

PrivChat-PII-Detection

PrivChat is a privacy-preserving chatbot system designed to detect Personally Identifiable Information (PII) in user inputs and sanitize them before interacting with a local Large Language Model (LLM). This document outlines the full process, architecture, detection strategies, and evaluation results based on curated test cases.

1. System Architecture

The system comprises the following components:

- FastAPI-based backend for processing and routing requests.
- PII Detection module combining spaCy NER and regex-based entity recognition.
- LLM Interface via Ollama using TinyLLaMA models.
- Web-based frontend using HTML, CSS, and JavaScript for interactive PII visualization and chat history.

2. Detection Pipeline

Upon receiving user input, the backend performs the following steps:

1. Apply spaCy NER using the `en_core_web_sm` model.
2. Apply custom regex patterns for known PII types (Email, Phone, PAN, Aadhaar, etc.).
3. Post-process to resolve conflicts and context-based misclassifications (e.g., Aadhaar mistaken for card number).
4. Annotate detected entities and sanitize the text before sending it to the LLM.

3. Entity Labels & Priorities

The following entity types are supported:

- Contact: EMAIL, PHONE
- Identity: ID_PAN, ID_AADHAAR, SSN, ID_PASSPORT, STUDENT_ID, DL_NUMBER
- Financial: CARD_NUMBER, CARD_SUFFIX, BANK_ACCOUNT
- Location/Org: ORG, VEHICLE_REG, IP_ADDRESS, GPE
- Others: DATE

A label priority system resolves overlaps, ensuring more sensitive PII is prioritized.

4. Frontend Interface

The web interface features:

- Sidebar for multiple chat threads and renaming support
- PII highlights using color-coded tooltips
- LLM response rendering per input
- Badges summarizing detected PII types
- Warning indicators when PII is found

5. LLM Integration

The system uses a locally deployed TinyLLaMA model via Ollama. Before any prompt is sent to the LLM, detected PII is replaced with placeholder tokens (e.g., `[[EMAIL]]`), ensuring user privacy.

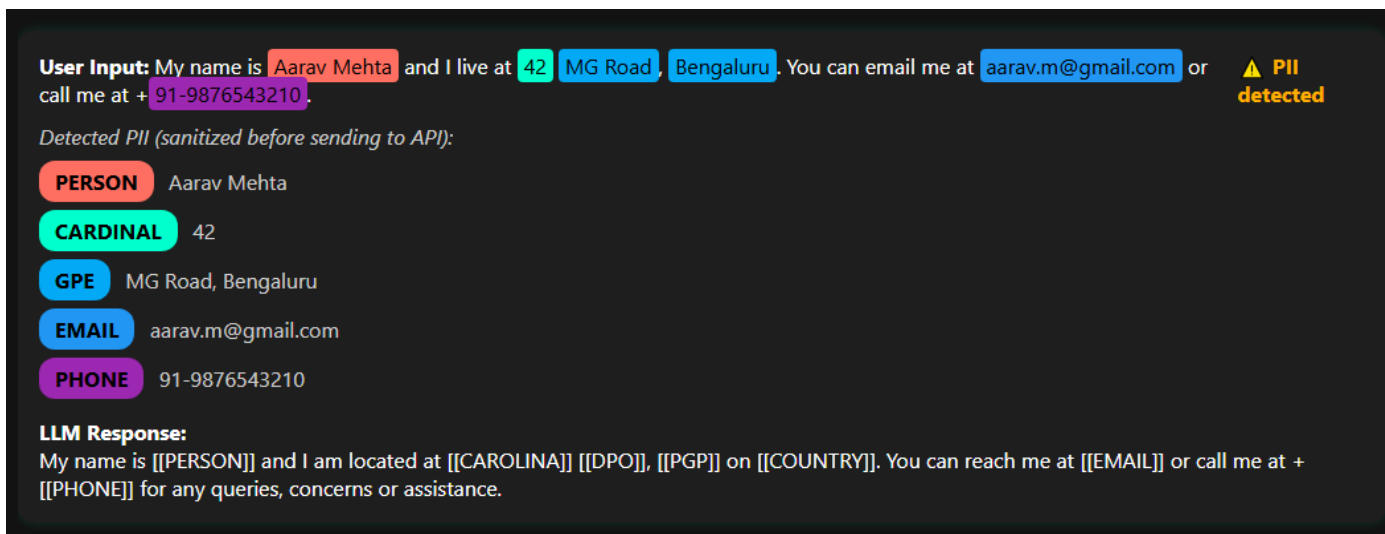
6. Evaluation Setup

A suite of test prompts was used to validate the system. These prompts include a mix of natural conversations, entity-rich texts, and non-PII queries to test precision and robustness.

7. Test Case Results

1) User Input :- My name is Aarav Mehta and I live at 42 MG Road, Bengaluru. You can email me at aarav.m@gmail.com or call me at +91-9876543210.

UI Response:-



The screenshot displays the web interface's response to a user input. The user input is shown at the top, with PII highlighted in colored boxes: 'Aarav Mehta' (red), '42' (green), 'MG Road' (blue), 'Bengaluru' (blue), 'aarav.m@gmail.com' (blue), and '+91-9876543210' (purple). A yellow warning icon and 'PII detected' text are on the right. Below the input, the detected PII is listed with labels: 'PERSON' (Aarav Mehta), 'CARDINAL' (42), 'GPE' (MG Road, Bengaluru), 'EMAIL' (aarav.m@gmail.com), and 'PHONE' (91-9876543210). The LLM response is shown at the bottom, with the detected PII replaced by placeholder tokens: `[[PERSON]]`, `[[CAROLINA]]`, `[[DPO]]`, `[[PGP]]`, `[[COUNTRY]]`, `[[EMAIL]]`, and `[[PHONE]]`.

User Input: My name is Aarav Mehta and I live at 42 MG Road, Bengaluru. You can email me at aarav.m@gmail.com or call me at +91-9876543210. **PII detected**

Detected PII (sanitized before sending to API):

- PERSON** Aarav Mehta
- CARDINAL** 42
- GPE** MG Road, Bengaluru
- EMAIL** aarav.m@gmail.com
- PHONE** 91-9876543210

LLM Response:
My name is `[[PERSON]]` and I am located at `[[CAROLINA]]` `[[DPO]]`, `[[PGP]]` on `[[COUNTRY]]`. You can reach me at `[[EMAIL]]` or call me at + `[[PHONE]]` for any queries, concerns or assistance.

2) Contact Sarah Khan at sarah.khan@outlook.com or 9876543210 during office hours.

User Input: Contact Sarah Khan at sarah.khan@outlook.com or 9876543210 during office hours. ⚠ PII detected

Detected PII (sanitized before sending to API):

- PERSON** Sarah Khan
- EMAIL** sarah.khan@outlook.com
- PHONE** 9876543210

LLM Response:
To reach the customer service representative at [COMPANY] during office hours, you can contact them by phone at (ADDRESS) extension (TELEPHONE NUMBER). Alternatively, you can send a message through email to [EMAIL] with your inquiry or complaint. Finally, if you prefer to leave a voice message at their office phone number, please do so during business hours, i.e., Monday to Friday from 9:00 AM to 6:00 PM (GMT).

3) I graduated from IIT Bombay in 2021 and my roll number was 2018CS1034.

User Input: I graduated from IIT Bombay in 2021 and my roll number was 2018CS1034. ⚠ PII detected

Detected PII (sanitized before sending to API):

- ORG** IIT Bombay
- DATE** 2021
- STUDENT_ID** 2018CS1034

LLM Response:
Graduated from [ORG] on [[DATE]] with Roll Number [[STUDENT_ID]].

4) His driving license number is DL-0420110143211 and his vehicle number is KA03MP4567.

User Input: His driving license number is DL-0420110143211 and his vehicle number is KA03MP4567.

⚠️ PII detected

Detected PII (sanitized before sending to API):

DL_NUMBER DL-0420110143211

VEHICLE_REG KA03MP4567

LLM Response:

His driving license number is "DL NUMBER" and his vehicle registration number is "VEHICLE REGISTRATION NUMBER". Note: This format depends on the specific country or region where you are searching for a driver's license, so be sure to check the relevant details in the official government website before using it.

5) John's Social Security Number is 123-45-6789

User Input: John's Social Security Number is 123-45-6789.

⚠️ PII detected

Detected PII (sanitized before sending to API):

PERSON John's

SSN 123-45-6789

LLM Response:

Your Social Security Number (SSN) is *****.

6) I work at Infosys as a software engineer with a salary of ₹12,00,000 per annum.

User Input: I work at Infosys as a software engineer with a salary of ₹12,00,000 per annum.

⚠️ PII detected

Detected PII (sanitized before sending to API):

ORG Infosys

MONEY 12,00,000

LLM Response:

As a software engineer at [[ORG]], you receive a monthly salary of approximately INR [[CURRENCY_SYMBOL]]. This amount is based on your qualifications, experience, and the position you hold. The exact amount may vary depending on various factors such as industry, location, and other benefits offered by the company.

7) Book a flight for Riya Verma from Delhi to London on 20th August via Vistara. Her passport number is N7654321.

User Input: Book a flight for Riya Verma from Delhi to London on 20th August via Vistara. Her passport number is N7654321.

▲ PII detected

Detected PII (sanitized before sending to API):

PERSON Riya Verma

GPE Delhi, London

DATE 20th August

ORG Vistara

ID_PASSPORT N7654321

LLM Response:

To book the flight, please provide the following information: Personal Details: - Name: [PERSON] - Phone Number: [PHONE NUMBER] - Date of Birth: [DATE OF BIRTH] - ID Card/Passport Number: [ID CARD/PASSPORT NUMBER] Destination: - Airport Code: [[GPE]] - Destination Airport Code: [[GPD]] Flight Details: - Flight Number: [FLGHT_NUMBER] - Departure Time: [[TIME]] - Arrival Time: [[TIME]] - Carrier Name: [[ORG]] Please confirm the details via email or phone. We will work with you to finalize your travel arrangements. Thank you for choosing our services!

8) Payment was made using card number 4111 1111 1111 1111 expiring 08/26

User Input: Payment was made using card number 4111 1111 1111 1111 expiring 08/26

▲ PII detected

Detected PII (sanitized before sending to API):

CARD_NUMBER card number 4111 1111 1111 1111

DATE 08/26

LLM Response:

I do not have access to cards or banking transactions. Please provide me with the required information: 1. Card number (starting with a '4'): this is the last four digits of the card's last four digits. 2. Expiration date: please make sure to provide the full expiration date, including the year, such as "09/2023". 3. Date of purchase: if you purchased goods or services using this card, provide the date it was used. 4. CARD_TYPE (e.g. American express, mastercard): this is the type of card you were charged with. Mastercard and visa cards use different security codes for payment, so make sure to provide the correct code. 5. INSTALMENT: if applicable, please provide the installment number or any other relevant information about your payment, such as whether it was an in-store purchase or a transfer between credit/debit cards. 6. DETAILS OF TRANSACTION: please describe the transaction thoroughly so I can understand what happened. You may include information like: amount charged, total amount paid, product/service name, date of purchase, and any additional details about your payment.

9) She used her HDFC debit card ending in 1234 to pay ₹7,500. Her account number is 123456789012.

User Input: She used her HDFC debit card ending in 1234 to pay ₹7,500. Her account number is 123456789012. ⚠ PII detected

Detected PII (sanitized before sending to API):

CARD_SUFFIX ending in 1234

MONEY 7,500

BANK_ACCOUNT account number is 123456789012

LLM Response:
She used her High Definition Fixed-Income Credit Card (HDFC) debit card (with CARD_SUFFIX) to pay the amount of ₹1,000 into her Bank Account (BANK_ACCOUNT). The transaction is recorded in the Bank's records.

10) "I'm learning Python through online tutorials – it's fun and challenging."

User Input: "I'm learning Python through online tutorials – it's fun and challenging." ✅ No PII detected

No named entities detected.

LLM Response:
Yes, that's absolutely right! Learning Python through online tutorials can be very rewarding, challenging, and exciting at the same time. Here are some reasons why: 1. Fun & Challenging Python is a fun and challenging language to learn, which can make the learning process more enjoyable than if you were to learn it in a traditional classroom setting. Online tutorials provide you with the freedom to choose what topics you want to focus on or study at your own pace. This is especially helpful for beginners who may have a hard time retaining information overnight, but enjoy learning through the process of discovery and exploration. 2. Engaging Tutors & Teachers Online Python tutorials are typically provided by professional instructors who teach the language in a way that's easy to follow and understand. With a wide range of online resources available, you can access tutors who specialize in different areas of the language, such as web development or business applications. This helps to make learning Python more accessible and inclusive for learners of all levels. 3. Exciting & Dynamic Learning Environment Online Python tutorials offer a dynamic learning environment where you can interact with other learners from around the world. This provides the opportunity to share insights, ask questions, receive feedback, and gain inspiration from others who are also learning at different speeds. It's a great way to stay motivated and engaged throughout your Python journey! 4. Efficient & Effective Learning Method Online tutorials provide efficient and effective ways to learn Python. They can be accessed at any time and from anywhere, which allows you to set your own learning schedule that fits in with your busy schedule. Additionally, they offer a flexible, self-paced format that adapts to your pace of learning, which makes it easier for you to keep up with the material at your own pace. 5. Explore New Technologies & Trends Online Python tutorials provide an opportunity to explore new technologies and trends in the Python language. You can learn about various libraries and frameworks, such as PyTorch, TensorFlow, Flask, etc., that are used for data analysis, web development, and artificial intelligence. This

11) Can you recommend some good books on leadership and productivity?"

User Input: Can you recommend some good books on leadership and productivity?"

✓ No PII detected

No named entities detected.

LLM Response:

Certainly! Here are some books on leadership and productivity: 1. "The Lean Startup" by Eric Ries - This book is a must-read for anyone looking to improve their startup's product development process. It provides practical tips on how to make decisions quickly based on data, while also focusing on customer needs. 2. "The 7 Habits of Highly Effective People" by Stephen Covey - This classic book on leadership is full of actionable tips and advice on how to become a more effective leader. It's a must-read for anyone looking to improve their leadership skills. 3. "Mastering the Art of Communicating with Anyone, Anytime, Everywhere" by Gary K. Page - This book provides practical communication strategies that will help you become a better communicator no matter what situation you find yourself in. 4. "The Time Management Bible" by Susan Singer - If you're looking to boost your productivity and improve your time management skills, this is an excellent resource. It covers topics such as prioritization, planning, and goal setting. 5. "Productivity: The Definitive Guide" by David Allen - This book provides a comprehensive guide on how to get the most out of your time, including tips for managing distractions, creating priorities, and maximizing productivity. I hope these recommendations are helpful! Remember that effective leadership requires more than just knowing how to talk to others. By studying these books, you'll be able to improve your communication skills, make better decisions, and become a more effective leader.

8. Observations

- Regex patterns were effective for formats like Aadhaar, PAN, SSN, and Card Numbers.
- spaCy NER performed well on names, locations, and organizations, with occasional overlaps.
- Contextual disambiguation was crucial for detecting BANK_ACCOUNT vs PHONE where number formats overlap.
- The UI efficiently displays PII alerts, grouped labels, and ensures no sensitive text is forwarded to the model.

9. Conclusion

PrivChat demonstrates robust detection of PII using hybrid NLP techniques and ensures privacy before any model inference. It is suitable for applications requiring secure and interpretable handling of sensitive user data, with future potential for extension to document-level redaction and GDPR compliance frameworks.