PII Redaction for Legal Documents – Project Documentation

Overview

This project focuses on automating **Personally Identifiable Information (PII) redaction** within legal documents using **Large Language Models (LLMs)** . PII redaction is an essential step in safeguarding sensitive information and ensuring organizations remain compliant with major privacy regulations such as **GDPR** , **CCPA** , and industry-specific confidentiality requirements. Legal documents often contain names, addresses, identification numbers, financial details, and other sensitive entities that must be masked before external sharing or archival.

Historically, PII detection relied heavily on traditional **Named Entity Recognition (NER)** models. While effective to some degree, these systems often struggle with context-dependent PII or domain-specific terminology found in legal contracts. With the advancement of modern LLMs, redaction tasks can now be handled more accurately—provided the prompts are properly engineered and the process is systematically structured. This project explores how different prompt designs impact detection accuracy, performance, and reliability.

Data Description

The project utilized four specific test legal documents: **Test_A** , **Test_C** , **Test_D** , and **Test_F** . These documents were sourced from an Excel file, loaded, and subjected to a preprocessing stage for initial data cleaning.

Crucially, each document was accompanied by a **target label** —a curated list of all true PII instances within that document. This ground truth was essential for accurately comparing and measuring the performance of the various prompt engineering strategies.

Prompt Engineering Approach

The project evaluates how prompt structure affects PII detection accuracy. The exploration began with **simple direct prompts** , such as instructing the model to “identify and redact PII,” and gradually evolved toward more sophisticated designs.

Enhancements included:

- Step-by-step task decomposition (“first identify PII, then redact”).
- Explicit lists of PII categories to track.
- JSON-structured output requirements for easier evaluation.
- Few-shot examples demonstrating correct redaction behavior.
- Error-checking instructions to reduce hallucinations or omissions.

This iterative process clarified how prompt clarity, constraints, and output formatting contribute to improved performance and consistency.

Model

The project uses the **Gemini free-tier model** as the inference engine. Despite being a cost-efficient model, it proves capable of performing entity recognition and redaction with reasonable accuracy when properly guided by well-designed prompts. Prompt engineering plays a key role in compensating for model limitations inherent in free-tier or lightweight LLM variants.

Evaluation Methodology

Model outputs were compared against ground-truth PII labels using standard information-retrieval metrics:

- **Precision** – proportion of correctly detected PII among all detected items
- **Recall** – proportion of ground-truth PII successfully identified
- **F1 Score** – harmonic mean of precision and recall, representing overall detection quality

These metrics allow for consistent comparison across different prompt strategies and highlight trade-offs between over-redaction and missed PII. The evaluation results guide which prompt patterns yield the best balance for legal-domain use cases.