iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82143

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Data Privacy in AI Chatbots: Understanding Risks and Ways to Protect User Information

Astha Rajput[1], Hema Sahu[2], Akhand Upadhyay[3], Yuvraj Singh[4], Ms. Heena Kausar[5]
[1 2 3 4]*Students,* [5]*Guide, Department of Computer Application, Shri Shankaracharya Professional University, Bhilai, India*

Abstract: ***The point is as follows: we are all discussing conversational AI based on huge language models, in the areas where privacy is incredibly important, such as healthcare and finance. However, there is a massive struggle between the necessity of these models to process huge quantities of information and the inability to violate the law regardless of the cost in terms of privacy. These LLMs also search through huge heaps of sensitive data, which increases the chances of spills by a significant margin. Throughout this paper, I take a close look at the privacy traps that are presented at every step of the life of an LLM, beginning with the manner in which we collect vast amounts of data, all the way to the ways in which people might attempt to actively attempt to peep inside. I compare all three leading privacy- enhancing tools: Differential Privacy, Federated Learning, and Secure Multi-Party Computation by discussing their strengths and weaknesses as well as the additional effort they require and the potential impact of extra effort on the utility of the model.***

I. INTRODUCTION

A. Background and Motivation

The development of LLDMs fully transformed the approach to natural language material, and now the emergence of cool projects occurs in virtually any industry. The magiclies in the fact that these models are taught insane complex language patterns on massive text piles. However, loading them with lots of data to feed them on training and finding answers, that is literally a privacy minefield, whether due to a data leak or even advanced attempts at attacks.

The greatest bane is the vast volumes of delicate information that flow through these models. By that I mean that LLaMs can contain terabytes or even petabytes of data in their training sets. And straining that all out of it? Basically a no-go. It implies that personal information, such as PII, health documents, financial information, and biometrics, is being peeped at by the model even before it is shipped. The larger the data, the larger the chance that something will leak or be abused, which is against the rights that all of us are concerned about. That is a clear indication that we badly require good tech remedies and governance strategies that are commensurate to the size and intricacy of how conversational AI actually reads.

B. Core Problem Statement

Hence the key research question is determining and demonstrating concrete, deployable methods of ensuring data security and also of maintaining models helpful (in both accuracy and performance). Much of the existing privacy solutions require you to make a bad trade-off, turn up the noise volume to make it private and lose a ton of quality. It is difficult to implement such models by companies and consumers because of that. Also, the conversational interface introduces an entire new range of attack angles, which require guardrails unlike those with which we operate in conventional cyber-security. After all, this paper attempts to unite the most recent mitigation strategies, juxtapose them in practice, by employing understandable metrics, and observe which ones reap their benefits.

C. Structure of the Paper

We are going to divide this step by step. Section II will take a step-by-step tour of the existing list of privacy risks, and categorize them as passive leaks and active attacks. In the third section, I present a way I approached studying and comparing the various tools and practices. The results are located in section IV: the strengths and weaknesses of each mitigation strategy against each other. Section V then considers the broad perspective of the basic PETs and regulatory systems on an international level. At last, Section VI brings it all together, indicates what must be done, both on the policy and the technical side, and where we should be seeking next.

II. LITERATURE REVIEW

1) Privacy Threats Taxonomy within the LLCM Data Lifecycle. The model of the data lifecycle can be used to classify privacy risks in conversational AI systems depending on their origin.

Data Governing and Risk of Data Collection.

The greatest privacy hits most of the time occur at the time of collecting and preparing the data. Leaving information in there without out the express consent or even awareness of people is a huge governance failure. In the modern world, users demand a clear description of the use of their data, and they want to have control over it. An example of the text book is LinkedIn: the users discovered that they were automatically subscribed to feed their data into generative AI models. That was a storm as default settings were poor, and people lost their confidence due to the absence of transparency. And since the volume of data intake is so immense, you inherently have systemic threats, such as unfettered surveillance or training pre-existing bias into models.

Cyber security and Systemic risks (Passive Leakage)

Such risks are based on its weak points in the model design, system settings, or the infrastructure. The problem of data leakage, i.e. the unintentional exposure of sensitive data, is an issue that AI models keep raising concerns about. A great example is the ChatGPT case because one of the big language models created by Open AI accidentally displayed to some users the name of the chat history of other users. It demonstrates that it is not just large public models that are at risk, but also small proprietary, in-house applications, such as a health-care diagnostic tool that may spill patient data. Something as harmless as a misconfigured model output, or even a message of error, may accidentally leak internal data, proprietary algorithms, or configuration information.

