Privacy and Security Challenges in Large Language Models

Vishal Rathod*, Seyedsina Nabavirazavi*, Samira Zad*, Sundararaja Sitharama Iyengar*
*Florida International University

2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC) | 979-8-3315-0769-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CCWC62904.2025.10903912

Abstract—**Large Language Models (LLMs) are at the forefront of artificial intelligence advancements, demonstrating exceptional capabilities in natural language understanding and generation across diverse domains such as healthcare, finance, and customer service. However, their deployment introduces substantial security and privacy risks, including prompt injection, data leakage, and unauthorized data disclosures. These vulnerabilities highlight the need for robust frameworks to safeguard sensitive data and prevent misuse. This paper provides a comprehensive analysis of the security and privacy challenges in LLMs, examines existing mitigation strategies such as intelligent LLM firewalls, differential privacy, and OWASP-based security principles, and discusses future directions for ethical and secure LLM deployment. By addressing these challenges in detail, we identify gaps in current practices and propose a roadmap for the secure and responsible deployment of LLMs in high-stakes applications. Our findings underscore the importance of tailored security frameworks and privacy-preserving techniques to ensure the ethical and reliable use of LLMs in sensitive environments. Additionally, this paper emphasizes the significance of a human-in-the-loop (HITL) approach to ensure accountability and accuracy, particularly in critical domains. The discussion extends to emerging technologies such as retrieval-augmented generation (RAG) and adaptive threat detection systems, which hold promise for enhancing the security and ethical deployment of LLMs.**

Keywords—**Artificial intelligence, Natural language processing, Large language models (LLM), Privacy, OWASP, Data Protection, AI Ethics, Firewall, Threat Modeling, Data Leakage, Ethical Bias Mitigation, Federated Learning, Healthcare AI, Human-in-the-Loop (HITL), Adaptive Security Frameworks, Privacy-Preserving Computation**

I. Introduction

Large Language Models (LLMs) such as Gemini, GPT and LLaMA, have reformed artificial intelligence (AI) applications across various domains including customer service [1], healthcare, finance, programming [2], and education. The ability of LLMs to understand and generate human-like text is a breakthrough in natural language processing (NLP) and automated systems, which is applied to different applications. However, the rapid deployment and widespread adoption of LLMs entail substantial security and privacy risks. LLMs is trained on vast amounts of data which are susceptible to various threats, including adversarial attacks, data leakage, and prompt injection vulnerabilities [3]. This can lead to unauthorized data disclosure or model manipulation .

The security risks in LLMs are further intensified by ethical and regulatory concerns. This includes issues such as bias in LLM response, lack of transparency in data handling, and compliance with privacy laws such as the General Data Protection Regulation (GDPR) highlight the need for robust security frameworks tailored to LLM environments. Traditional approaches to AI security are often inadequate in addressing the unique challenges presented by LLMs, where threats can range from direct attacks on the model's architecture to indirect misuse of model-generated content in real-world applications.

This paper aims to address critical gaps by examining vulnerabilities in LLMs, evaluating existing mitigation strategies, and discussing future directions for LLM's secure and ethical deployment. Our contributions include a comprehensive analysis of the risks specific to LLMs, a review of privacy-preserving methods and frameworks, and recommendations for adapting security practices to manage LLM-specific threats.

The unprecedented adoption of LLMs is also accompanied by concerns regarding their integration with third-party applications and plugins, which can lead to supply chain vulnerabilities and increase security risks [4]. Furthermore, since LLMs are used in decision-critical scenarios, such as autonomous systems and predictive analytics, their reliability and interpretability become increasingly vital to promote trust between stakeholders [5].

II. Related Work

A. Security Challenges in Large Language Models

The adoption of LLMs has introduced unique security challenges due to their large-scale data-driven training processes and application in sensitive domains. A significant concern is their vulnerability to adversarial attacks, where maliciously crafted inputs exploit weaknesses in the model's response logic, potentially leading to harmful or unintended outcomes. For instance, adversarial inputs may manipulate prompts to cause LLMs to reveal confidential information or generate biased content. These vulnerabilities are particularly problematic in applications like customer service or healthcare, where LLMs may inadvertently share sensitive information or provide inaccurate guidance.

