PII Redaction for Legal Documents

1. Overview

This project builds an automated PII redaction system for legal documents using LLMs, with a core focus on how prompt engineering directly impacts extraction accuracy. Traditional NER models often miss context-dependent PII, but LLMs can perform significantly better when guided by strict, well-structured prompts. The goal is to evaluate the effect of different prompt designs on precision, recall, and reliability, particularly using the Gemini free-tier model, where strong prompt constraints are crucial due to model limitations.

2. Prompt Engineering

The project follows an iterative prompt-engineering workflow, continuously enhancing structure, specificity, and constraints to improve model behavior.

- More constraints: Tight instructions greatly reduce false positives and random outputs.
- Structured formats enable objective evaluation: JSON schemas prevent formatting drift and improve metric computation.
- Examples reduce confusion: Helps models correctly detect rare legal PII (ID types, contract numbers).
- Role-based instructions increase compliance: "Extraction engine" framing minimizes conversational behavior.

Version	What Changed	Key Additions / Techniques	Impact on Performance
v1 – Baseline Prompt	Simple instruction: "Extract PII."	None – minimal constraints	• Inconsistent categories • Missing spans • Invalid JSON
v2 – Structured Output Prompt	Introduced strict formatting	• JSON-only output • Fixed PII categories • Start/end character indexing • No explanations allowed	• More consistent outputs • Fewer formatting errors • Better evaluation accuracy
v3 – Definitions + Examples Prompt	Added clarity + training-like examples	• Formal definitions per category • Real legal-domain examples • Tricky cases explained (embedded IDs) • Strict schema enforcement	• Reduced ambiguity • Lower hallucination rate • Higher precision
v4 – Extraction Engine Prompt	Switched to machine-instructio n tone	• Step-by-step extraction rules • Zero-hallucination policy • Explicit taxonomy • Exact span + casing enforcement	• Major recall improvement • Highly controlled outputs
v5 – Final Optimized Prompt	Combined all prior improvements	• Mandatory extraction for all categories • No rewriting/normalization • Unified strict JSON schema • Zero commentary rule	• Best precision + recall balance • Most stable and reliable version

3. Dataset

Four legal documents (A, C, D, F) were extracted from an Excel dataset. Each contains ground-truth PII labels, enabling quantitative evaluation of different prompt versions.

4. Evaluation Method

Each prompt version was tested using Precision, Recall, and F1 Score. Ground-truth labels made it possible to measure exactly how each prompt revision improved performance. Prompt engineering compensated for: shallow context, higher hallucination probability, and inconsistent JSON formatting. Well-designed prompts significantly closed the performance gap between free-tier and larger models.

5. Conclusion

Prompt engineering, not the model alone, determines PII extraction accuracy. Structured constraints, definitions, and strict schemas turned an unreliable baseline into a high-precision redaction system, even using a free-tier LLM.

6. Future Work
    a. Few-shot prompt variants
    b. Self-consistency prompting with a hidden chain-of-thought
    c. Prompt ensembles for robustness