Memorization (Model-Inherent Risks)

Due to the fact that LLMs are trained to analyze large sources of text, they tend to memorize certain sections of their training samples. The fact that memorizing provides an opportunity to leech sensitive information in case it is included in the training data. Membership Inference Attacks (MIA). Are aimed at leveraging that with a view to attempting to guess whether a specific record was included in the training set. In case of MIAs victory, this will prove that ML models leak memorized information, and it is a serious privacy warning flag, particularly when the training data includes super-secret information.

2) Conversation System Active Privacy Attacks. Active privacy attacks exploit the fact that conversational AI can be so interactive and language-focused.

Prompt Injection Attacks

Fast injection kills by using the mechanism by which LLMs execute complex and natural- language instructions. Attackers create inputs that are malicious and they deceive the LLM to tell or do things it is not supposed to do, bypassing safeguard filters and output restrictions. They are able to conceal commands which lead to the exposure of leaky information such as API keys, credentials, or internal notes in sensitive data contexts. Because the attack entry is primarily through the chat interface, we require more than standard network security measures and are interested in verifying the intention of prompts and the character of what the model says.

Membership Inference Attacks (MIA)

MIA is a major form of attack, which attempts to decide whether the information of a particular individual was included in the training set. People typically measure MIAs based on measures such as Area Under the Curve(AUC), accuracy, or the True Positive Rate at low False Positive Rate (TPR @low FPR), however, the most recent studies indicate that the measures are capable of overlooking significant differences between attack strategies. To decide on privacy correctly, it is necessary to have a more comprehensive range of measures that can represent the great diversity and multiplicity of data leaks of various model types.

3) Basic Privacy-Enhancing Technologies(PETs)

The technological community has developed a few state-of-the-art privacy-saving systems in the cases of big-data.

Differential Privacy (DP)

DP is a statistical method which provides a quantifiable privacy assurance, conventionally indicated with a parameter epsilon. Noisiness is added to the raw data or the gradients during training, and it is difficult to know whether any particular record was used or not. The lower the epsilon, the higher the privacy, and this is usually done with an increase of noise amount.

Federated Learning (FL)

FL divides the training to the local devices and therefore you do not have to centralize raw data. The model is trained at each device, and model updates (or gradients) are provided to a central server to be added. FL reduces the risk of data leaks at a central store, however, research has indicated that FL alone cannot eliminate the risk, an attacker can still use the updates to make inferences about sensitive information.

Secure Multi-Party Computation (SMPC).

SMPC allows different parties to calculate a common function without disclosing the privacy of each other. By using SMPC and FL together, the central server is able to safely compile encrypted model updates and it is difficult to reverse engineer or detect which participant provided a specific update.

TEEs or Trusted Execution Environments.

Specialized hardware is used to isolate sensitive data and compute in relation to the rest of the OS and host; it is called TEEs or secure enclaves. TEEs form a component of a more mature privacy toolkit and can be used with PETs such as homomorphic encryption or SMPC to establish cryptographic resilience in digital systems.

Table I summarizes the operational trade-offs for deploying leading PETs in conversational AI architectures

Technology	Privacy Guarantee	Primary Technical Challenge	LLM Application Context
Differential Privacy (DP)	Mathematically bounded individual data influence (noise injection)	Significant trade-off with model accuracy and high computation/batch size overhead	Protecting user contributions during centralized model training and aggregation
Federated Learning (FL)	Keeping raw data localized on client devices (decentralization)	Vulnerability to gradient leakage and high communication costs	Collaborative training across multiple organizations/devices without sharing raw data
Secure Multi-Party Computation (SMPC)	Enabling secure computation on encrypted data from multiple parties	High communication overhead and complex cryptographic implementation	Securely aggregating sensitive client updates or querying encrypted databases

III. RESEARCH METHODOLOGY

A. *System Model Definitions and scope.*

This discussion is based on the privacy consideration of three major architectural models used to deploy conversational AI nowadays:

Cloud-based LLM: The implementation is based on the centralized infrastructure with high comprehensive performance and massive scalabilities. The data is usually relayed to other external servers.

On-Device/ Edge LLM: It is calculated on the user device. This provides extremely low latency and in-built privacy since the information is not transmitted.

Hybrid LLM: This is a combination of types of models that apply on-device processing to sensitive data sanitization and real-time response to address urgent needs, and centralized and secure cloud resources to run complex models.

