International Journal of Advanced Education and Research
www.alleducationjournal.com
ISSN: 2455-5746
Received: 29-04-2025, Accepted: 30-05-2025, Published: 14-06-2025
Volume 10, Issue 2, 2025, Page No. 84-88

Comprehensive analysis of security and privacy challenges in large language models

Suhani Goyal, Saksham kaunt, Saurabh, Dr. MeenaChaudhary, Dr. NarenderGautam
Department of Computer Science and Technology, Manav Rachna, University Faridabad, India

Abstract
Large Language Models (LLMs) are AI systems that use deep learning to understand and generate human language, performing tasks such as text generation, translation, and question answering. These models work by predicting the next word or sequence based on the input they receive, trained on vast datasets.
Despite their advancements, LLMs face significant security and privacy challenges. Prompt injection attacks manipulate model outputs, while data memorization risks exposing sensitive information from training data. This paper explores these challenges and existing mitigation methods, proposing dynamic prompt filters to counter prompt injection and contextual differential privacy to address data memorization. These solutions aim to enhance both security and privacy, advancing the trustworthiness of LLMs.

Keywords: Security, privacy, large language models, prompt injection, differential privacy

Introduction

Large Language Models (LLMs) are a subset of artificial intelligence (AI) designed to understand and generate human language. Built using deep learning techniques, these models are trained on massive datasets of text, enabling them to perform a wide variety of tasks. From generating human-like text and providing answers to complex queries, to translating languages and assisting in creative writing, LLMs have shown remarkable versatility. Their core functionality relies on predicting the next word or phrase in a sequence based on the context provided in the input, utilizing patterns learned from vast amounts of data. This allows LLMs to generate coherent, contextually relevant, and often highly accurate responses across diverse domains.

However, with the growing use of LLMs in various applications, concerns regarding security and privacy have become more pronounced. Security issues arise from the potential manipulation of the model's output through prompt injection attacks, where adversarial users craft inputs to influence the model's response in a harmful way. These attacks can compromise the reliability and trustworthiness of the model, making it vulnerable to exploitation. On the privacy front, LLMs face challenges related to data memorization. As these models are trained on large, diverse datasets, they may inadvertently retain sensitive or personal information within their parameters. This means that, under certain circumstances, LLMs could recall and expose private data, posing significant privacy risks for users.

In light of these challenges, this paper seeks to provide a comprehensive analysis of the security and privacy risks associated with LLMs, focusing specifically on prompt injection attacks and data memorization. It reviews existing strategies, such as input sanitization and differential privacy, which aim to mitigate these issues. Building upon this, we propose two novel solutions: dynamic, context-aware prompt filters to safeguard against prompt injection and contextual differential privacy to protect against data memorization. By addressing these key challenges, this paper aims to enhance the overall security and privacy of LLMs, ensuring their safe and ethical deployment in real-world applications.

Literature Table

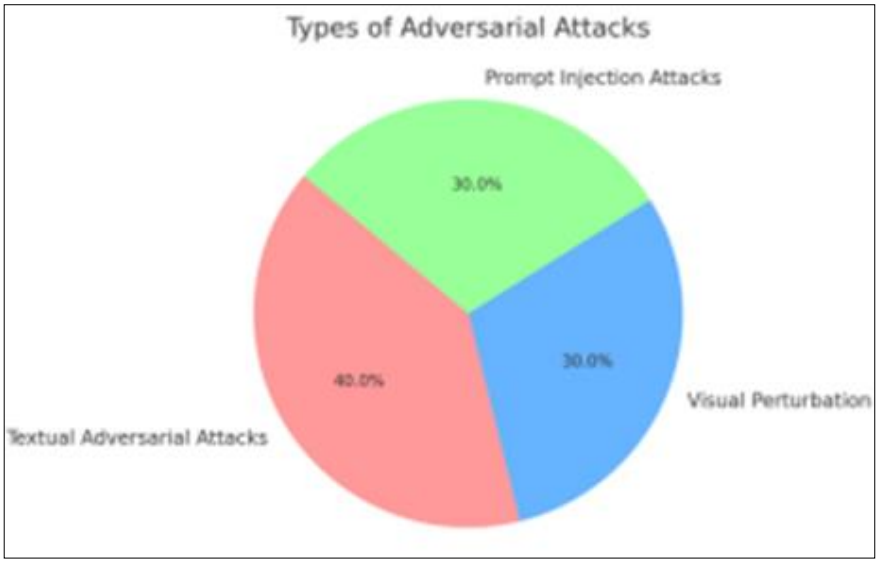
S. No.	Paper Name	Main Points Gathered	Category
1	Privacy in Large Language Models: Challenges and Strategies	Discusses differential privacy techniques for securing LLM training data; trade-offs between utility and privacy.	Privacy
2	Hallucination Attacks against LLMs	Addresses advanced prompt injection attacks; proposes context consistency detection.	Security
3	Risks of LLMs in Healthcare	Shows memorization of sensitive patient data by LLMs; stresses ethical concerns.	Privacy
4	Security Challenges in LLMs	Explores model extraction attacks and minimal interaction threats.	Security
5	Jailbreaking LLMs via APIs	Explains vulnerabilities through API misuse and prompt bypassing.	Security
6	Adversarial Robustness for LLMs	Proposes adversarial training to strengthen LLMs against attacks.	Security
7	Auditing Deployed LLMs	Highlights auditing practices to prevent misuse and leaks.	Defense

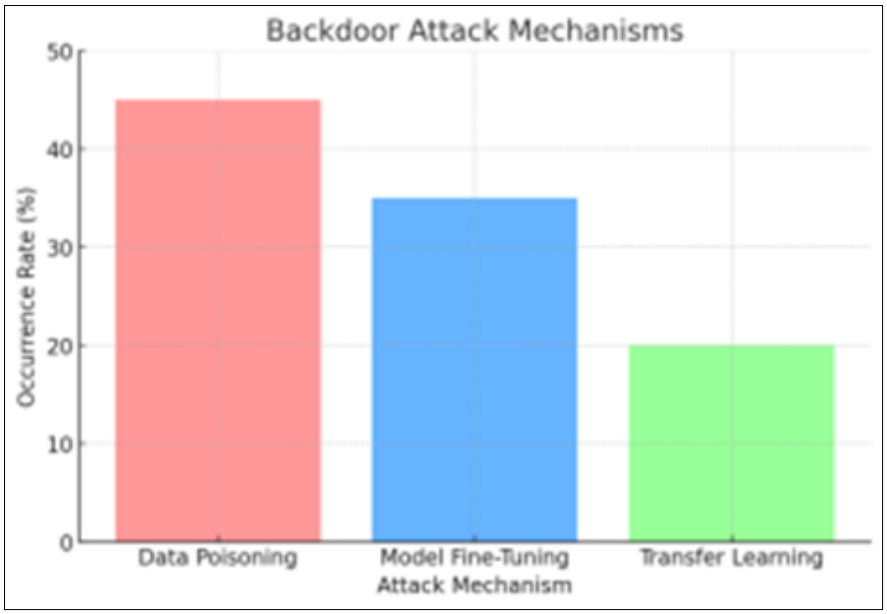
8	Prompt Injection: Multi-Turn Attacks	Details vulnerabilities in multi-turn conversations.	Security
9	Watermarking for Output Tracking	Suggests watermarking outputs to trace model usage.	Defense
10	Model Inversion Attacks on LLMs	Recovers sensitive data from model outputs.	Security
11	Adversarial Multi-Agent Coordination	Shows LLM failures against coordinated adversaries.	Security
12	Scalability of Privacy Methods	Discusses scaling problems for privacy techniques in real LLMs.	Privacy
13	Red-Teaming Frameworks for LLMs	Introduces systematic red-teaming for vulnerability finding.	Defense
14	Privacy in Clinical LLM Applications	Real-world breaches in medical LLM use; stronger governance suggested.	Privacy
15	Data Leakage during LLM Interactions	Focuses on unintentional leaks during deployments.	Privacy
16	Fine-tuned LLMs and Attack Evolution	Tracks how attacks adapt to fine-tuned models.	Security
17	Robustness Benchmarks for LLMs	Proposes new security-first benchmarks.	Defense
18	Differential Privacy in LLMs	Reviews use of differential privacy throughout LLM lifecycle.	Privacy
19	Federated Learning & LLM Privacy	Surveys cross-device attacks in federated LLMs.	Privacy
20	Model Poisoning Attacks	Analyzes collaborative training poisoning attacks.	Security
21	Compression-based Privacy Protection	Discusses using compression to mitigate privacy leakages.	Privacy