Data leakage is another critical issue. LLMs trained on extensive datasets, which may include proprietary or sensitive information, can unintentionally memorize and reproduce fragments of these datasets, thereby exposing private information during use.

Kshetri (2023) highlights these risks, emphasizing the role of LLMs in enabling cybercrime through social engineering attacks such as phishing and impersonation. The advanced language generation capabilities of LLMs make it easier for

979-8-3315-0769-5/25/$31.00 ©2025 IEEE.

attackers to create persuasive and authentic-looking phishing messages or impersonate individuals and organizations for malicious purposes. Their work demonstrated that social engineering attacks, exploit the realistic language generation capabilities of LLMs, with very high success rates in certain scenarios. [6].

In addition to security risks, Wu et al. (2024) examine the ethical implications of LLM-generated content. They highlight concerns about models reinforcing biases inherent in their training data, which can result in discriminatory outputs or ethical dilemmas. Coupled with user overreliance on LLMs, these issues may lead to data leaks, misinformation, or unethical decision-making in automated systems [7].

Furthermore, Mozes et al. (2023) explore the use of LLMs for illicit purposes and their associated security concerns. It categorizes risks into three domains: threats, prevention measures, and vulnerabilities. The threats include the misuse of LLMs for fraud, misinformation, malware generation, and social engineering. Author also discusses about preventive measures focus on techniques like reinforcement learning from human feedback (RLHF), content filtering, and watermarking. The study highlights LLMs' susceptibility to adversarial attacks and data poisoning, emphasizing that current safety measures are often imperfect. [8].

B. OWASP Top 10 Security Risks

The Open Web Application Security Project (OWASP) has adapted its Top 10 list to address the unique security risks posed by LLM applications. These risks include prompt injection, insecure output handling, model theft, and sensitive data exposure, arising from the text-based and context-sensitive nature of LLM interactions.

Fasha et al. (2024) propose using intelligent agents to monitor and mitigate these vulnerabilities in real time, focusing on prompt injection attacks. These agents leverage rule-based filtering, semantic analysis, and pattern matching to neutralize adversarial inputs before reaching the model. This approach aligns with OWASP's emphasis on adapting traditional security measures to the unstructured nature of LLM-generated content [9].

OWASP also emphasizes secure output handling. Since LLM-generated responses can include unexpected or unsafe content, sanitizing outputs through encoding, filtering, and manual review is essential. These steps mitigate risks such as Cross-Site Scripting (XSS), code injection, and privacy violations [10].

C. Privacy and Compliance Concerns

LLM applications, especially in sectors such as healthcare and finance, raise significant privacy concerns due to their handling of sensitive information. Regulations like the General Data Protection Regulation (GDPR) in Europe impose strict requirements on data handling, retention, and transparency. Violations of these regulations can result in substantial financial penalties and reputational damage.

Rahman (2023) underscores the importance of privacy-preserving techniques, particularly in healthcare. Methods like differential privacy, which adds statistical noise to data, and federated learning, which allows distributed training without centralizing sensitive information, are key strategies to mitigate privacy risks [11].

Vu and Hoang (2024) highlight the importance of aligning LLMs with privacy policies and regulatory standards like GDPR. Their approach employs transfer learning to ensure compliance, enabling LLMs to analyze privacy policies and align with regulatory requirements, thereby minimizing risks associated with data retention and processing [12].

Derner and Batistič (2023) discuss how LLM platforms like ChatGPT raise unique concerns regarding plugin ecosystems, emphasizing the need for secure sandboxing and rigorous third-party vetting to ensure compliance and data safety [13].

III. Security Threats in LLMs

The deployment of LLMs in sensitive applications introduces several security vulnerabilities. In the following, we discuss key threats, including prompt injection, training data poisoning, insecure output handling, and denial-of-service (DoS) attacks.

A. Prompt Injection

Prompt injection exploits the dynamic nature of LLM-generated responses. Attackers craft malicious prompts to manipulate the model's behavior, potentially leading to the unauthorized disclosure of information. For example, a benign-looking prompt embedded with malicious instructions may cause the LLM to reveal confidential details or perform unintended actions [9].

Direct prompt injection targets the system prompt itself, overriding default instructions, while indirect prompt injection involves embedding malicious instructions in external content, such as web pages, that the LLM processes [10].