B. *Mechanism Comparison Criteria.*

In order to properly evaluate the effectiveness and deployability of privacy mechanisms, the following quantitative and qualitative criteria are applied:

Privacy Guarantees: Characterizing the theoretical and practical of protection (e.g.,sample -level DP, user -level DP, or cryptographic confidentiality).

Computational Overhead: Evaluation of resource intensity, in terms of increased latency and processing time and communication overhead, which is what makes it feasible in reality.

Utility Degradation: The measure of the adverse effect of the mechanism on the model performance (accuracy, perplexity, and most importantly semantic utility (the consistency of the generated output and its contextual correctness).

C. *Privacy Evaluation Measures and Standards.*

Assessment must have metrics that determine the technical assurances of the PETs as well as the operational maturity of the governance processes.

Meanwhile, Quantifiable Privacy Metrics.

Privacy loss is measured using the Differential Privacy parameters. The lower value indicates increased privacy which is generally obtained by adding more noise. In addition, adversarial robustness is quantified by evaluating the effectiveness of active attacks in particular Membership Inference Attacks and the effectiveness of blocking immediate injections.

Organizational Resiliency Measures (Governance)

Since LLMs are complex and because human error happens quite often (asarule,about15% of employees typed sensitive data in open LLMs), it is not enough to use only the static prevention. The industry interest has turned in to the organizational resilience, which means the possibility to quickly spot and react to the unavoidable events.

This is evaluated by industry standards of AI governance models. Key metrics include:
Data Classification Coverage: The portion of the critical datasets that is covered by Classification (PII / PHI / Biometrics), and the industry standards should be over 95 percent coverage. This validates the fact that sensitive inputs are detected before they are consumed, and this reduces risks linked to unauthorized gathering.
Data Sanitization Effectiveness: The proportion of sensitive fields that are masked or tokenized during ingestion, preferably more than 90.
Response Time: The Mean Time to Detect (MTTD) privacy events, the practical benchmark target of which is less than 15 minutes. This indicator is used to gauge the speed of the containment process after the accidental exposure of data, i.e. the conversation history leakage in the ChatGPT incident.

IV. FINDINGS: PRIVACY MECHANISMS AND ARCHITECTURES.

A. Training and aggregation A. Differential Privacy (DP).

The point is that DP is really huge when you have to train LLMs, and particularly when you are combining it with Federated Learning. The concept is to add a noise by applying two big tricks: DP- SGD which shakes up the derivatives you obtain at each device (that is at the sample-level privacy) and DP-FedAvg that adds noise to the central averaging step (that is at the user-level privacy).

The Utility - Privacy Dilemma
It appears that the consensus among all seems to be that vanilla DP arrangements are rather harsh slices into model performance. Noising that privacy bar has the effect of damaging the performance of the model in a rather fundamental way. It is the timeless invertizing correlation - the higher the noise, the less the utility. This complicates the provision of super accurate models into manufacturing.

Progressive Methodologies of improving this Trade-off.
To overcome this, scholars have been exploring improved mechanisms of adding noise and doing the aggregation. The new parameters of the vanilla algorithms work with such stuff as Haar wavelet transforms and other more cool noise tricks. The goal? Reduce the noise variance to ensure the model converts more and the model maintains high utility at the same time satisfying the privacy requirements. The utility of chatbots is not merely the correctness of words, but the fact that the meaning is not ruined after all the privacy circuses, such as pseudonyms and pseudonym reversal.

B. Secure Cryptographic Protocols: TEEs and SMPC.

That is where SMPC (Secure Multi-Party Computation) and TEEs (Trusted Execution Environments) enter the picture they are the standards of maintaining the privacy of the data during the calculations.

SMPC Implementation and Security.
SMPC can be used when it comes to collaborative AI work, such as joint training of a medical LLM by multiple hospitals without leaking the personal information of anyone. It ensures that the aggregated updates remain encrypted when fed into FL preventing these intermediate gradients that might be secrets leakers. SMPC is also capable of storing user queries encrypted even when inference is being carried out until the correct user reaches the point of seeing the answer.

The Basic Overwhelming Intensity of Computations.
Sadly enough, SMPC is very sluggish. Additional arithmetic and the associated bandwidth may substantially slow down the process of training and inference, particularly with large LLMs. Such a speed-v-cost trade-off means that SMPC typically finds itself in high-stakes are as such as health and finance, in which failure would be disastrous irrespective of the cost. The regulators are essentially coercing the companies to embrace SMPC out of necessity.

