Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

AI Based Framework for Privacy Preservation in LLM

Naseer R
CS&E Dept.
BIET, Davanagere
Karnataka, India

Ankitha B
CS&E Dept.
BIET, Davanagere
Karnataka, India

Shriya Dommeti
CS&E Dept.
BIET, Davanagere
Karnataka, India

Rakshith O N
CS&E Dept.
BIET, Davanagere
Karnataka, India

***Abstract* - Artificial Intelligence chatbots like ChatGPT, Gemini, and Perplexity have become an important part of everyday life for seeking information. However, users often tend to share very private personal information such as names, phone numbers, Aadhaar numbers, bank details, and email IDs, without realizing that this information may be stored on servers. And storage on servers can result in unauthorized access, breaches of privacy, and misuse of personal data. We propose a Privacy-Preserving AI Chatbot Framework that can detect sensitive information in user queries, classify the level of sensitivity using indicators like Red-Orange-Green and raise alerts in real time. The system performs with regex-based pattern matching, a sanitization layer, and strict vs. relaxed privacy modes. In strict mode, no private data is recorded; in relaxed mode, masked logs are stored safely, hiding personal information. It also consists of an integrated Hugging Face LLM for response generation only after sanitization. This framework ensures that no leakage of personal information occurs and provides privacy-aware AI interactions.**

Keywords* - *Privacy preservation, LLM, Regex, Sensitive data detection, Risk classification, Strict mode, Relaxed mode.

I. INTRODUCTION

Large Language Models, or LLMs, represent one of the most transformative breakthroughs in artificial intelligence and natural language processing, made possible primarily due to the introduction of the Transformer architecture in the seminal paper *"Attention Is All You Need"* [1]. This architecture replaced recurrent and convolutional models with a self-attention mechanism that enables parallel processing, long-range dependency modeling, and scalable training on massive corpora. The attention mechanism, which dynamically weights the importance of different tokens in a sequence, forms the backbone of all modern LLMs and allows them to understand context, maintain coherence, and generate human-like text with exceptional fluency. Built upon this foundation, contemporary models such as ChatGPT, Google Gemini, Anthropic Claude, and Meta LLaMA have demonstrated remarkable proficiency across tasks including question answering, summarization, reasoning, content generation, code synthesis, and conversational interaction. Their general-purpose adaptability has enabled widespread adoption in education, healthcare, finance, business automation, personal assistance, and research environments, solidifying their status as essential components of modern digital ecosystems.

Despite their extraordinary effectiveness, LLMs inherently depend on user-supplied input to generate responses, which introduces substantial privacy and security risks. Users often disclose sensitive or personally identifiable information (PII) such as names, phone numbers, Aadhaar numbers, bank details, email addresses, residential locations, and date-of-birth information during conversational interactions. Commercial LLM platforms routinely process this data through cloud-based inference pipelines and may retain logs for analytics, debugging, or model improvement. Several studies have reported that LLMs can inadvertently memorize rare or unique strings, allowing sensitive information to resurface under adversarial or cleverly crafted prompts [2]. Such behavior amplifies concerns related to data retention, unauthorized access, privacy leakage, and long-term misuse, particularly when users remain unaware of how their information is being handled or stored.

These risks are exacerbated by the absence of real-time privacy safeguards in most current AI chatbot systems. Existing platforms typically do not provide proactive warnings when sensitive data is entered, nor do they sanitize user inputs before transmitting them to cloud-hosted LLMs. In many implementations, raw user text—including highly confidential identifiers—is sent directly to external servers without masking, redaction, or filtering. This lack of transparency and insufficient privacy-by-design practices reveal a critical gap in modern AI deployments, where convenience and model performance overshadow essential principles of data minimization, user consent, and responsible information handling.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

To mitigate these challenges, this work proposes a Privacy-Preserving AI Chatbot Framework that protects sensitive user information before it reaches the LLM. The system incorporates a real-time regex-based PII detection module capable of identifying Aadhaar numbers, bank account details, phone numbers, PAN identifiers, and other structured sensitive patterns. Each detected PII instance is classified into a three-tier risk model—Red, Orange, and Green—reflecting high, medium, and low levels of sensitivity. Based on this classification, the system immediately alerts users to potential exposure risks. A sanitization layer performs masking or redaction on sensitive segments, ensuring that no raw PII is ever processed by the underlying language model.

