INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Journal ID: 234A-56Z1

- Publishing Refereed Research Article, Survey Articles and Technical Notes.

IJAIRD

IAEME Publication
Chennai, India
editor@iaeme.com/ iaemedu@gmail.com
https://iaeme.com/Home/journal/IJAIRD

International Journal of Artificial Intelligence Research and Development (IJAIRD)
Volume 3, Issue 2, July-December 2025, pp. 1-17, Article ID: IJAIRD_03_02_001
Available online at https://iaeme.com/Home/issue/IJAIRD?Volume=3&Issue=2
Impact Factor (2025): 19.25 (Based on Google Scholar Citation)
Journal ID: 234A-56Z1; DOI: https://doi.org/10.34218/IJAIRD_03_02_001

SAFEGUARDING SENSITIVE DATA IN LLM RACE - AWARENESS AND PROTECTION

Sagar Patel
Independent Researcher, Virginia, USA.
ORCID: https://orcid.org/0009-0002-9738-2182

ABSTRACT

The rapid growth of large language models (LLMs) have brought significant advancements in how individuals and organizations generate and process information. The ease with which these LLMs integrate into our everyday applications introduces new risks of exposing sensitive data to the world and poses challenges to properly safeguard it from potential leaks. As users often unknowingly transmit personal, medical, financial, and proprietary information into these LLMs without fully understanding the risks involved, the potential for data breaches, privacy & regulatory violations, monetary damage and reputational damage continues to grow. While the traditional methods of protecting data enforced by organizations are effective in controlled environments, the dynamic and unstructured nature of information flowing through LLMs renders these methods ineffective. This paper highlights the importance of sensitive data awareness in the context of LLM usage, examines the risks associated with data exposure, and proposes information safeguarding strategies. In the era where AI is omnipresent and integration of LLMs continues to accelerate in critical everyday workflows, protecting sensitive information should be recognized as a fundamental need for both individuals and organizations.

Keywords: AI Governance, AI Risk Mitigation, Data Privacy, Generative AI, Large Language Models (LLMs), PII Detection, Sensitive Data Protection, Regulatory Compliance

Cite this Article: Sagar Patel. (2025). Safeguarding Sensitive Data in LLM Race - Awareness and Protection. *International Journal of Artificial Intelligence Research and Development (IJAIRD)*, 3(2), 1–17. DOI: https://doi.org/10.34218/IJAIRD_03_02_001

1. Introduction

Large language models (LLMs) such as Open AI's ChatGPT, Gemini, GitHub Co-pilot, Meta AI, Claude, Microsoft Co-pilot and various other advanced LLMs are becoming increasingly popular in a wide range of industries such as finance, healthcare, information technology, education, and law [1]. Their ability to process human-readable language, support decision-making, summarize information and generate human-like responses has led to a smooth flow of information between humans and machines, providing unprecedented innovation and efficiency. However, this rapid adoption of LLMs also introduces significant risks in the handling of sensitive data.

Sensitive information - including personally identifiable information (PII), medical records, financial transactions, confidential business data, and intellectual property is often unknowingly shared with LLMs [2]. Unlike traditional software systems that operate within well-defined security constraints and structured data flows, allowing the data management rules to be well-defined, LLMs interact with pure text data provided by users, which is highly unstructured and difficult to govern with the same rules. This introduces new risks for data leaks, unauthorized data retention, and potential regulatory compliance breaches, mandated by the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA) and other privacy frameworks [3] [4].

Also, traditional security measures such as in-flight or at-rest encryption, access control and data classification are rendered ineffective in addressing the risks, as all of the data from users is fed in as plain text and the traditional software systems are not well-equipped to detect sensitive data before being let into these LLMs. Furthermore, the LLMs could have little to no sensitive data detection capabilities baked in at the point of ingestion and combining that with their ability to store and accidentally reproduce sensitive information requires a re-evaluation of our traditional data protection strategies.