B. Training Data Poisoning

Training data poisoning involves introducing malicious data into the training dataset to influence the model's behavior. Such attacks can embed biases, unethical responses, or erroneous outputs in the LLM. This threat is particularly concerning for models trained on publicly available or minimally curated datasets, as harmful content can be difficult to detect and remove [14], [15].

C. Insecure Output Handling

LLM-generated responses, if not properly sanitized, can result in downstream vulnerabilities like XSS or remote code execution. Since LLM outputs are often dynamic and unpredictable, rigorous sanitization is essential before using these responses in applications [10].

D. Model Denial of Service (DoS)

Due to the computational intensity of LLMs, malicious users can exploit resource limitations through DoS attacks. Overloading the model with excessive or resource-intensive queries can degrade its performance, resulting in service outages or increased costs for cloud-based deployments [16].

IV. Privacy Concerns in LLMs

LLMs inherently process vast amounts of data, raising privacy concerns across various domains. This section addresses key concerns, including data leakage, user privacy risks, and ethical implications.

A. Data Leakage

LLMs trained on large datasets may inadvertently reveal proprietary or sensitive information, posing significant privacy risks in domains such as healthcare and finance [7]. Differential privacy offers a practical solution by obscuring individual data points in training datasets [11] [17].

B. User Privacy in Automated Systems

Automated systems powered by LLMs, such as chatbots, must adhere to privacy regulations. Ensuring data minimization and implementing privacy-compliant retention policies can mitigate risks [12], [18].

C. Ethical Bias and Discrimination

LLMs trained on biased datasets may perpetuate societal biases. Bias audits and adversarial debiasing techniques are crucial to mitigate these risks [7].

D. Over-Reliance on LLM Outputs

Over-reliance on LLM-generated content without oversight poses additional privacy risks, as users may place undue trust in the accuracy and confidentiality of responses. In critical sectors, reliance on LLMs without human validation can lead to misinformation or data misuse, especially if the model's outputs are taken as definitive answers. To prevent these issues, a "human-in-the-loop" approach, where human oversight is integrated into the model's response evaluation, can help maintain accountability and accuracy in LLM applications [10], [19].

V. Mitigation Strategies

The security and privacy risks associated with LLM requires robust mitigation strategies that are specifically cater to the unique vulnerabilities of these systems. In this section, we discussed key strategies to mitigate risk.

A. LLM Firewalls

An LLM firewall is a specialized security mechanism that filters inputs and outputs in real-time to prevent harmful content from being processed or generated by the model. Huang et al. (2024) stated an LLM firewall model that combines rule-based filtering with semantic analysis to detect potential threats. This firewall uses a multi-engine detection strategy, where inputs are analyzed for harmful instructions, embedded code, or other indicators of prompt injection attacks. A multi-engine detection strategy means that the firewall uses multiple detection techniques or engines simultaneously to analyze inputs. Each engine is optimized to catch different types of threats, making the detection process more comprehensive and robust [16]. LLM firewalls are designed to recognize and block content patterns associated with prompt injection, data leakage, and other adversarial manipulations.

B. Differential Privacy and Federated Learning

Differential privacy is a technique that adds statistical noise to data, making it difficult to trace individual data points back to their source, thereby preserving user anonymity. This technique is especially beneficial in applications where LLMs handle sensitive data, such as healthcare and finance. Rahman (2023) highlights differential privacy as an effective way to prevent data leakage by ensuring that personal data is less likely to be memorized and reproduced by the model [11], [20].

Federated learning, on the other hand, allows LLMs to train on decentralized datasets without directly accessing raw data. In federated learning, data remains on the user's device, and only aggregated model updates are shared, preserving user privacy. This approach is particularly useful for LLMs used in distributed systems, where centralizing data could pose significant privacy risks. Both differential privacy and federated learning contribute to building privacy-preserving LLM architectures [21].

C. OWASP-Based Security Guidelines

The Open Web Application Security Project (OWASP) provides guidelines for managing security in LLMs. Key recommendations from the OWASP Top 10 for LLMs include secure output handling, access control, and sandboxed plugin design [9].