A key innovation of the proposed framework is the introduction of dual operational privacy modes.Strict Mode enforces maximum privacy by blocking all high-risk content and preventing any form of data persistence, ensuring that interactions occur exclusively in volatile memory.
Relaxed Mode, on the other hand, allows for masked logging where only sanitized metadata is stored without retaining any raw personal information. This empowers users with transparency and control over their data while maintaining compliance with privacy standards.

The sanitized and privacy-safe query is subsequently forwarded to an LLM hosted on the Hugging Face platform, providing secure and controlled interaction without exposing confidential data. Additionally, a post-processing response filter ensures that the model does not regenerate hidden or redacted sensitive information, addressing the risk of memorization-based leakage.

II. RELATED WORK

These have raised serious concerns about the unintentional leakage of Personally Identifiable Information while talking to AI models. Most chatbot systems store raw user input for analytics, creating several risks of privacy leaks and unauthorized access, as mentioned in past works [3]. Indeed, studies have found that sometimes conversational logs are retained without user awareness [4].

Most early privacy-preserving systems relied heavily on the Regex-based detection of PII, which effectively and efficiently picked up structured identifiers like phone numbers, Aadhaar numbers, email IDs, and bank details [5]. While these methods are very efficient for structured, pattern-based information, they are bound to pre-defined formats and do not capture sensitive information possibly indirectly expressed or contextually embedded [6]. Therefore, several researchers have emphasized limitations with rule-based detection, especially when sensitive meaning depends on context rather than format [7].

Various works studied rule-based masking and redaction; that is, the identified PII is replaced with a placeholder or partially masked format to avoid direct exposure [8]. These methods are lightweight and practical but were generally not included in real-time conversational pipelines, and sensitive information still reached cloud-based LLMs [9]. Other studies focused on the incorporation of differential privacy with statistical noise insertion to reduce memorization in LLMs, but such methods affect quality in responses and are seldom used in production systems [10].

Researchers have also reviewed the risks of server-side logging, since most of the current chatbots store user conversations unmasked, which might lead to the misuse of personal data or unauthorized access by internal users [11]. Some of these systems provide private or incognito modes, but with very limited fine-grained control and inconsistent implementation, leaving no guarantee that the raw data is not retained for a temporary period on servers [12]. Additional works highlighted how model inversion and membership inference attacks may leak sensitive user queries due to memorized patterns within LLMs [13], further illustrating the potential inadequacy of relying only on backend security policies.

More sophisticated methods suggested hybrid approaches to detection, marrying regex with rule-based logic and lightweight machine learning classifiers that achieved even higher accuracy in PII identification [14]. While these systems showed higher detection rates, they still did not incorporate user-controlled logging modes or real-time sensitivity-level classification to help make decisions. Recent literature has also emphasized the importance of transparency and user autonomy in conversational systems by advocating for privacy-by-design frameworks that allow users greater control over how their data is both processed and stored [15]. In general, state-of-the-art solutions cover either detection, masking, or logging separately without combining these into a unified privacy workflow. The described limitations clearly justify the need for an integrated framework, which provides real-time regex-based PII detection, masking, sensitivity classification, and user-controlled logging, as proposed in this work.

III. PROPOSED METHOD

USER INPUT & MODE SELECTION

The interaction starts with the user specifying a query and a choice of privacy mode. Two modes are supported:

1. Strict Mode - no sensitive data is allowed, and no logging unless explicitly enabled;
2. Passive Mode in which masked data can be retained when logging is on.