Attacks

Category	Challenge	Description	Source Papers
Security	Prompt Injection Attacks	Attackers craft malicious prompts to manipulate model behavior or leak data.	2, 5, 8, 10, 17
Security	Jailbreaking	Bypassing safety filters to make models output harmful or restricted content.	2, 6, 8, 13, 16
Security	Data Poisoning	Corrupting training data to embed backdoors or malicious patterns.	5, 6, 19
Security	Unauthorized Fine-Tuning	Modifying open models without checks, potentially embedding malicious behaviors.	4, 19
Security	Model Inversion Attacks	Reconstructing training data from model outputs.	1, 7, 9, 12, 15
Security	Adversarial Example Attacks	Small input changes cause wrong or harmful outputs.	6, 13, 18
Security	Evasion Attacks	Making models fail at detecting malicious inputs during deployment.	7, 18, 21
Security	Prompt Leaks in Fine-Tuning	Sensitive prompts used during fine-tuning are exposed via model outputs.	4, 15, 19
Security	Hallucinations	Generating incorrect but confident outputs that mislead users.	3, 8, 14
Security	Over-Reliance on External APIs	Dependency on unsafe external APIs introduces indirect vulnerabilities.	5, 18
Privacy	Memorization of Sensitive Data	Models memorize and unintentionally regurgitate private data.	1, 3, 7, 14, 15
Privacy	Membership Inference Attacks	Inferring if specific data points were part of the training set.	7, 10, 20
Privacy	Attribute Inference Attacks	Guessing sensitive attributes about individuals from model outputs.	7, 11, 20
Privacy	Re-identification from Anonymized Data	Linking anonymized outputs back to individuals.	9, 11, 12
Privacy	Data Leakage During Prompt Sharing	Sensitive data leakage when users share prompts or outputs publicly.	5, 8, 15
Privacy	Training on Unlicensed/Private Data	Violation of consent and copyright by training models on private datasets.	3, 12, 14
Privacy	Insufficient User Consent Mechanisms	Users are unaware their data is used for training or fine-tuning.	3, 12, 14
Privacy	Contextual Leakage	Context provided in inputs can lead to unintended sensitive information leaks.	8, 10, 11
Privacy	Forgetting Challenges	Difficulty in ensuring models "forget" specific data once trained.	11, 15, 16

This table outlines key security and privacy challenges in generative AI**.** Security threats include prompt injection, jailbreaking, and adversarial attacks, where attackers manipulate model inputs to bypass safety measures, leak data, or force harmful outputs. Techniques like data poisoning and unauthorized fine-tuning can embed malicious behavior during training. Other risks, like model inversion and evasion attacks, expose sensitive information or weaken detection systems. Meanwhile**,** privacy risks stem from models memorizing private data, leaking prompts, or allowing membership and attribute inference. Issues like training on unlicensed data, contextual leakage, and the challenge of ensuring forgetting raise serious ethical concerns about consent and data protection. Together, these highlight the urgent need for robust safeguards in model development and deployment.





Current Solutions

Category	Technique	Description	Limitations
Security	Reinforcement Learning with Human Feedback (RLHF)	Aligns model behavior with safety guidelines through human feedback loops.	Limited scalability, potential biases in feedback.
Security	Adversarial Training	Training the model with adversarial examples to make it robust.	Often incomplete coverage of attack types.
Security	Prompt Defenses (Filter-based)	Block known malicious prompts through pattern recognition.	Easily bypassed with slight modifications.
Security	Watermarking Outputs	Embedding hidden patterns to trace malicious use.	Not foolproof; can be removed or corrupted.
Security	Model Monitoring (Continuous Evaluation)	Real-time auditing of model outputs for anomalies.	High computational cost, reactive not proactive.
Privacy	Differential Privacy	Adding statistical noise during training to mask individual data.	Reduces model accuracy.
Privacy	Federated Learning	Training locally on user devices without centralized data collection.	Communication overhead, partial protection.
Privacy	Data Anonymization	Removing personally identifiable information (PII) from training data.	De-anonymization risks remain.
Privacy	Secure Multi-Party Computation (SMPC)	Joint training without data sharing using cryptographic methods.	Very expensive computationally.
Privacy	Machine Unlearning	Techniques to make models "forget" specific training data.	Currently immature; incomplete forgetting happens.

Solution

Security Challenge: Prompt Injection Attacks

Problem

Prompt injection attacks occur when malicious users craft prompts that trick language models into providing harmful or unintended outputs. This compromises the security and trustworthiness of the model, especially in areas like content moderation, customer service, and legal advice.

