1. Introduction to the Domain / Motivation for the Problem

At Ai4Privacy, we are committed to building global solutions to enhance privacy protections in the era of Artificial Intelligence. With the rapid advancement of AI assistants and Large Language Models (LLMs), there's an increasing need to ensure that sensitive and personally identifiable information (PII) is not inadvertently exposed or mishandled. Protecting user privacy is not just a regulatory requirement but also a trust imperative in today's data-driven world.

The motivation behind this problem stems from the necessity to develop robust solutions that can detect and mask PII across multiple languages and jurisdictions. By creating effective PII masking models, we can enhance privacy protections in various applications such as chatbots, customer support systems, email filtering, and data anonymization processes. Solving this problem is crucial for maintaining user trust, complying with global privacy regulations, and enabling the safe deployment of AI technologies that interact with personal data.

---

2. Problem Statement

The challenge for participants in this hackathon is to develop a machine learning model capable of accurately detecting and masking personally identifiable information (PII) in text data across multiple languages and locales. Using the provided synthetic dataset, participants will train and evaluate models to automatically identify and redact 17 types of PII within natural language texts. The goal is to create a solution that can be integrated into various systems to enhance privacy and data protection without compromising the usability of the data.

---

3. Explanation of the Data

Participants will have access to the **world's largest open dataset for privacy masking**, specifically designed for training and evaluating PII detection and masking models.

- **Total Entries**: 406,896
- **Total Tokens**: 20,564,179
- **Total PII Tokens**: 2,357,029
- **Number of PII Classes**: 17 in the public dataset (63 in the extended dataset)
- **Locales:** 8

Dataset Contents:

Each entry in the dataset includes:

- **source_text**: Natural language text containing synthetic PII.

- **target_text**: The masked version of the source text, where PII is replaced with placeholders (e.g., `[USERNAME]`, `[TIME]`).
- **privacy_mask**: Detailed labels of PII, including the value, start and end positions in the text, and the type of PII.
- **span_labels**: Exact character spans of PII within the text.
- **mberttokens**: Tokens generated using multilingual BERT tokenization.
- **mbert_bio_labels**: BIO (Beginning, Inside, Outside) labels corresponding to the tokens for sequence tagging tasks.
- **id**: Unique identifier for each entry.
- **language**: The language of the text (English, Italian, French, German, Dutch, Spanish).
- **locale**: The regional locale associated with the text (e.g., United Kingdom, United States, Italy).
- **split**: Indicates whether the entry is part of the training or validation set.

Key Facts:

- The dataset is synthetic and generated using proprietary algorithms—no real personal data is used.
- It covers 6 languages with localization in 8 jurisdictions.
- Designed for tasks like token classification and text generation.

---

4. Potential Solution

Participants are expected to develop a machine learning solution that can perform PII detection and masking effectively. Here's what the solution might entail:

- **Model Training**:
  - Utilize the provided dataset to train a token classification model.
  - Choose an appropriate pre-trained model (e.g., BERT, RoBERTa, mDeBERTa) and fine-tune it on the dataset.
  - Implement sequence tagging techniques to identify PII tokens within the text.
- **Scope of the MVP**:
  - **Essential Features**:
    - Accurate detection of PII across multiple languages.
    - Masking of identified PII in the text.
  - **Optional Enhancements**:
    - Develop a user interface (UI) or an API for the model for a specific use-case
    - Handle longer text sequences efficiently using extension such as BELT: https://www.reddit.com/r/MachineLearning/comments/19edzov/project_belt_bert_for_longer_texts/
    - Post the developed model on HuggingFace for community use and feedback.

- **Technical Considerations**:
    - Use transformer-based architectures suitable for token classification.
    - Employ multilingual models to handle the diversity of languages in the dataset.
    - Implement evaluation metrics to assess model performance (e.g., Precision, Recall, F1-Score).

Participants should aim to create a solution that is both effective and efficient, potentially ready for integration into applications like chatbots, data anonymization tools, or content moderation systems.

---

5. Additional Information

Useful Resources:

- **Dataset Access**:
    - **AI4Privacy PII Masking Dataset**: https://huggingface.co/datasets/ai4privacy/pii-masking-400k

- **Sample Submission**:
    - **Piiranha-v1 Model**: https://huggingface.co/iiiorg/piiranha-v1-detect-personal-information
        - An example of a model trained for PII detection, including performance metrics and training details.

- **Guides and Documentation**:
    - **Information on the PII Masking Problem**: https://p5y.org
    - **HuggingFace Token Classification Task**: https://huggingface.co/docs/transformers/tasks/token_classification