Proposal: Automatic Masking of Personally Identifiable Information (PII) by Large Language Models

Abstract

Large language models (LLMs) are now widely used in education, business, and personal applications, but their rapid popularity raises pressing privacy concerns. Users frequently include Personally Identifiable Information (PII) in prompts, often unintentionally. Current practice places full responsibility on users to recognize and remove such information, an approach that is neither reliable nor sustainable. This proposal recommends the development of built-in masking capabilities that automatically detect and obscure PII before storage or processing. Shifting this responsibility from the user to the system will mitigate risks of data leakage, strengthen regulatory compliance, and foster user trust in AI technologies.

Introduction

The widespread use of LLMs such as Chat GPT or Gemini has increased communication and productivity across industries. However, this creates vulnerabilities, particularly when PII is entered into prompts. Examples include names, phone numbers, place of origin, and such. At present, users are expected to self-censor or manually redact such inputs.

This expectation is unrealistic. Many users lack expertise in privacy standards, while even trained professionals can overlook sensitive details. High-profile cases, such as corporate employees inadvertently submitting confidential information to these LLMs, illustrate the risks of relying solely on user diligence. Current strategies, like issuing reminders have proven insufficient. A system-level solution is necessary: automatic PII detection and masking performed directly within the LLM.

Project Description

The proposed system integrates a PII detection-and-masking module at the interface level of LLMs. Its primary functions would include:

1. **Detection** – Identifying PII in real time through separately trained LLM

2. **Masking** – Replacing detected items with placeholders such as “[REDACTED]” or sudo information

3. **Transparency** – Informing users when masking has occurred and providing an option to override if the information is essential.

4. **Continuity** – Ensuring that masking does not interrupt workflow or reduce the model's usefulness.

This approach addresses privacy risks at their source while preserving the functionality of LLM applications.

Rationale

Automatic PII masking is important for several reasons:

- **Privacy Protection**: Reduces the likelihood of accidental exposure by minimizing reliance on individual users

- **Regulatory Compliance**: Supports organizations in meeting legal requirements under data privacy regulations such as HIPAA.

- **Trust in Adoption**: Encourages broader use of AI tools by embedding privacy principles that demonstrate responsibility in LLM usage.

Plan of Work

Scope: Initial development will prioritize detection of the most common PII categories: names, contact information, place of origin, current place of living, government documented information and such.

Methods:

- Conduct a review of existing PII detection technologies.
- Develop a hybrid detection system combining rule-based and statistical models using Hugging Face.
- Pilot the module across varied input types (e.g., emails, customer service chats, and personal documents).

Problem Analysis: The core issue is a mismatch between user expectations and system design. Users assume AI tools are safe by default, while current systems place the task of privacy protection entirely on individuals. This proposal fixes that gap by embedding automatic safeguards.

Conclusion

LLMs offer incredible potential but also introduce new privacy risks. Requiring users to identify and remove PII is impractical and unsafe. By incorporating automatic detection and masking capabilities, AI developers can create systems that protect individuals, assist organizations in meeting compliance standards, and build public trust in AI technology. Privacy is not an optional feature; it is a necessary standard for the responsible deployment of large language models.