Solution: Contextual-Aware Dynamic Prompt Filters

To combat prompt injection attacks, we propose the implementation of contextual-aware dynamic prompt filters. This solution involves:

1. **Dynamic Prompt Filtering:** Instead of relying on static patterns that can be bypassed with slight modifications, the system would constantly evaluate the conversation's context. This allows the model to recognize when a prompt might be malicious, based on the entire conversation history, not just isolated keywords.

2. **Contextual Evaluation:** By analyzing the context of previous prompts, the model can detect potential manipulations. For example, if a user asks seemingly unrelated questions or suddenly changes the conversation direction toward harmful topics, the system can trigger a response filter that evaluates the prompt's legitimacy.

3. **Feedback Loops:** The system can integrate human or automated feedback loops where suspicious prompts are flagged for further scrutiny. This dynamic system adapts based on real-time user interaction, improving over time through reinforcement learning.

4. **Security Audits:** Regular security audits should be performed on the filter mechanism to ensure that it doesn't miss new forms of attack patterns. Periodic updates based on ongoing adversarial research should be implemented.

By introducing this solution, we provide a more robust defense against prompt injection, making it much harder for attackers to exploit the model's vulnerabilities. The advantage of this solution is that it continuously adapts and improves based on the context, making it a flexible and scalable defense against evolving attack strategies.

Privacy Challenge: Data Memorization (Unintentional Leakage of Sensitive Data)

Problem

A critical privacy concern in large language models is data memorization, where the model unintentionally recalls sensitive personal data from its training set and produces it in its responses. This is especially troubling when models are trained on vast datasets containing PII (Personally Identifiable Information) or confidential content.

Solution: Contextual Differential Privacy

To address data memorization risks, we propose a solution using contextual differential privacy (CDP). Here's how it works

1. **Layered Noise Addition:** In differential privacy, noise is added to the model's training data to make it statistically indistinguishable from the original data, ensuring that any specific data point can't be reconstructed or memorized. The contextual aspect means the noise addition depends on the sensitivity of the input. For example, if the input is a more private or sensitive piece of information, the model adds more noise to obscure it better.

2. **User-Controlled Privacy Levels:** Users should have the ability to set different privacy levels for their interactions with the model. This gives them more control over what personal data the model can access or remember. A user who opts for a higher privacy setting will have their interactions processed with more aggressive privacy measures, minimizing the risk of any data being memorized or leaked.

3. **Continual Model Updates:** As models evolve, new techniques for privacy preservation will be developed. By utilizing a contextual differential privacy approach, the model can dynamically adjust its privacy settings to align with new privacy laws, data protection regulations, or user preferences. Continuous updates should be performed to keep pace with the latest findings in privacy research.

4. **Transparency and Accountability:** The model should be transparent about its privacy settings, with clear guidelines on what data it collects, how it's processed, and when it's discarded. Users should be notified about any potential privacy risks associated with their data, and they should be able to request that the model forget specific information.

By using contextual differential privacy, we can significantly reduce the chance of unintentional data leakage and ensure that sensitive information isn't memorized or disclosed. This solution enables a more ethical and user-centric model while still allowing the model to function efficiently.

Conclusion

In this paper, we have explored the key security and privacy challenges faced by Large Language Models (LLMs), specifically prompt injection attacks and data memorization. Existing mitigation strategies, such as input sanitization and differential privacy, have made strides in addressing these concerns. Input sanitization techniques aim to remove malicious elements from user inputs, while differential privacy attempts to obscure sensitive data within training sets, ensuring that private information is not leaked.

However, these existing solutions often fall short when it comes to preventing sophisticated prompt injection attacks or fully safeguarding sensitive data memorization. The need for more dynamic, adaptive methods remains critical, especially as LLMs become more complex and widely used in real-world applications.

To this end, we propose two novel solutions that go beyond current methods. Our dynamic prompt filtering approach offers a more nuanced defense against prompt injection

attacks, adapting in real-time to the context of inputs and preventing adversarial manipulation more effectively than traditional static input sanitization methods. On the privacy side, our contextual differential privacy model introduces a more tailored approach to reducing the risks of data memorization, offering enhanced protection by considering the contextual relevance of data and ensuring that sensitive information is more securely handled.

