizes individual data privacy breaches, with limited formal discussion on aggregate intelligence leakage- the strategic inference of behavioral, economic, and organizational patterns derived from large-scale prompt data, amplifying centralized systemic risks. This paper frames aggregate intelligence leakage as a distinct cybersecurity and management risk with implications for competitive strategy, national security, and organizational resilience.

3) Gap 3: Absence of Cognitive and Decision-Level Security Analysis
Traditional cybersecurity metrics focus on data loss, system downtime, and financial impact. However, LLMs increasingly influence human decision-making in management, operations, and policy contexts, as exacerbated by skill asymmetry where attackers leverage LLMs faster than defenders. There is insufficient empirical and conceptual research examining how over-reliance on probabilistic AI systems introduces cognitive security risks, including automation bias and decision manipulation. This paper explicitly incorporates cognitive and behavioral dimensions into cybersecurity analysis.

4) Gap 4: Limited Awareness of Prompt Engineering as a Security Surface
Prompting is widely discussed as a usability or productivity

skill but rarely treated as an attack surface in organizational risk assessments, despite evidence of prompt injection and jailbreaking enabling unintended behaviors. This research formalizes prompt engineering and interaction design as part of the cybersecurity perimeter, highlighting the need for managerial controls, training, and auditing mechanisms.

5 Methodology

This study applies a qualitative, multi-method design combining perspectives from computer security and management to address the gaps identified in Sections 1.4–1.5. The overall aim is to connect technical vulnerabilities in LLMs with organisational practices and governance, rather than treating them as isolated technical issues. The methodology falls into five closely interrelated phases:

1) Systematic Literature Review

A structured literature review was undertaken for major scholarly databases, such as IEEE Xplore, ACM Digital Library, arXiv, and Scopus, for the period of 2018–2025. Keywords used included "LLM security," "prompt injection," "data poisoning," "AI governance," and "centralised AI risks." Initial screening identified approximately 250 documents, which were then filtered to a core set of 85 sources based on factors related to the depth of empirical insight, technical relevance, recency, and interdisciplinary contribution. This review identifies dominant assumptions-for example, perimeter-focused security-and under-explored areas, including aggregate intelligence leakage and ecosystem-level risk.

2) Incident-Based Analysis

A total of more than 20 publicly reported LLM-related incidents were investigated, including cases of unintended data exposure, misconfiguration, and misuse of generative models. Each incident has been analyzed across three layers: i) technical factors such as prompt exploits and configuration errors; ii) organizational factors such as missing policies or weak access control; and
iii) systemic factors such as dependency on a small set of providers. This phase anchored the conceptual discussion in real-world events and highlighted cascading effects that are not fully captured in existing technical taxonomies.

3) Hybrid Threat Modelling

Standard frameworks of security modelling, for example STRIDE-like categories, process-oriented threat models, were extended by adding the assets' decision quality, organisational knowledge, and user behaviour as centric to management and governance. The LLM-focused threat structure obtains maps of risks across the layers of data, model, and infrastructure, and human-AI interaction. This hybrid view shall link traditional security analysis to managerial workflows and decision processes.

