1. Team Details

- **Team Name:** Viksit
- **Team Members:**
    - Arnav Majithia
    - Ishan Rai
    - Ritika Sinha
    - Prince Agrahari

2. Problem Understanding and Scope

- Problem Statement:

  In today's digital landscape, users frequently encounter complex privacy policies written in legal jargon. This makes it difficult to:

    - · Understand what personal data is being collected,
    - · Know how long it's stored,
    - · Determine whether it is being shared, and
    - · Exercise rights like opt-out or data deletion
    - · With increased focus on privacy (e.g., GDPR, HIPAA, PDPB), transparency and user education are crucial.

- **Target Documents & Formats:**
    - **Medical Documents:** Patient intake forms, fictional medical records, and health insurance forms.
    - **Financial Documents:** Financial statements and related forms.
    - **Identification:** ID cards and other official identification documents.
    - **Specialized Data:** EDI X12 data streams for health transactions[1].
- **Types of Identifiable Data (PII) to Detect & Redact:**
    - **Personal Details:** Names (e.g., Carlos E. Rodriguez) [2222], Dates of Birth (e.g., June 22, 1967) [3333], Gender [4444], Ethnicity[5555].
    - **Contact Information:** Addresses (e.g., 1442 Palm Tree Blvd, Miami, FL 33176) [6666], Phone numbers [7777], Email addresses[8888].
    - **Sensitive Identifiers:** Social Security Numbers (SSN) [9999], Member IDs [10101010], Group Numbers [11111111], National Provider Identifiers (NPI)[12121212], and barcodes.
    - **Medical Information:** Diagnoses [13131313], ICD-10 Codes [14141414], and Physician names[15151515].
    - **Visual Data:** Handwritten notes and signatures.
- **User Personas / End Users:**
    - **Hospitals & Healthcare Providers:** To anonymize patient records for research or sharing while ensuring HIPAA compliance.

    - **Legal Teams:** To redact sensitive information from documents during the discovery phase of litigation.
    - **Government Offices:** To protect citizen data when processing public records or forms.
    - **Insurance Companies:** To secure client data in claims processing and internal analysis.

3. Proposed Solution & Approach

- High-Level Architecture (with Diagram):

Our solution will follow a sequential pipeline designed for accuracy and efficiency.

1. **Input:** The user uploads a document (PDF, JPG, PNG).
2. **OCR Engine:** An Optical Character Recognition (OCR) model extracts both the text and its coordinates from the document.
3. **Hybrid Detection:**
    - **NER Model:** A Named Entity Recognition (NER) model scans the extracted text for PII/PHI.
    - **Object Detection Model:** A computer vision model (e.g., YOLO) simultaneously scans the document image to find non-textual elements like signatures and barcodes.
4. **Redaction:** The system uses the coordinates of the detected sensitive data to draw redaction boxes over them.
5. **Output:** The system generates a redacted PDF/image and an accompanying log file.

- **AI/ML Models Considered:**
    - **OCR:** Tesseract for a baseline, with consideration for more advanced models like LayoutLMv3 or Donut, which understand document layout.
    - **PII/PHI Detection:** Transformer-based models (e.g., spaCy with custom rules or a fine-tuned BERT model) for NER.
    - **Signature/Object Detection:** A YOLO (You Only Look Once) model fine-tuned on the **SignverOD** dataset for detecting signatures and other visual elements.
- Data Strategy:

The primary dataset for training our visual element detector will be the SignverOD dataset, which is specifically designed for signature object detection. For PII/PHI text detection, we will use a combination of publicly available NER datasets and the provided sample documents (e.g., Carlos_E_Rodriguez.pdf, Fake Test Data.xlsx - X12 - 278.csv) to create a custom, annotated evaluation set. This will ensure our model performs well on the target document types.

- Innovation / Unique Selling Point (USP):

Our key innovation is a hybrid detection model. Unlike solutions that run OCR and NER sequentially, our approach runs textual NER and visual object detection in parallel. This allows the system to cross-reference findings—for example, confirming that a detected handwritten element is indeed a signature located in a designated signature box—which significantly reduces false positives and improves redaction accuracy.

4. User Experience & Accessibility

- Intended UI/UX Design (if applicable):

We will develop a simple and intuitive drag-and-drop web interface. Users can upload their files directly in the browser, see a preview of the redactions, and download the final secure document. For power users and integration, a Command-Line Interface (CLI) will also be offered for batch processing.

- **Input & Output Format Expectations:**
    - **Input Formats:** PDF (single and multi-page), PNG, JPEG, TIFF.
    - **Output Formats:** Redacted PDF or image files. A JSON log file detailing the type and location of each redacted element will be generated for auditing purposes.
- **Accessibility / Ease of Use Considerations:**
    - **Non-Technical Users:** The UI will be designed with simplicity in mind, requiring no prior technical knowledge.
    - **Multilingual Support:** While the initial model will focus on English, the architecture will be flexible to incorporate multilingual models in the future.
    - **Low Resource Settings:** The solution will be offered as a cloud service to ensure that users do not need powerful local hardware to perform redactions.