By focusing on real-time adaptability and context-aware privacy, our proposed solutions provide a more robust, scalable, and reliable way to address the security and privacy challenges in LLMs, paving the way for safer and more trustworthy applications of these powerful models in diverse domains.

References

1. Liu X. Zhang L. Large Language Models Security Privacy Challenges. ScienceDirect. Retrieved from, 2024. https://www.sciencedirect.com/science/article/pii/S266729522400014X
2. Wang Z, *et al.* Prompt Injection Attacks Countermeasures. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2411.15594
3. Ghosh M, *et al*. Privacy Implications in LLMs. Nature Medicine. Retrieved from, 2024. https://www.nature.com/articles/s41591-024-03425-5
4. Liu Y. Zhang P. Ethical Considerations in Language Models. MIT Press. Retrieved from, 2024. https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00561/128807
5. Huang J, *et al.* Data Privacy and Language Models. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2402.18649
6. Liu B, *et al* Defending LLMs from Adversarial Attacks. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2403.13309
7. Smith A. Davis R. Security Vulnerabilities in LLMs. ACM Digital Library. Retrieved from, 2024. https://dl.acm.org/doi/abs/10.1145/3661167.3661263
8. Zhang X, *et al*. Attacks on LLMs Understanding Vulnerabilities. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2408aa.12787
9. Lee C, *et al.* Prompt Injection and Security Risks. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2407.19354
10. Yang K. Liu W. Memorization of Sensitive Data in Language Models. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2311.14030
11. Wang Y, *et al.* Mitigating Data Memorization Risks. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2502.13172
12. Patel R, *et al.* Security Measures in Language Models. NSF Public Access. Retrieved from, 2024. https://par.nsf.gov/biblio/10448809
13. Kumar S. Singh A. Trustworthy AI Challenges in LLM Security. AAAI. Retrieved from, 2024. https://ojs.aaai.org/index.php/AAAI/article/view/32088
14. Gupta P, *et al*. Privacy in Medical AI Models. NEJM AI. Retrieved from, 2024. https://ai.nejm.org/doi/full/10.1056/AIdbp2400537
15. Zhang J, *et al*. Enhanced Security in LLMs. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2502.08966
16. Chen Y, *et al*. Differential Privacy for LLMs. ArXiv. Retrieved from, 2024. https://arxiv.org/abs/2503.11232
17. V. Rathod S. Nabavirazavi S. Zad SS, Iyengar, Privacy Security Challenges in Large Language Models, IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, 00746-00752, doi: 10.1109/CCWC62904.2025.10903912. keywords {Industries;Ethics;Privacy;Technological innovation;Firewalls (computing);Large language models;Computational modeling;Medical services;Threat assessment;Security;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}, https://arxiv.org/abs/2503.09780
18. Morris C, *et al* Threats to Large Language Models. IEEE Xplore. Retrieved from, 2024. https://ieeexplore.ieee.org/abstract/document/1090391
19. T. Su B. Zhang C. Zhang L. Wei, Privacy Leak Detection in LLM Interactions with a User-Centric Approach, IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Sanya, China, 2024, 1647-1652. doi: 10.1109/TrustCom63139.2024.00226. keywords: {Privacy Data privacy Large language models Filling Security Leak Detection General Data Protection Regulation Protection Monitoring Faces Privacy leak detection privacy in large language models}, https://ieeexplore.ieee.org/document/10945196
20. T. HG, Vu XB, Hoang, User Privacy Risk Analysis within Website Privacy Policies, International Conference on Multimedia Analysis Pattern Recognition (MAPR), Da Nang, Vietnam, 2024, 1-6. doi 10.1109/MAPR63514.2024.10660854. keywords: {Training; Learning systems Privacy Data Privacy Law Large language models Transfer learning Privacy policy legal compliance LLM RAG privacy risk analysis}, https://ieeexplore.ieee.org/document/10660854
21. J. Sun B. Suleiman I. Ullah, Effectiveness of Privacy-Preserving Algorithms for Large Language Models A Benchmark Analysis, 21st Annual International Conference on Privacy, Security and Trust (PST), Sydney, Australia, 2024, 1-8. doi 10.1109/PST62714.2024.10788045. keywords {Training Measurement; Data privacy Adaptation models Privacy Organizations Benchmark testing; Data models; Security Protection large language model's privacy-preserving algorithms differential privacy benchmarks}, https://ieeexplore.ieee.org/document/10788045