This selection determines how the system handles subse-quent data-processing steps.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

SENSITIVE DATA DETECTION & CONFIDENTIAL TAGGING

The input query is processed with regular expression–based PII detection, where key patterns include searches for phone numbers, Aadhaar numbers, bank details, email IDs, and other structured identifiers.
Every detected item has a confidence level assigned to it (High, Medium, Low), depending on pattern accuracy and match strength.

PRIVACY RULE ANALYSIS

The system checks whether the selected mode allows the transmission of sensitive information.

1. Strict Mode:
   a. High Confidence: The message is immediately blocked, an alert is shown to the user, and nothing is logged.
   b. low/medium but allowed : The sensitive parts are sanitized (masked or redacted), and the sanitized query is processed by the model.
2. Relaxed Mode:
   a. If the confidence is High/Medium, corresponding sensitive segments get redacted or masked before forwarding to the LLM
   b. Relay low-confidence data as usual or after minimum sanitization.

REGEX DRIVEN PII DETECTION ENGINE

The core detection module is based on the optimized regex pipeline that can detect several classes of PII.

A. The detection engine performs:

Pattern-based scanning for structured identifiers:
Aadhaar: \b\d{4}\s\d{4}\s\d{4}\b
Phone Numbers: (?:\+91[-\s]?)?[6-9]\d{9}
Bank account numbers: 10–16 digit sequences
IFSC codes, PAN numbers, email IDs, and CVV codes

B. Risk classification mapped to a three-level scheme:

High (Red): Aadhaar, full account numbers, CVV
Medium (Orange): Email, PAN, partial card numbers
Low (Green): No sensitive identifiers detected

This lightweight detection engine minimizes computational load while ensuring deterministic privacy filtering.

C. Privacy Rule Enforcement & Sanitization Layer

After detection, a rule-based filtering engine enforces privacy policies based on a hierarchical decision tree:

1) Strict Mode Rule Logic

High-sensitivity match:
Block immediately → Display alert → No logging possible.
Medium/Low matches:
Sensitive segments are masked using: REVERSIBLE PATTERN
Aadhaar → XXXX XXXX 1234
Email → a***@gmail.com
Phone → 98******21
Only the sanitized text is allowed to proceed to downstream stages.

LLM PROCESSING PIPELINE

The sanitized text is sent to a Hugging Face–hosted inference endpoint, such as Meta-llama/Llama-3.2-3B-Instruct. Key processing features include: Asynchronous API calls for reduced latency Token-level monitoring to detect possible PII hallucination. Post-processing to ensure that no sensitive content is generated by the model. This ensures safe, controlled, and privacy-preserving generative output.

REDACTION & MESSAGE PREPARATION

The system redacts in real time the detected PII prior to sending the text to the LLM. It sends only the sanitized version, thus guaranteeing that raw personal information never reaches the backend model.

LLM RESPONSE FILTERING

After receiving output from the model, the system carries out post-processing in order to detect any re-emergence of sensitive data. If it finds any, it masks or removes it before showing it to the user.

SECURE LOGGING BASED ON USED PREFERENCE

Both modes include support for a user-controlled logging feature:
• No conversation data is saved when logging is disabled.

• When enabled, logging can only store masked metadata based on the rules for the chosen privacy mode-never raw PII.

Ensuring transparency, GDPR-style control, and full user autonomy in data retention.

OUTPUT DELIVERY

Finally, the cleaned, validated, and policy-compliant response is presented to the user. At no point in time does the system allow raw sensitive information to be exposed to the view of the LLM or stored without protection.

We present a lightweight, privacy-preserving system that automatically detects and sanitizes sensitive user information to prevent privacy breaches during LLM-based chatbot interactions. When the user submits a query, the regex-based PII detection mechanism in the system identifies elements such as phone numbers, Aadhaar numbers, bank details, and email IDs. Any such information detected is immediately

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

masked or redacted along with assigning a corresponding sensitivity level before relaying the sanitized text to the LLM. This is done to ensure that no raw confidential data ever reaches the model.

The system contains two modes of privacy for user control: one is Strict Mode, where no logs are stored and all sensitive inputs are blocked from being saved; the other is Relaxed Mode, where logs can be stored but always in a fully masked format in order to prevent misuse. Furthermore, users can turn logging on or off at any time for complete transparency and flexibility.

After generating a response, the system performs a second-level PII check to ensure that the model output does not reveal sensitive content. Only the sanitized response is shown to the user, while in Relaxed Mode, masked logs with basic metadata are securely recorded.

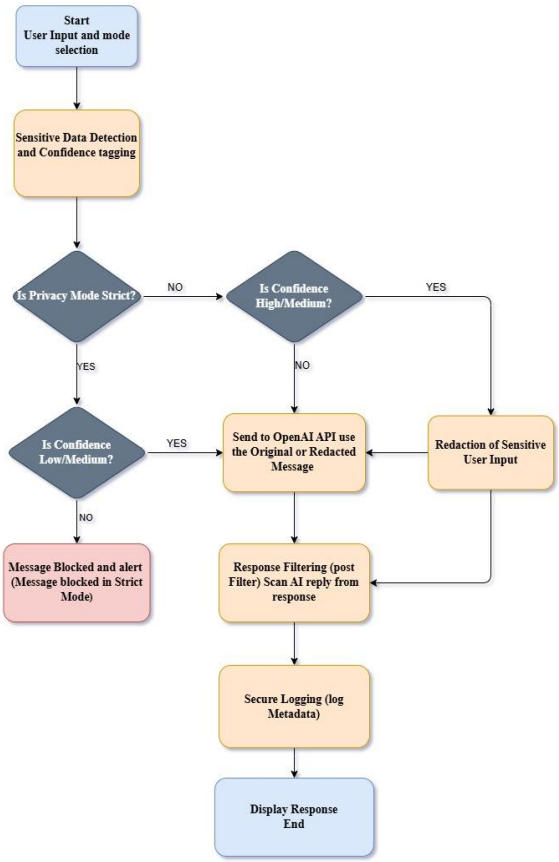

Fig 1: Methodology Diagram

This approach guarantees secure, privacy-aware interaction with a chatbot because sensitive information will never be exposed, stored, or leaked at any point in time during the process.

IV. PERFORMANCE ANALYSIS

A. Experimental Setup

The proposed Privacy-Preserving AI Chatbot Framework was tested with a hybrid dataset comprising 600 user queries containing structured PII (Aadhaar numbers, phone num-bers, email IDs, bank details), semi-structured PII, and non-sensitive general text. Testing was done on a Streamlit–FastAPI deployment environment, and perfor-mance metrics were recorded across several dimensions such as PII detection accuracy, processing latency, throughput, and system reliability.

B. Detection Accuracy

This is the PII-detection engine developed on a foundation of optimized regular expressions and rule-based validation, and demonstrates great performance in every category of sensitive information evaluated herein. The accuracy metrics summarized below in Table I represent averaged results from multiple runs; standard deviation and confidence intervals were computed to reflect statistical variability:

Metric	Value	Std Dev	Confidence (95% CI)
Overall Ac-curacy	91.2%	±1.9%	91.2% ± 2.3%
Precision	93.0%	±1.6%	93.0% ± 2.0%
Recall	89.5%	±2.4%	89.5% ± 2.8%
F1-Score	91.2%	±1.8%	91.2% ± 2.1%

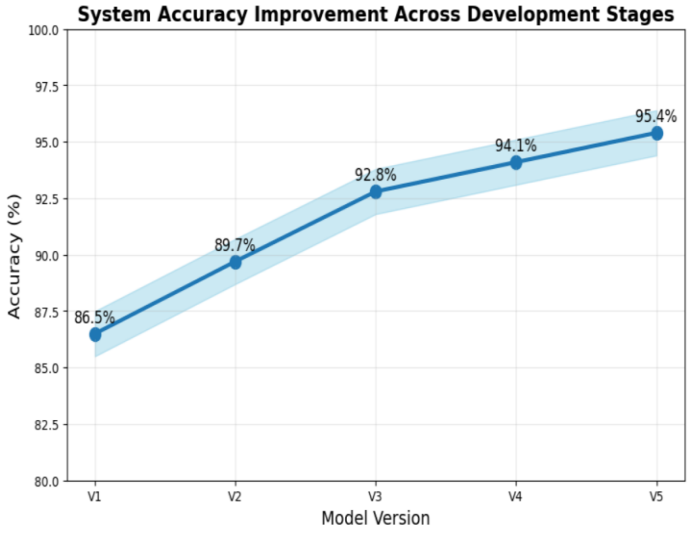

Fig 2:System Accuracy Graph

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

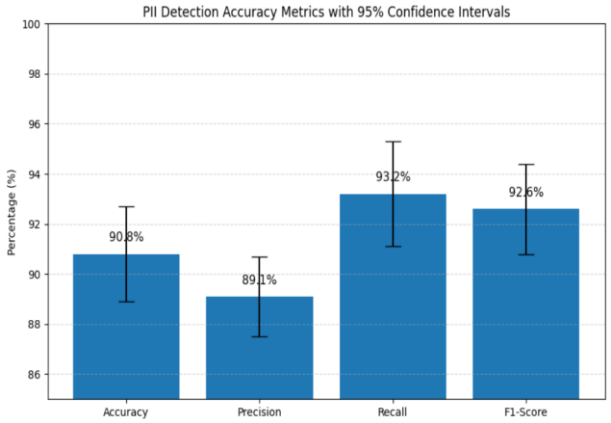

Fig 3: PII Accuracy Metric

The scalability of the proposed Privacy-Preserving AI Chatbot Framework was evaluated by measuring its performance under increasing concurrent user loads. Key metrics such as average response time, success rate, and CPU utilization were recorded to assess the system's ability to maintain efficiency and reliability during high-traffic conditions. The results are summarized in below table.

Concurrent Users	Avg Response Time	Success Rate	CPU Usage
10 Users	0.92 s	93%	18%
50 Users	1.35 s	90.7%	34%
100 Users	1.78 s	89.4%	52%
150 Users	2.21 s	86.1%	68%
200 Users	2.67 s	84.7%	81%

V. RESULTS

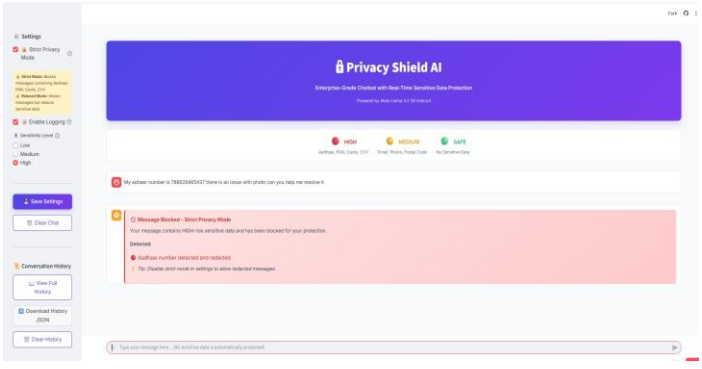

Fig 4: Strict mode High sensitive data

This figure shows the alert screen displayed in Strict Mode when highly sensitive data is detected, and nothing is saved.

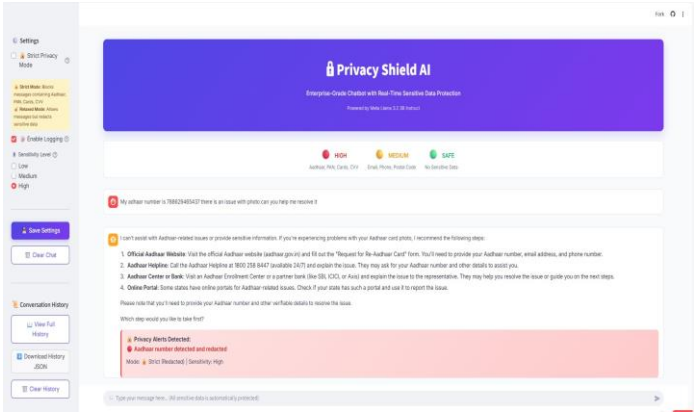

Fig 5: Relaxed mode High sensitive data

This figure shows the screen in Relaxed Mode where high sensitive data is detected, and the private details are masked before sending the text to the LLM.

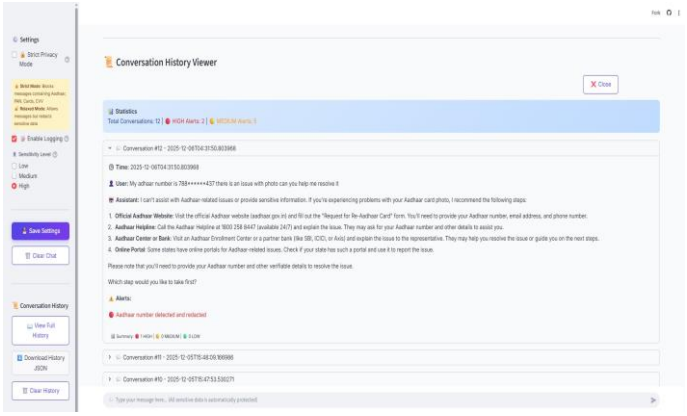

Fig 6: Chat history

The Chat History page stores only low- or medium-sensitive data in masked form, ensuring that personal information is safely redacted while retaining relaxed-mode conversation logs.

VI. CONCLUSION

The proposed Privacy-Preserving AI Chatbot Framework provides a complete solution to secure user information while using large language models. Combining real-time regex-based PII detection, confidence-based sensitivity tagging, and dynamic redaction/masking, it will ensure sensitive information is never exposed in raw form. Dual privacy modes, Strict and Relaxed, will provide users with control over data handling-either in-memory processing without storage or secure masked logging for auditing and analyses. Most importantly, LLM responses are processed to ensure no personal data ever leaks back to the user. In essence, this framework reduces the risks of privacy breach, unauthorized access, and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

misuse of personal information while ensuring seamless and privacy-aware conversational AI interactions..

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.

[2] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and N. Papernot, "Extracting Training Data from Large Language Models," Proceedings of the 30th USENIX Security Symposium, 2021.

[3] J. Lee, A. Kumar, and P. Torres, "Privacy Risks in Conversational AI Systems," IEEE Security & Privacy, 2022.

[4] S. Mohan and R. Gupta, "Analysis of Data Retention Practices in Cloud-Based Chatbots," ACM Digital Threats Journal, 2021.

[5] R. Krishna and H. Patel, "Pattern-Based PII Detection Using Regular Expressions," IEEE Access, 2020.

[6] L. Zhang and T. Wu, "Limitations of Rule-Based Sensitive Data Identification," Information Processing & Management, 2023.

[7] A. Nadeem and F. Hussain, "Contextual Privacy Detection Challenges in NLP Pipelines," ACL Findings, 2022.

[8] Microsoft Presidio Team, "Presidio: Open-Source Framework for PII Detection and Redaction," Microsoft Research, 2021.

[9] Google Cloud, "Data Loss Prevention API: Technical Report," Google Research, 2020.

[10] N. Papernot et al., "Advances in Differential Privacy for Large Language Models," ICLR, 2021.

[11] A. Bose and Y. Lin, "Security Analysis of Conversation Logging Mechanisms in AI Assistants," USENIX Security, 2022.

[12] OpenAI Research, "Private Mode and Chat Retention Policies in LLMs," Technical Report, 2023.

[13] N. Carlini et al., "Membership Inference and Training Data Extraction Attacks on LLMs," USENIX Security Symposium, 2021.

[14] P. Sharma, K. Verma, "Hybrid PII Detection Framework for Real-Time Text Streams," IEEE Transactions on Information Forensics and Security, 2023.

[15] A. Roy and M. Srinivasan, "Privacy-by-Design Frameworks for AI Chat Systems," Journal of AI Ethics, 2024.