- **Output Handling**: To prevent issues like Cross-Site Scripting (XSS) and code injection, OWASP recommends sanitizing and encoding all LLM outputs, especially if they are directly displayed or executed in downstream applications.
- **Access Control**: Limiting access to LLM functions based on user roles can prevent unauthorized individuals from manipulating sensitive model parameters or outputs. Access control mechanisms reduce the risk of unauthorized prompt injection or data exfiltration.
- **Sandboxed Plugins**: For LLMs that interact with external plugins, OWASP suggests isolating plugins in sandboxed environments. This minimizes the risk of security breaches due to vulnerabilities in third-party extensions.

By following these OWASP guidelines, LLM applications can strengthen their defenses against common security risks, ensuring that generated content and interactions remain secure.

D. Ethical Bias Mitigation

LLMs are susceptible to propagating biases present in their training data, leading to ethical issues when these models produce discriminatory or biased content [22]. Wu et al. (2024) suggest that ethical bias audits, adversarial debiasing, and regular evaluation of model outputs are essential for maintaining fairness in LLM applications [7].

Adversarial debiasing is a technique where LLMs are trained to minimize biased outputs actively. This can be achieved by incorporating counterfactual data examples that challenge the model's biases during training. Additionally, frequent audits of model responses can identify emerging biases, allowing developers to address these issues through targeted re-training or fine-tuning. Ensuring ethical outputs in LLMs is crucial in sectors like law and healthcare, where biased responses could have significant real-world implications.

E. Summary of Security Threats and Privacy Concerns and its Mitigating

Table I provides an overview of the primary security threats and privacy concerns in LLMs, along with suggested mitigation strategies.

VI. Case Study: LLM Security in Healthcare

To highlight the critical importance of security and privacy in LLM deployment, this case study explores LLMs' applications within the healthcare sector. In this field, the handling of sensitive patient information mandates strict compliance with data privacy and security standards. LLMs are employed for various tasks, including responding to patient queries, assisting in diagnoses, and analyzing medical data. Given these responsibilities, ensuring robust security and privacy measures is of paramount importance.

To illustrate the importance of security and privacy in LLM deployments, this case study examines the use of LLMs in healthcare, a domain where handling sensitive patient information requires strict compliance with data privacy and security standards. LLMs in healthcare are utilized for tasks such as patient query responses, diagnosis assistance, and medical data analysis, making security and privacy a top priority.

A. Application Scenario

In healthcare sector, LLMs can be used as a virtual assistants to enhance patient care. The virtual assistant can respond to patient inquiries, provide assistance in diagnoses, and provide preliminary recommendations based on patient's symptoms. For instance, an LLM-based chatbot can help patients in managing chronic conditions by answering questions about medication schedules, dietary modifications, or symptom administration.

However, this virtual assistant requires strict adherence to data privacy and security regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe, ensuring the protection of sensitive health data is important, as failing to do so could result in severe legal consequences, damage to reputations, and loss of patient trust.

B. Privacy and Security Analysis

Healthcare applications of LLMs face a high risk of data leakage, as these models may unintentionally reveal personal health information if appropriate measure are not in place. Threat modeling frameworks like STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privileges) can help identify potential vulnerabilities in LLM deployments [14]. For example:

- **Information Disclosure**: Data leakage due to model responses containing traces of sensitive patient information [25].
- **Denial of Service (DoS)**: High demand on LLM resources during peak hours may degrade service quality, impacting patient access to healthcare advice.
- **Spoofing and Tampering**: Unauthorized access to the system could result in tampering with responses or unauthorized data access.

By using following recommendation, healthcare providers can systematically measure security risks and design mitigation strategies that ensure the safety of patient information.

C. Mitigation Recommendations

To secure LLMs in healthcare, several strategies should be implemented:

- **Differential Privacy**: Adding noise to the training data ensures that patient records are not directly memorized by the model, reducing the risk of data leakage. [11], [26].
- **Retrieval-Augmented Generation (RAG)**: By incorporating RAG, healthcare LLMs can retrieve only the necessary, non-sensitive information from external databases, ensuring privacy while answering queries [12], [24].
- **LLM Firewalls and Secure Output Handling**: Implementing an LLM firewall for real-time monitoring of inputs and outputs can help prevent inadvertent disclosures and secure interactions. Secure output handling also ensures that LLM-generated responses do not inadvertently reveal private information [16].
- **Human-in-the-Loop (HITL)**: In sensitive domains like healthcare, a HITL approach ensures that healthcare professionals review LLM-generated responses before they reach patients, maintaining accuracy and accountability [19].