C. Architectural Resilience: Trade-offs of Deployment Model.

Make your selection where to run your model and you do indeed set a pace of how private you can be.
Table II: LLM Deployment Architecture Trade- offs for Privacy and Performance

Feature/Metric	On-Device AI	Cloud AI	Hybrid/Edge Approach
Raw Data Exposure Risk	Low (Data Stays on Device)	High (Data Sent to Servers)	Moderate (Sensitive inputs stay local)
Computational Capacity	Limited (Device Capabilities)	Near-Unlimited (Cloud Resources)	Flexible, Scalable (via Cloud integration)
Latency/Response Speed	Ultra-low (Real-time)	Higher (Internet Dependent)	Optimized (Local processing for speed)
Compliance/Control	Highest (Ideal for regulated environments)	Dependent on Provider Agreements	Requires strong governance coordination

On-Device AI: Through this model, everything is stored on the user device, and as a result, data does not leave the device and reduces the chances of transferring it to any location, and also, the device provides an extremely quick response since no networking is necessary. It is also very useful when the team requires absolute control over the data, but the flip side is that you are confined to the limited processing of the phones or laptops and that updating the model everywhere may be a logistical nightmare.
Cloud AI: the cloud provides you with nearly unlimited computing power and real-time updates, although the downside is that you must transfer data to remote servers and therefore reduce your privacy inherent to the cloud.
Hybrid Models: the majority of these individuals are currently going into hybrids where the sensitive data should remain in house (Edge Safety) and the cloud takes the heavy but not sensitive workloads. This is a compromise of privacy, size, and the distinct regulations that you are subjected to.

D. Data Sanitization and Run-time Defense.

Fancy crypto is not a thing you can count on; you need tools which operate end-to-end data protection.

Pre- Training Defense
In order to prevent this model memorizing data, we clean the data prior to the training. The use of techniques such as masking and tokenization is used to strip off PII/PHI of the final training set.

Runtime Defense Layers
Even on live runs you are required to guard against such things as prompt injection. Important ones are middleware, reverse proxies, and output filters. A reverse proxy can cleanse inputs or outputs until they reach logs preventing sensitive leaks. Good prompt validation and filtering is used to identify harmful input and stop rogue disclosures.