This paper argues that in the context of widespread LLM usage, an evolution of the traditional data protection techniques is a must and should become a top priority for Data Management organizations within any business. Through this paper, we analyze the nature of sensitive data risks introduced by LLMs, explore examples of potential vulnerabilities and propose both procedural and technical measures to mitigate the risks. By fostering awareness and taking proactive steps, organizations and individuals can better navigate the challenges of using large language models safely and responsibly in a data-driven world.

2. WHAT IS SENSITIVE DATA?

Sensitive data refers to any information that, if improperly handled, exposed, or misused, could negatively impact an individual, organization, or entity [5]. This impact may include identity theft, financial loss, reputational damage, regulatory penalties, or privacy breaches.

While interacting with LLMs, sensitive data can often be embedded within casual conversations, input prompts or document summaries, without the user's explicit awareness or consent.

Sensitive data can be broadly categorized into the following groups:

- **Personally Identifiable Information (PII):** Information including but not limited to names, Social Security numbers, addresses, phone numbers, email addresses, national ID numbers, and other biometric identifiers that can directly or indirectly reveal an individual's identity are termed as personally identifiable information.
- **Financial Information:** Details about individuals or businesses, such as bank account numbers, credit card numbers, banking transactions, tax records, investment information, and other data that could be exploited for financial gain.
- **Health and Medical Information:** Covers patient records, diagnoses, medical history, prescriptions, insurance details, and any data protected under regulations like HIPAA.
- **Proprietary and Intellectual Property (IP):** Includes trade secrets, confidential business strategies, unpublished research, source code, product designs, and other forms of non-public company knowledge [6].
- **Confidential Communications:** Refers to sensitive discussions, contracts, legal documents, internal emails, negotiation materials, and privileged communications that require confidentiality.

- **Regulated Data:** Data specifically protected by legal or industry regulations, such as GDPR-covered personal data, Payment Card Industry (PCI) information, or classified government materials [7].

Each category carries unique regulatory, ethical, and operational implications. As LLMs increasingly handle unstructured data from diverse sources, the boundaries between these categories may blur, further complicating the identification and protection of sensitive information.

3. LLMs and Their Risks

Large language models (LLMs) are trained using a lot of text as input data, which helps them give smooth and relevant answers. Their capabilities have made them valuable tools for automating tasks such as drafting emails, summarizing documents, analyzing unstructured text, and even providing legal or medical suggestions [8]. However, these same capabilities introduce new and complex risks when sensitive data is involved.

3.1. Unstructured Input and Informal Data Sharing

LLMs are typically accessed via natural language interfaces, encouraging informal user interactions. Users may input sensitive information without recognizing the implications, such as pasting full documents into a prompt for summarization or including client details while drafting correspondence. These unstructured inputs bypass many of the traditional enterprise controls used to monitor and filter sensitive data.

3.2. Data Retention and Model Memory

While most publicly available LLMs are designed not to retain user data beyond the session, some implementations, particularly fine-tuned models or enterprise versions, may store interactions for future improvement or auditing [9]. If sensitive information is not filtered or anonymized beforehand, it may become embedded in logs or training datasets, potentially leading to data leaks or unauthorized access.

3.3. Prompt Injection and Model Leakage

Adversarial techniques such as prompt injection can trick LLMs into revealing information from prior sessions or unintended internal behaviors. In enterprise deployments where LLMs are used to interact with internal systems, such attacks could extract proprietary or confidential data [10].

3.4. Shadow AI and Unauthorized Use

Employees may use public LLMs for productivity without formal approval or oversight—often referred to as "shadow AI." This practice circumvents IT governance and increases the risk of exposing confidential data to external platforms [11].

3.5. Regulatory and Legal Exposure

If sensitive data is shared with LLMs hosted by third parties, organizations may unknowingly violate data protection laws, especially when user data crosses jurisdictions or lacks appropriate consent. Fines under GDPR or HIPAA violations can be substantial, with additional consequences to reputation and stakeholder trust.

These risks highlight the urgent need for new approaches to data awareness, governance, and tooling in LLM-integrated environments. Unlike traditional systems where access and data boundaries are well defined, LLMs operate in contexts where information flows are often opaque, decentralized, and user-driven.

4. Why Traditional Data Protection Measures are not enough

Conventional data protection strategies, such as encryption, firewalls, role-based access controls, and endpoint security, have long served as the foundation for safeguarding sensitive information in structured systems. These tools are designed to protect data at rest and in transit within clearly defined boundaries. However, when it comes to interactions with large language models (LLMs), these mechanisms are increasingly insufficient.

4.1 Lack of Structure in LLM Interactions

Traditional data loss prevention (DLP) tools are optimized for structured data and predictable data flows, such as email servers, databases, or document repositories. In contrast, LLMs operate on unstructured natural language inputs, making it difficult for legacy systems to detect and filter sensitive content embedded in free-form text [9]. Sensitive data may be hidden within contextually rich but semantically ambiguous prompts, preventing detection.

4.2 Absence of Boundary Enforcement

LLMs are often accessed via APIs, chat interfaces, or third-party integrations, where the distinction between internal and external systems is blurred. Data boundaries that were once enforced by network perimeter security or access control layers may no longer apply. Sensitive information can accidentally leave an organization without being noticed or raising any alarms [12].

4.3 No Built-In Sensitivity Awareness

Most LLMs do not natively understand or classify the sensitivity of the data they process. They treat all input as generic language, lacking mechanisms to recognize or enforce constraints based on data classification, compliance levels, or privacy designations. This increases the likelihood of inappropriate processing or disclosure of regulated information [13].

4.4 Difficulty Auditing and Monitoring Usage

Monitoring LLM interactions is more complex than tracking conventional user actions within enterprise applications. The conversational nature of LLMs means that data may be revealed gradually, through multiple prompts, or inferred indirectly. Standard audit trails and logging systems cannot reconstruct these interactions meaningfully or flag risks in real time [14].

4.5 Rapid and Decentralized Adoption

The adoption of LLMs often occurs faster than IT governance structures can adapt [11]. Technical and non-technical stakeholders may integrate these tools into workflows without consulting security teams. The result is a fragmented risk surface, where traditional protections are bypassed entirely or implemented inconsistently.

Together, these factors demonstrate that existing data protection controls must be complemented by new context-aware and language-sensitive approaches, integrated directly into LLM workflows. The next section outlines practical vulnerabilities and real-world scenarios where these challenges manifest, emphasizing the need for proactive measures.

5. Real World Examples

To better understand the risks of using LLMs without proper data safeguards, this section presents real-world cases in which sensitive information was exposed or placed at risk due to the inappropriate use of language models.

5.1. Samsung: Source Code Leak via ChatGPT

In 2023, Samsung engineers accidently leaked confidential source code and internal meeting notes by pasting them into ChatGPT to get help with debugging and summarization.

The inputs included proprietary code and performance data related to semiconductor manufacturing. Because the platform was external and lacked an enterprise agreement, the data was processed and potentially retained by OpenAI, raising serious concerns about IP leaks and

compliance violations. Samsung responded by restricting internal use of generative AI tools [2].

5.2. Apple: Restriction on ChatGPT Usage Due to Data Leak Concerns

In May 2023, Apple restricted employees from using ChatGPT and similar AI tools over concerns that confidential company data could be inadvertently leaked. The company feared that employees might input sensitive information into these platforms, which could then be stored and potentially accessed by others [15].

5.3. OpenAI: ChatGPT Bug Exposes User Chat Histories

In March 2023, a bug in ChatGPT allowed some users to see the titles of other users' chat histories. Although the content of the conversations was not exposed, the incident raised concerns about data privacy and the potential for sensitive information to be unknowingly shared [16].

5.4. Google's Gemini AI Exploited by State-Sponsored Hackers

In early 2025, reports emerged that state-sponsored hackers were leveraging Google's Gemini AI to enhance their cyberattack capabilities against U.S. targets. These actors used Gemini to craft more convincing phishing emails and generate malicious code, thereby increasing the sophistication and effectiveness of their attacks. The exploitation of Gemini AI by malicious entities raised serious concerns about the potential misuse of advanced AI tools and the need for robust safeguards to prevent such scenarios [17].

5.5. Meta's Use of Personal Data for AI Training Without Explicit Consent

In June 2024, Meta faced significant scrutiny after it was revealed that the company planned to use years of personal posts, private images, and online tracking data from European users to train its AI technologies without obtaining explicit consent. The privacy advocacy group NOYB filed complaints in 11 European countries, arguing that Meta's actions violated users' fundamental rights to data protection and privacy [18]. Critics highlighted concerns over the lack of transparency and the inability for users to opt out or have their data removed once included in the training datasets. This incident underscores the ethical and legal challenges of using personal data in AI development without clear user consent.

6. Best Practices and Mitigation Strategies

As large language models (LLMs) become increasingly integrated into enterprise, healthcare, educational, legal, and consumer facing systems, safeguarding sensitive data

requires security control beyond the conventional approaches. This section outlines techniques, combining technical defenses, architectural design, organizational policies and AI driven strategies to minimize the data exposure risks.

6.1 Input-Level Controls

6.1.1 Data Minimization

Only necessary and non-sensitive data should be shared with LLMs. Sensitive information such as personal identifiers, credentials, or proprietary code should be removed or masked prior to processing. Submitting full documents or records containing proprietary business information should be avoided.

6.1.2 Pre-processing and Redaction Pipelines

Build orchestration pipelines to periodically check for sensitive data across your organization's databases using techniques such as Regex pattern matching, positive and negative keywords checks, or open source libraries like Microsoft Presidio and its integration with ML models such as Spacy, Stanza, Transformers, Gliner etc. to identify and redact PII/PHI before inputting data into an LLM [19].

6.1.3 Prompt Risk Scoring

Establish a process for risk scoring the LLM prompts based on a sensitivity score based on detected PII/PHI or business context. High-risk prompts can be blocked entirely or redirected to private models for safer handling.

6.1.4 Prompt Injection Defense

Mitigate prompt injection attacks by sanitizing user inputs, limiting prompt chaining, and implementing instruction boundary hardening [20].

6.1.5 Embedding-Based Prompting (Privacy-Preserving Queries)

Use vector embeddings instead of raw data to query LLMs, minimizing direct exposure of source content [21].

6.2 Middleware and Architectural Safeguards

6.2.1 Controlled Interfaces and Proxy Firewalls

Deploy secure, company-controlled LLM gateways that can filter and log traffic, enforce access policies or implement rules based blocking for unsafe input or output patterns [22].

6.2.2 Stateless and Memory-Free LLM Inference

Using configurations to clear out user data after sessions could help avoid long-term model memory retention and minimize breach surfaces [23].

6.2.3 On-Prem or VPC Deployment of Models

Deploying LLMs in secure, internal environments for full control over data flow and access. This is essential for high-compliance industries (e.g., healthcare, finance, defense).

6.3 Output Moderation and Model Behavior Control

6.3.1 Output Post-Processing and Redaction

Apply detection and redaction tools similar to pre-processing pipelines to model output before presenting it to users. Use NLP-based NER checks or custom rule sets to flag hallucinated sensitive content [24][25].

6.3.2 Model Fine-Tuning and RLHF Guardrails

Fine-tune LLMs using curated, sanitized datasets. Use Reinforcement Learning from Human Feedback (RLHF) to discourage sensitive content generation and increase refusal accuracy [26].

6.3.3 Output Mutation Limits

Restrict model outputs that match known sensitive data formats (e.g., credit cards, SSNs) through regex scanning or decoder penalties [27].

6.4 Organizational and Governance Controls

6.4.1 Employee Training and Awareness

Periodically train employees on LLM risks, including prompt safety, sensitive data recognition, and acceptable use. Tailor training to user roles.

6.4.2 Acceptable Use Policies (AUPs)

Create and enforce usage policies for generative AI at organizational level. Clearly define prohibited input content and secure usage standards.

6.4.3 Role-Based Access Control (RBAC)

Limit who can access LLMs and under what data contexts. Restrict access to models trained on proprietary or regulated datasets.

6.4.4 Policy-as-Code Enforcement

Use frameworks like Open Policy Agent (OPA) to define AI usage rules programmatically. Integrate with CI/CD pipelines and LLM access APIs [28].

6.4.5 Shadow AI Detection and Blocking

Detect unauthorized use of public AI tools using proxy or SaaS monitoring. Block access to consumer LLM sites unless authorized [11].

6.4.6 Third-Party Vendor Evaluation

Vet external LLM providers for security controls, data retention policies, and regulatory alignment (SOC 2, ISO 27001, HIPAA, etc.).

6.4.7 Compliance Monitoring and Audits

Schedule audits for AI use, data handling, and prompt logging. Ensure adherence to GDPR, CCPA, HIPAA, or local privacy laws.

6.5 Mitigation Strategies Classification

The table [1] below summarizes all the mitigation strategies discussed above based on the complexity vs impact assessment and the layer at which each strategy can be applied in an LLM based architecture.

Table 1 Mitigations Strategies by Layer, Complexity and Impact

Strategy	Layer	Complexity	Impact
Data Minimization	Input	Low	High
Prompt Redaction (NER, Presidio)	Input	Medium	High
Prompt Risk Scoring	Input	Medium	Medium
Prompt Injection Defense	Input	Medium	Medium
Controlled Interfaces / Proxies	Middleware	High	High
Stateless LLM Configuration	Middleware	Medium	High
Embedding-Based Querying	Input/Model	High	High
Output Moderation and Filters	Output	Medium	Medium
RLHF-Based Model Guardrails	Model	High	High
RBAC for LLM Access	Governance	Low	Medium
Shadow AI Monitoring	Governance	Medium	High
Policy-as-Code (OPA, Rego)	Governance	Medium	Medium
Vendor Risk Review	Compliance	Low	Medium
AI Usage Audits	Compliance	Medium	High

6.6 In-depth Mitigation Layers Flow Diagram

The figure [1] below shows a simplified flow diagram for each stage in the lifecycle of LLM based software systems, providing guidance on where the above discussed strategies can be applied when building such systems and making them safer by design.

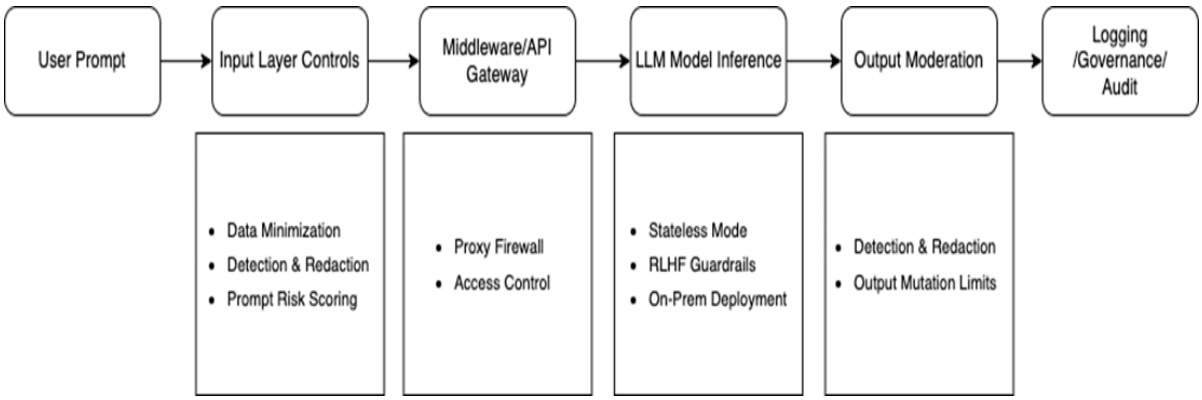

Figure 1 Mitigation layers flow diagram

7. Future Considerations and Regulatory Outlook

As generative AI continues to evolve rapidly, the protection of sensitive data in AI systems is drawing increasing attention from policymakers, regulators, and standards organizations. This section explores anticipated developments and the role of regulation in mitigating data risks associated with large language models (LLMs).

7.1 Emerging Regulatory Trends

7.1.1 AI-Specific Legislation (EU AI Act, U.S. Executive Orders)

The European Union has introduced the AI Act, which classifies AI systems by risk level and imposes specific obligations on high-risk use cases [29]. Systems processing biometric or sensitive personal data may require transparency, human oversight, and strict documentation. Similarly, the United States has issued an executive order encouraging responsible AI development, with agencies such as the FTC and NIST playing key roles in defining trustworthy AI [30].

7.1.2 Data Sovereignty and Cross-Border Controls

Governments are increasingly enacting data localization laws that require sensitive data to be stored and processed within national borders, especially for healthcare, defense, and financial data [31]. This has significant implications for cloud-hosted LLMs trained on globally sourced data.

7.1.3 Consent and Transparency Requirements

Future regulations are expected to mandate clearer user consent mechanisms, audit trails, and disclosures when AI systems are used to make decisions or when user data contributes to training models. This aligns with the GDPR's emphasis on lawful, transparent processing.

7.2 Ethical and Societal Considerations

7.2.1 Bias, Fairness, and Discrimination

LLMs may unintentionally reinforce biases present in training data, leading to harmful or discriminatory outputs [32]. Ensuring ethical use will require bias detection, representative datasets, and ongoing evaluations of societal impact.

7.2.2 Digital Trust and User Control

Maintaining trust in AI systems depends on granting users agency over their data, including the ability to opt out, view how their data is used, and request deletion [18]. These rights are central to future frameworks on digital self-determination.

7.2.3 Environmental Responsibility

The large computational footprint of LLMs also raises concerns about sustainability.

Future data protection policies may intersect with environmental regulations, encouraging more efficient, responsible training and deployment practices [33].

7.3 Anticipated Industry Shifts

- Wider adoption of AI Assurance Standards such as ISO/IEC 42001 to enforce trustworthy AI practices [34]
- Increased demand for on-premise/private LLMs for sensitive domains (e.g., healthcare, defense) to comply with jurisdictional data laws
- Increased regulatory scrutiny for consumer-facing AI assistants
- Development of "AI firewalls": platforms that vet, sanitize, and redact inputs/outputs around AI systems [22]
- Federated learning and synthetic data techniques will gain traction to enable safer model training without real data exposure [35].
- Global AI registries may emerge to document model usage, risk levels, and compliance status

As regulatory frameworks evolve, it is imperative for organizations to adopt a privacy-by-design approach in their AI strategy. Integrating legal, ethical, and technical safeguards early in the development lifecycle will not only facilitate compliance but also build long-term trust with users, partners, and regulators.

8. Conclusion

The rapid adoption of large language models (LLMs) in business, education, healthcare, and software development has brought significant efficiency and innovation, but it has also introduced substantial data privacy risks. Recent real-world incidents illustrate that sensitive data can easily be exposed, misused, or mishandled due to a lack of awareness and insufficient safeguard mechanisms.

This paper has outlined the categories of sensitive data, common risk scenarios involving LLMs, and multiple high-impact breaches from major organizations such as Samsung, Meta, and Google. It also highlighted actionable best practices, from a technical perspective like input redaction and access controls, to organizational measures such as policy enforcement and employee training. Going forward, regulatory frameworks such as the EU AI Act [29] and evolving U.S. policies are expected to demand greater accountability, transparency, and data protection in AI systems. Organizations must be proactive in embedding privacy by design into their AI workflows and staying ahead of both compliance requirements and public expectations.

Protecting sensitive data is no longer just a legal or technical requirement, it is the foundation to ethical AI usage and sustainable innovation in the age of generative models.

References

[1] Wikipedia contributors, “ChatGPT,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/ChatGPT

[2] T. Claburn, “Samsung Engineers Paste Sensitive Data into ChatGPT, Prompting Security Warnings,” Dark Reading, Apr. 2023. [Online]. Available: https://www.darkreading.com/vulnerabilities-threats/samsung-engineers-sensitive-data-chatgpt-w arnings-ai-use-workplace

[3] U.S. National Institute of Standards and Technology (NIST), “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),” NIST SP 800-122, Apr. 2010. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-122/final

[4] European Union, “General Data Protection Regulation (GDPR),” EU Law, 2016. [Online]. Available: https://gdpr-info.eu/

[5] U.S. Department of Health and Human Services, “The HIPAA Privacy Rule,” 2024. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

[6] World Intellectual Property Organization (WIPO), “What is Intellectual Property?” [Online]. Available: https://www.wipo.int/about-ip/en/

[7] U.S. Government, “Classified National Security Information,” Executive Order 13526, Dec. 2009. [Online]. Available: https://www.archives.gov/isoo/policy-documents/cnsi-eo.html

[8] H. Park and D. Ahn, “The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications,” in Proc. 2024 CHI Conf. Hum. Factors Comput. Syst. (CHI ’24), Honolulu, HI, USA, May 2024, pp. 1–30. [Online]. Available: https://doi.org/10.1145/3613904.3642785

[9] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, et al., (2020). Extracting Training Data from Large Language Models. arXiv:2012.07805. https://arxiv.org/abs/2012.07805

[10] E. Debenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini, D. Fabian, C. Kern, C. Shi, A. Terzis, and F. Tramèr, “Defeating Prompt Injections by Design,” arXiv:2503.18813, Mar. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.18813

[11] S. Moisset, “Shadow AI: Hidden Risks and Challenges”, Feb. 2025. [Online]. Available: https://www.freecodecamp.org/news/shadow-ai-hidden-risks-and-challenges/

[12] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, Y. Liu, “Prompt Injection Attack against LLM-Integrated Applications,” arXiv:2306.05499, Jun. 8, 2023. [Online]. Available: https://arxiv.org/abs/2306.05499

[13] S. Goyal, M. Hira, S. Mishra, S. Goyal, A. Goel, N. Dadu, K. DB, S. Mehta, N. Madaan, “LLMGuard: Guarding Against Unsafe LLM Behavior,” arXiv:2403.00826, Feb. 27, 2024. [Online]. Available: https://arxiv.org/abs/2403.00826

[14] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, E. Kiciman, “Defending Against Indirect Prompt Injection Attacks With Spotlighting,” arXiv:2403.14720, Mar. 20, 2024. [Online]. Available: https://arxiv.org/abs/2403.14720

[15] J. Vincent, “Apple restricts employees from using ChatGPT over fear of data leaks,” The Verge, May 19, 2023. [Online]. Available: https://www.theverge.com/2023/5/19/23729619/apple-bans-chatgpt-openai-fears-data-leak

[16] A. Ghosh, “ChatGPT Bug Leaked Users’ Conversation Histories,” All Tech Magazine, Mar. 23, 2023. [Online]. Available: https://alltechmagazine.com/chatgpt-bug-leaked-users-conversation-histories/

[17] A. Vakulov, “Hackers Hijack AI: Google Warns of Gemini Misuse by Cybercriminals,” Forbes, Feb. 3, 2025. [Online]. Available: https://www.forbes.com/sites/alexvakulov/2025/02/03/hackers-hijack-ai-google-warns-of-gemini-misuse-by-cybercriminals/