These strategies collectively contribute to creating a secure environment for deploying LLMs in healthcare, protecting patient data while allowing efficient and informative interactions.

VII. Discussion and Future Directions

The adoption of Large Language Models (LLMs) in sensitive domains underscores the need for robust security and privacy measures. While current mitigation strategies address many immediate vulnerabilities, the evolving landscape of LLM applications presents new challenges that require continuous innovation in security frameworks and ethical considerations. In this section, we discuss anticipated threats,

TABLE I
SUMMARY OF SECURITY THREATS AND PRIVACY CONCERNS IN LLMS AND THEIR MITIGATION STRATEGIES

Threat/Concern	Description	Mitigation Strategy
Prompt Injection	Attackers craft input prompts to manipulate LLM responses, potentially causing unauthorized actions or data disclosure.	Input validation, strict prompt filtering, and employing intelligent LLM firewalls for real-time monitoring [9], [10], [23].
Training Data Poisoning	Adversaries introduce malicious data into training sets, embedding biases or harmful behaviors that compromise model integrity. Emerging concerns also point to ”data poisoning at scale,” where adversaries target open datasets used in training to inject malicious patterns across diverse domains, increasing the difficulty of detection [5].	Data validation, adversarial training, periodic audits of training data to identify and remove malicious inputs, and improving dataset curation standards [14], [23].
Insecure Output Handling	Unfiltered LLM outputs can lead to Cross-Site Scripting (XSS) or other code injection attacks when transferred downstream.	Output sanitization, encoding, and secure handling practices following OWASP guidelines to prevent misuse of generated content [10].
Model Denial of Service (DoS)	High computational demands of LLMs can be exploited by malicious users submitting excessive requests, leading to service slowdowns.	Rate limiting, priority queuing, anomaly detection, and resource allocation policies to manage high-volume queries [16].
Data Leakage	LLMs may inadvertently reveal sensitive information from training data in response to certain user prompts.	Implement differential privacy to protect sensitive data points and minimize the risk of data leaks in responses [11].
User Privacy in Automated Systems	Privacy risks arise when LLMs handle personal data, especially in compliance-heavy fields like healthcare.	Privacy-compliant data handling protocols, anonymization, and data retention policies; regular privacy impact assessments [12].
Retrieval-Augmented Generation (RAG) for Privacy	RAG integrates LLMs with external knowledge bases to minimize the use of sensitive training data. By retrieving specific, domain-relevant information at runtime, RAG reduces the model’s dependency on memorized content, mitigating data leakage risks. Particularly effective in healthcare for retrieving anonymized medical guidelines [12].	Use retrieval-augmented generation techniques to ensure that sensitive or private data is not stored or used in training. Employ anonymized and domain-specific external knowledge bases [18], [24].
Ethical Bias and Discrimination	LLMs trained on biased datasets may unintentionally reinforce societal biases, resulting in discriminatory outputs.	Bias audits, adversarial debiasing, and ethical guidelines to identify and mitigate biased or unfair responses [7].
Over-Reliance on LLM Outputs	Over-reliance on LLMs without human oversight can lead to misinformation or data misuse.	Employ a ”human-in-the-loop” approach to review and validate LLM-generated outputs, especially in critical applications [10].

recommendations for secure LLM deployment, and future research opportunities.

A. Anticipated Threats and Emerging Risks

As LLM technology advances, the sophistication and scale of potential threats grow exponentially, presenting new challenges for secure and ethical deployment. One of the most concerning trends is the emergence of adversarial attacks, which are becoming increasingly complex and harder to detect. Advanced techniques, such as prompt injection and training data poisoning, enable attackers to manipulate LLMs in ways that evade traditional security measures [14]. These attacks not only compromise the integrity of LLM responses but also pose significant risks to applications in sensitive domains like healthcare and finance. Furthermore, as LLMs integrate with external systems and plugins, supply chain vulnerabilities become a critical point of failure. Exploiting third-party dependencies, attackers can infiltrate and compromise LLM-driven applications, emphasizing the urgent need for end-to-end security frameworks.