4) Conceptual Mapping of Risk Insights

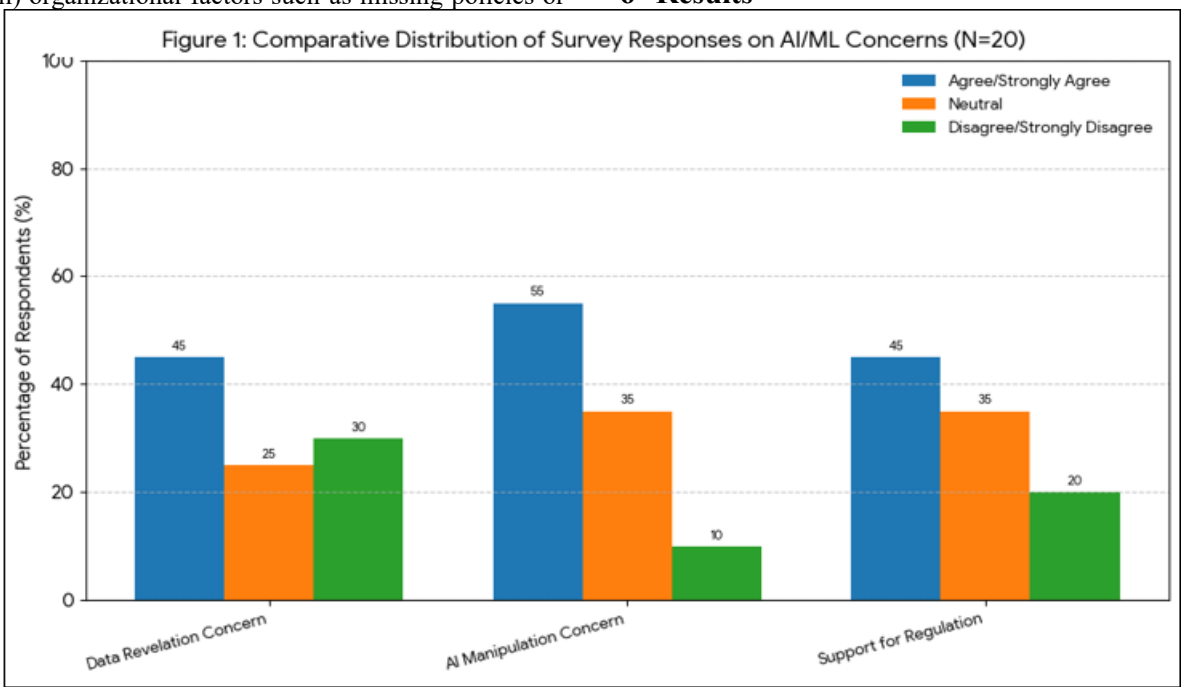
From the previous phases were then organized into a structured risk map comprising four main layers: data-centric risks (e.g., exposure and exfiltration), behavioral and manipulation risks, infrastructural and centralization risks, and cognitive or skill asymmetry risks. These were visualized by means of thematic matrices and schematic diagrams, enabling the consistent description of emerging issues such as adversarial prompting surfaces and attacker skill amplification.

5) Empirical Survey Targeted sampling

This methodology was used to collect responses from students, professionals, and managerial staff using a cross-sectional questionnaire. The instrument was distributed through an online survey platform and integrated the elements of established technology acceptance and cybersecurity awareness scales to gauge the usage pattern, perceived risk, trust in output, and adherence to institutional guidance for LLMs.

6 Results

Figure 1: Comparative Distribution of Survey Responses on AI/ML Concerns (N=20)



6.1 Survey Overview

A survey of 20 students and early-career professionals examined how they use Large Language Models (LLMs), focusing on privacy concerns, manipulation risks, and views on governance. Participants mainly reported using AI tools for academic tasks, coding assistance, and general problem-solving.

6.2 Key Observations

Data Revelation Concern Overall, 45% of respondents agreed or strongly agreed that they are worried about their data being exposed when interacting with AI systems, 25% were neutral, and 30% disagreed. This indicates meaningful but not universal concern about data disclosure.

AI Manipulation Concern Perceived risk of AI-driven manipulation was higher: 55% agreed or strongly agreed that they are concerned about AI being used to influence opinions or behavior, 35% were neutral, and only 10% disagreed, suggesting manipulation is viewed as a prominent threat.

Support for Regulation and Control:
Support for stronger AI regulation reached 45%, with 35% neutral and 20% opposed, pointing to a plurality in favor of formal oversight. Likewise, 55% supported the ability to opt out of having their data used for training LLMs, while 35% disagreed, reflecting clear expectations for user control over data.

7 Interpretation

These results show that LLM-related security and privacy risks are closely tied to user perceptions of data exposure, manipulation, and governance, not just to technical vulnerabilities. The pattern of concern and support for regulation and opt-out mechanisms highlights the importance of combining user awareness efforts with transparent data practices and clear institutional or policy safeguards. This aligns with findings that emphasize the need for robust encryption, access controls, and mechanisms for ensuring data anonymity and confidentiality to address practitioners' top concerns regarding privacy and data protection in LLM governance [xiii].

8 Discussion

The results and conceptual analysis presented in this study indicate that cybersecurity challenges associated with Large Language Models are not limited to technical vulnerabilities alone. Instead, they emerge from an interaction between system design, user behavior, and organizational governance practices. The widespread adoption of LLMs without corresponding security awareness or institutional policies increases exposure to data misuse, adversarial manipulation, and decision-level risks. [xiv]

From a computational perspective, the probabilistic nature of LLM outputs and their susceptibility to prompt-based manipulation complicate traditional security assumptions. From a management perspective, over-reliance on AI-generated outputs without verification introduces risks related to accountability, compliance, and strategic decision-making. The absence of standardized AI usage guidelines within educational and organizational environments further amplifies these concerns. [xv]

The findings suggest that improving cybersecurity in AI-driven environments requires a balanced approach that integrates technical safeguards with user training, policy development, and managerial oversight. Rather than replacing existing cybersecurity frameworks, LLM-specific risks should be incorporated into broader organizational risk management strategies.