[18] NOYB - European Center for Digital Rights, “NOYB urges 11 DPAs to immediately stop Meta’s abuse of personal data for AI,” noyb.eu, Jun. 6, 2024. [Online]. Available: https://noyb.eu/en/noyb-urges-11-dpas-immediately-stop-metas-abuse-personal-data-ai

[19] Microsoft, “Microsoft Presidio,” GitHub, 2024. [Online]. Available: https://github.com/microsoft/presidio

[20] L. Beurer-Kellner, B. Buesser, A.M. Creţu, E. Debenedetti, D. Dobos, D. Fabian, M. Fischer, D. Froelicher, K. Grosse, D. Naeff, E. Ozoani, A. Paverd, F. Tramèr, and V. Volhejn, “Design patterns for securing LLM agents against prompt injections,” arXiv:2506.08837, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.08837

[21] G. Izacard, E. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering.” arXiv, 2021, Jul. 2020. [Online]. Available: https://arxiv.org/abs/2007.01282

[22] LlamaFirewall, “LlamaFirewall: An open source guardrail system for building secure AI agents” Meta, Apr. 2025. [Online]. Available: https://ai.meta.com/research/publications/llamafirewall-an-open-source-guardrail-system-for-buil ding-secure-ai-agents/

[23] X. Zhang, Y. Pang, Y. Kang, W. Chen, L. Fan, H. Jin and Q. Yang, "No Free Lunch Theorem for Privacy-Preserving LLM Inference," arXiv:2405.20681, Feb. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2405.20681

[24] Microsoft, "What is Azure AI Language Personally Identifiable Information (PII) detection?" , Mar. 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-infor mation/overview?tabs=text-pii

[25] LiteLLM, "Guardrails, Logging, and Rate Limits for LLM APIs," 2025. [Online]. Available: https://docs.litellm.ai/docs/proxy/guardrails/quick_start#:~:text=1.,LLM%20call%2C %20on%20 input%20&%20output

[26] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," arXiv:2204.05862, Apr. 2022. [Online]. Available: https://arxiv.org/abs/2204.05862

[27] A. Hassan et al., "Data Protection in LLM Interfaces: Threats and Defenses," ACM Queue, vol. 22, no. 1, pp. 54–66, Jan. 2024.

[28] R. Balchandani, "Enforcing Policy as Code: Open Policy Agent (OPA)," Medium, Aug. 15, 2022. [Online]. Available: https://raunakbalchandani.medium.com/enforcing-policy-as-code-open-policy-agent-opa-508883 d6c0e8

[29] European Commission, "EU Artificial Intelligence Act," European Union Law, 2024. [Online]. Available: https://artificialintelligenceact.eu/ai-act-explorer/

[30] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, Oct. 30, 2023. [Online]. Available: https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-or der-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[31] Government of India, "Digital Personal Data Protection Act, 2023," Gazette of India, Aug. 11, 2023. [Online]. Available: https://www.fpf.org/blog/the-digital-personal-data-protection-act-of-india-explained/

[32] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "On the Dangers of Stochastic Parrots," FAccT 2021, Mar. 2021. [Online]. Available:https://dl.acm.org/doi/10.1145/3442188.3445922

[33] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, Jul. 2019, pp. 3645–3650. [Online]. Available: https://aclanthology.org/P19-1355

[34] International Organization for Standardization, "ISO/IEC 42001:2023 Artificial Intelligence Management system," ISO, Dec. 2023. [Online]. Available: https://www.iso.org/standard/81228.html

[35] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3298981

Citation: Sagar Patel. (2025). Safeguarding Sensitive Data in LLM Race - Awareness and Protection. International Journal of Artificial Intelligence Research and Development (IJAIRD), 3(2), 1–17.

Abstract Link: https://iaeme.com/Home/article_id/IJAIRD_03_02_001

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJAIRD/VOLUME_3_ISSUE_2/IJAIRD_03_02_001.pdf

✉ **editor@iaeme.com**