Another pressing issue is model theft, where attackers replicate or modify a deployed model for unauthorized use. This threat is particularly alarming in high-value applications, such as proprietary financial analytics or healthcare systems, where intellectual property and sensitive functionalities are at stake. To address this, future research must focus on developing resilient encryption techniques and robust authentication mechanisms to safeguard model integrity and prevent unauthorized replication. Additionally, the integration of AI Security Posture Management (AI-SPM) frameworks can provide continuous monitoring and adaptive defense mechanisms, ensuring comprehensive protection for LLMs against these sophisticated threats.

B. Recommendations for Secure LLM Deployment

To address these challenges, LLM deployments must follow best practices tailored to secure their unique architecture and application environment:

- **Proactive Monitoring and Anomaly Detection**: Implementing continuous monitoring systems with irregularity detection capabilities can help identify suspicious activities in real-time. Machine learning-based threat detection algorithms that adapt to evolving usage patterns are valuable for distinguishing legitimate queries from potential threats. Furthermore, implementing self-supervised learning mechanisms for anomaly detection within LLM architectures can enhance proactive threat monitoring. This includes training models to identify unusual patterns or unexpected behaviors in real-time [5], [20].
- **Human-in-the-Loop (HITL) Integration**: Especially in high-stakes applications, a HITL approach ensures that model-generated outputs are reviewed and validated by

human experts. This layer of oversight can prevent the unintended consequences of biased, inaccurate, or harmful content reaching end users [12].

- **Regular Bias Audits and Ethical Evaluations**: With ethical and societal impact considerations, it is recommended that organizations conduct periodic bias audits. Additionally, training practices should include ethical evaluations to reduce the risk of discriminatory or harmful output [7], [27].
- **Secure Data Handling Protocols**: As LLMs are increasingly deployed in regulated industries, privacy compliance is essential. Techniques such as differential privacy, federated learning, and secure output handling should be implemented to reduce the risks of data leakage and unauthorized access.

These best practices provide a foundation for deploying LLMs responsibly and securely across diverse application areas. In addition to technical measures, regulatory bodies should be involved in establishing industry standards and guidelines for LLM security and privacy, particularly in sectors where LLM outputs have a direct impact on individual welfare.

C. Research Opportunities

There are several promising directions for future research to address emerging challenges in LLM security and privacy:

- **Adaptive Security Frameworks for LLMs**: Developing adaptive security frameworks that dynamically adjust to new threat patterns can significantly enhance the robustness of LLMs. These frameworks could leverage advanced machine learning models for real-time threat detection and response, enabling systems to proactively identify and mitigate novel attack vectors [28].
- **Privacy-Preserving Computation and Federated LLM Training**: Further exploration into privacy-preserving techniques, such as federated training of LLMs, can allow models to benefit from large datasets without compromising user privacy. Research in secure multi-party computation and homomorphic encryption for LLM training is particularly promising, as these techniques ensure data confidentiality while maintaining model performance [11], [20], [29], [30].
- **Improved Ethical and Fairness Metrics**: Current bias evaluation metrics for LLMs are limited in scope and often fail to account for the nuanced ethical considerations inherent in diverse applications. Future research should focus on developing comprehensive fairness metrics that address biases not only at a syntactic level but also within semantic and contextual dimensions. These metrics should align with real-world deployment scenarios to ensure equitable outcomes [31], [32].
- **Responsible AI Development and Deployment Standards**: Establishing clear standards for responsible AI development is critical for fostering trust and accountability in LLM deployments. Transparent model documentation, ethical guidelines, and rigorous oversight mechanisms are essential for achieving these objectives. Collaborative efforts between academia, industry, and regulatory agencies will play a pivotal role in defining and enforcing these standards [27], [33].

Advances in these areas will ensure that LLMs continue to provide societal benefits while mitigating potential harms. By addressing these research opportunities, stakeholders can further enhance the ethical, secure, and effective deployment of LLMs across various critical domains.

VIII. Conclusion

Our study focuses on the security and privacy risks of Large Language Models (LLMs) and how can these challenges be tackled. LLMs have brought significant advancements in understanding and generating natural language, revolutionizing many fields. However, as we use them more widely, we also face major security and privacy issues like prompt injection attacks, data leaks, and ethical problems like bias and discrimination. The paper comprehensively covers these problems and suggests some strong solutions, such as LLM firewalls, differential privacy, federated learning, and following OWASP guidelines.

By addressing these challenges, this study adds valuable insights to ongoing research aimed at building robust, secure, and ethically sound AI systems. This study can be very useful to fill gaps in current security systems by highlighting specific vulnerabilities in LLMs, which will offer practical solutions for stakeholders to effectively mitigate risks. Also, it stresses the approach that combines technical innovation with ethical and regulatory considerations.

This paper also anticipates emerging threats, which will require continuous innovation in security and privacy techniques. Advances like adaptive threat detection and privacy-preserving computation can help protect LLMs from future risks. Collaboration among researchers, industry, and policymakers is crucial to establish strong standards for ethical and secure LLM use. As LLMs become more integrated into critical areas like healthcare, finance, and education, ensuring their potential aligns with accountability and transparency is essential. By adopting these strategies and staying proactive, organizations can leverage LLMs' capabilities while building trust and ensuring responsible use in sensitive environments.

Finally, the study highlights the ongoing need for continuous auditing and collaboration between academia, industry, and regulatory bodies. Such efforts are vital to developing universal standards for LLM transparency, ethical accountability, and security enforcement, especially as these models are used in more sensitive and complex areas [13]. Striking this balance is key to unlocking the full benefits of LLMs while minimizing unintended consequences.

IX. Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0264. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] M. Esmaeili, M. Ahmadi, M. D. Ismaeil, S. Mirzaei, and J. Canales Verdial, “Advancements in ai-driven customer service,” in *2024 IEEE World AI IoT Congress (AIIoT)*, 2024, pp. 1–5.

[2] B. Nadimi and H. Zheng, “A multi-expert large language model architecture for verilog code generation,” *arXiv preprint arXiv:2404.08029*, 2024.

[3] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on large language model (llm) security and privacy: The good, the bad, and the ugly,” *High-Confidence Computing*, vol. 4, no. 2, p. 100211, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.hcc.2024.100211

[4] U. Iqbal, T. Kohno, and F. Roesner, “Llm platform security: Applying a systematic evaluation framework to openai’s chatgpt plugins,” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 611–623, Oct. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AIES/article/view/31664

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682*, 2022. [Online]. Available: https://arxiv.org/abs/2206.07682

[6] N. Kshetri, “Cybercrime and privacy threats of large language models (llms),” *IT Professional*, vol. 25, no. 3, pp. 9–13, 2023.

[7] X. Wu, R. Duan, and J. Ni, “Unveiling security, privacy, and ethical concerns of ChatGPT,” *Journal of Information and Intelligence*, vol. 2, no. 2, pp. 102–115, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949715923000707

[8] M. Mozes, X. He, B. Kleinberg, and L. D. Griffin, “Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities,” 2023. [Online]. Available: https://arxiv.org/abs/2308.12833

[9] M. Fasha, F. A. Rub, N. Matar, B. Sowan, M. Al Khaldy, and H. Barham, “Mitigating the OWASP top 10 for large language models applications using intelligent agents,” in *2024 2nd International Conference on Cyber Resilience (ICCR)*, 2024, pp. 1–9.

[10] “OWASP Top 10 for Large Language Model Applications,” https://owasp.org/www-project-top-10-for-large-language-model-applications/, [Accessed: 2024-11-08].

[11] M. A. Rahman, “A survey on security and privacy of multimodal LLMs – connected healthcare perspective,” in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1807–1812.

[12] T.-H.-G. Vu and X.-B. Hoang, “User privacy risk analysis within website privacy policies,” in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2024, pp. 1–6.

[13] E. Derner and K. Batistič, “Beyond the safeguards: Exploring the security risks of chatgpt,” 2023. [Online]. Available: https://arxiv.org/abs/2305.08005

[14] L. Ruhländer, E. Popp, M. Stylidou, S. Khan, and D. Svetinovic, “On the security and privacy implications of large language models: In-depth threat analysis,” in *2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics*, 2024, pp. 543–550.

[15] H. E. Oskouie, C. Chance, C. Huang, M. Capetz, E. Eyeson, and M. Sarrafzadeh, “Leveraging large language models and topic modeling for toxicity classification,” *arXiv preprint arXiv:2411.17876*, 2024.

[16] T. Huang, L. You, N. Cai, and T. Huang, “Large language model firewall for aigc protection with intelligent detection policy,” in *2024 2nd International Conference On Mobile Internet, Cloud Computing and Information Security (MICCIS)*, 2024, pp. 247–252.

[17] H. Li, Y. Chen, J. Luo, J. Wang, H. Peng, Y. Kang, X. Zhang, Q. Hu, C. Chan, Z. Xu, B. Hooi, and Y. Song, “Privacy in large language models: Attacks, defenses and future directions,” 2024. [Online]. Available: https://arxiv.org/abs/2310.10383

[18] S. Wang, T. Zhu, B. Liu, M. Ding, X. Guo, D. Ye, W. Zhou, and P. S. Yu, “Unique security and privacy threats of large language model: A comprehensive survey,” 2024. [Online]. Available: https://arxiv.org/abs/2406.07973

[19] A. S. Inamdar and S. Eswaran, “A comprehensive review of security and privacy issues in large language models,” *SSRN Electronic Journal*, 2024.

[20] G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, “Privacy-preserving techniques in generative ai and large language models: A narrative review,” *Information*, vol. 15, no. 11, 2024. [Online]. Available: https://www.mdpi.com/2078-2489/15/11/697

[21] H. Kim, M. Song, S. H. Na, S. Shin, and K. Lee, “When llms go online: The emerging threat of web-enabled llms,” 2024. [Online]. Available: https://arxiv.org/abs/2410.14569

[22] I. H. Sarker, “Llm potentiality and awareness: a position paper from the perspective of trustworthy and responsible ai modeling,” *Discover Artificial Intelligence*, vol. 4, no. 1, May 2024.

[23] B. C. Das, M. H. Amini, and Y. Wu, “Security and privacy challenges of large language models: A survey,” 2024. [Online]. Available: https://arxiv.org/abs/2402.00888

[24] M. Fatehkia, J. K. Lucas, and S. Chawla, “T-rag: Lessons from the llm trenches,” 2024. [Online]. Available: https://arxiv.org/abs/2402.07483

[25] J. Haltaufderheide and R. Ranisch, “The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms),” *npj digital medicine*, vol. 7, no. 1, Jul 2024.

[26] T. U. Islam, R. Ghasemi, and N. Mohammed, “Privacy-preserving federated learning model for healthcare data,” in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 2022, pp. 0281–0287.

[27] M. Chakraborti, B. J. Prestoza, N. Vincent, and S. Frey, “Responsible ai in open ecosystems: Reconciling innovation with risk assessment and disclosure,” 2024. [Online]. Available: https://arxiv.org/abs/2409.19104

[28] J. Wen, V. Hebbar, C. Larson, A. Bhatt, A. Radhakrishnan, M. Sharma, H. Sleight, S. Feng, H. He, E. Perez, B. Shlegeris, and A. Khan, “Adaptive deployment of untrusted llms reduces distributed threats,” 2024. [Online]. Available: https://arxiv.org/abs/2411.17693

[29] J. Zheng, H. Zhang, L. Wang, W. Qiu, H. Zheng, and Z. Zheng, “Safely learning with private data: A federated learning framework for large language model,” 2024. [Online]. Available: https://arxiv.org/abs/2406.14898

[30] R. Ye, W. Wang, J. Chai, D. Li, Z. Li, Y. Xu, Y. Du, Y. Wang, and S. Chen, “Openfedllm: Training large language models on decentralized private data via federated learning,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6137–6147. [Online]. Available: https://doi.org/10.1145/3637528.3671582

[31] N. B. Brown, “Enhancing trust in llms: Algorithms for comparing and interpreting llms,” 2024. [Online]. Available: https://arxiv.org/abs/2406.01943

[32] I. M. Serouis and F. Sèdes, “Exploring Large Language Models for Bias Mitigation and Fairness,” in *IJCAI 2024 Workshop AIGOV*, ser. IJCAI 2024 Workshop AIGOV, Jeju Island, South Korea, Aug. 2024, a workshop which aims to delve into the critical aspects of AI governance with a specific focus on the contribution of Large Language Models (LLMs) in shaping ethical and responsible AI practices. [Online]. Available: https://hal.science/hal-04667517

[33] O. Friha, M. Amine Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine, “Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness,” *IEEE Open Journal of the Communications Society*, vol. 5, pp. 5799–5856, 2024.