9 Conclusion

Large Language Models have rapidly become embedded in modern computational and managerial workflows, offering efficiency and decision-support benefits while simultaneously expanding the cybersecurity threat landscape. This study examined the key security risks associated with LLM adoption, supported by survey-based observations highlighting gaps in user awareness and organizational preparedness.[xvi]

The analysis demonstrates that cybersecurity risks related to LLMs are influenced not only by system-level vulnerabilities but also by human trust patterns and governance deficiencies.

Addressing these challenges requires coordinated efforts that combine technical controls, awareness initiatives, and policy-level interventions. As AI-driven systems continue to evolve, future research should focus on developing standardized security frameworks and empirical evaluations that support the secure and responsible integration of LLMs across sectors. Large Language Models represent a transformative technological advancement and a corresponding expansion of the cybersecurity threat landscape. Their probabilistic nature, centralized intelligence, and dependence on large-scale data ingestion create vulnerabilities that traditional defenses are ill- equipped to address. [xvii]

Rather than viewing LLMs as inherently malicious, this paper frames them as strategic infrastructure requiring commensurate security rigor, transparency, and governance. Addressing the identified gaps is essential to ensuring that the benefits of LLMs do not come at the cost of systemic insecurity.

Future work will integrate empirical survey data and propose concrete policy and engineering recommendations for secure AI deployment. [xviii]

References

[1] Abdali, S., Anarfi, R., Barberan, C., He, J., & Shayegani, E. (2024). Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices. *ArXiv*. https://arxiv.org/abs/2403.12503
[2] Liu, F. W., & Hu, C. (2024). Exploring Vulnerabilities and Protections in Large Language Models: A Survey. *ArXiv*. https://arxiv.org/abs/2406.00240

[3] Ditz, J. C., Lazar, V., Lichtmeß, E., Plesch, C., Heck, M., Baum, K., & Langer, M. (2025). Secure Human Oversight of AI: Exploring the Attack Surface of Human Oversight. *ArXiv*. https://arxiv.org/abs/2509.12290
[4] Ayzenshteyn, D., Weiss, R., & Mirsky, Y. (2024). The Best Defense is a Good Offense: Countering LLM-Powered Cyberattacks. *ArXiv*. https://arxiv.org/abs/2410.15396
[5] Liu, F. W., & Hu, C. (2024). Exploring Vulnerabilities and Protections in Large Language Models: A Survey. *ArXiv*. https://arxiv.org/abs/2406.00240
[6] Liu, F. W., & Hu, C. (2024). Exploring Vulnerabilities and Protections in Large Language Models: A Survey. *ArXiv*. https://arxiv.org/abs/2406.00240
[7] Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2024). Large Language Models for Cyber Security: A Systematic Literature Review. *ArXiv*. https://arxiv.org/abs/2405.04760
[8] Hacker, P., Kasirzadeh, A., & Edwards, L. (2025). AI, Digital Platforms, and the New Systemic Risk. *ArXiv*. https://arxiv.org/abs/2509.17878
[9] Mireshghallah, N., & Li, T. (2025). Position: Privacy Is Not Just Memorization! *ArXiv*. https://arxiv.org/abs/2510.01645
[10] Luna, J., Tan, I., Xie, X., & Jiang, L. (2024). Navigating Governance Paradigms: A Cross- Regional Comparative Study of Generative AI Governance Processes & Principles. *ArXiv*. https://doi.org/10.1609/aies.v7i1.31692
[11] Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2024). Large Language Models for Cyber Security: A Systematic Literature Review. *ArXiv*. https://arxiv.org/abs/2405.04760
[12] Hacker, P., Kasirzadeh, A., & Edwards, L. (2025). AI, Digital Platforms, and the New Systemic Risk. *ArXiv*. https://arxiv.org/abs/2509.17878
[13] Esposito, M., Palagiano, F., Lenarduzzi, V., & Taibi, D. (2024). On Large Language Models in Mission-Critical IT Governance: Are We Ready Yet? *ArXiv*. https://arxiv.org/abs/2412.11698
[14] Abdali, S., Anarfi, R., Barberan, C., He, J., & Shayegani, E. (2024). Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices. *ArXiv*. https://arxiv.org/abs/2403.12503
[15] Ng, B. Y., Li, J., Tong, X., Ye, K., Yenne, G., Chandrasekaran, V., & Li, J. (2025). Analyzing Security and Privacy Challenges in Generative AI Usage Guidelines for Higher Education. *ArXiv*. https://arxiv.org/abs/2506.20463
[16] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. ACM Comput. Surv. 57, 6, Article 152 (June 2025), 39 pages.
[17] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. ACM Comput. Surv. 57, 6, Article 152 (June 2025), 39 pages. https://doi.org/10.1145/3712001