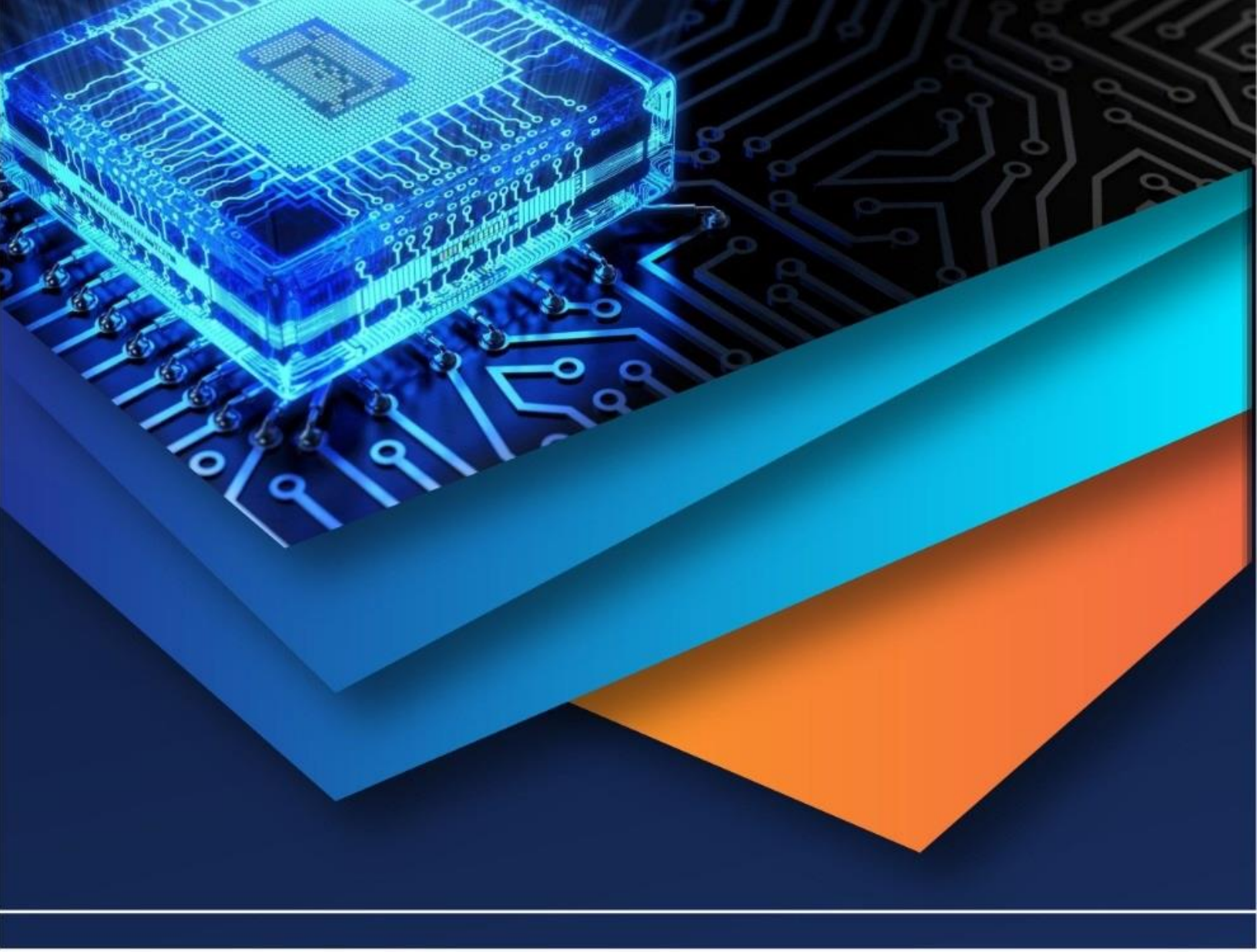
V. CONCLUSION

The issue of privacy of data on AI chatbots essentially falls on the cross of technical design and operation, regulation and organizational culture. There are special difficulties in conversations since they are abundant and unstructured. Threats such as inference attacks, regulatory fines and shadow AI information leakage are not far-fetched as I
Have read, however, a combination of privacy- enhancing technology, good governance, and user-focused controls can mitigate these risks significantly.
In the future, studies ought to explore ways of improving the utility privacy trade-off of differentiating privacy of language models, enhancing secure and robust federated learning systems, and establishing more practical and effective unlearning systems to support deletion rights. To enable policymakers and practitioners to coordinate around more transparent, explainable, and accountable conversational AI, it will be necessary to ensure that chatbots develop in a way that supports our rights without compromising on the ability to provide genuine user and institutional benefits.

REFERENCES

[1] [2406.07973] Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey-arXiv,accessedonNovember26,2025, https://arxiv.org/abs/2406.07973
[2] A Complete Survey on LLM-based AI Chatbots - arXiv, accessed on November 26, 2025,https://arxiv.org/pdf/2406.16937?
[3] AI Privacy Risks & Mitigations–Large Language Models (LLMs) - European Data Protection Board, accessed on November 26, 2025,https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf
[4] The Right to Be Forgotten—But Can AI Forget?|CSA-Cloud Security Alliance, accessed on November 26, 2025, https://cloudsecurityalliance.org/blog/2025/04/11/the-right-to-be-forgotten-but-can-ai-forget
[5] Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data-arXiv,accessedonNovember26,2025, https://arxiv.org/html/2503.01630v2

[6] Supervised Fine-Tuning(SFT) for PII Masking Using Axolotl - Medium,accessedonNovember26,2025, https://medium.com/@aakulkarni/supervised-fine-tuning-sft-for-pii-masking-using-axolotl-c306f3245bc6
[7] A Case Study on Samsung's ChatGPT Incident-Human Firewall, accessed on November 26, 2025,https://humanfirewall.io/case-study-on-samsungs-chatgpt-incident/
[8] Be Careful What You Tell Your AI Chatbot| Stanford HAI, accessed on November 26, 2025, https://hai.stanford.edu/news/be-careful-what-you-tell-your-ai-chatbot
[9] Security Concerns for Large Language Models: A Survey-arXiv,accessedon November 26, 2025, https://arxiv.org/html/2505.18889v2
[10] Unveiling AI Agent Vulnerabilities Part III: Data Exfiltration | Trend Micro(US), accessed on November 26, 2025, https://www.trendmicro.com/vinfo/us/security/news/threat-landscape/unveiling-ai-agent-vulnerabilities-part-iii-data-exfiltration



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY