# **Week 2 Report: Contact Document Explainer (Invoice/Contact Info Parser)**

**Prepared by:** Model Development Team  
**Date:** [25/June/2025]

**📌 Objective**

To develop an AI tool that processes unstructured contact documents (e.g., invoice PDFs, email threads, scanned letters) and extracts:

* Contact person’s name, company, and role
* Phone numbers and emails
* Addresses (billing/shipping)
* Any contextual remarks or inquiries
* Structured data for downstream use

**🧱 System Architecture**

PDF/Text Input ➝ OCR (if scanned) ➝ Preprocessing ➝ LLM-based Information Extraction ➝ JSON Output

**🧩 Components & Implementation**

**1. Input Handling**

* Supports .pdf, .txt, and OCR on images/scanned documents
* Integration with Google Colab or local directory

**2. Text Preprocessing**

* Cleans document text (removes headers/footers, noise)
* Splits into paragraphs for context-aware extraction

**3. Information Extraction Engine**

* LLM used: OpenAI GPT-4-turbo or Mistral (depending on config)
* Prompt-based parsing for entities like:
  + Name
  + Email
  + Phone number
  + Company
  + Designation
  + Billing & shipping addresses
  + Any instructions/queries

**4. Output Format**

* JSON dictionary containing structured fields
* Can be exported as:
  + .json for integration
  + .csv for tabular review
  + .txt summary report

**5. Fallback Extraction**

* Regex-based parser for key fields:
  + Email
  + Phone
  + Names (via capitalized patterns)
* Triggered if LLM fails or returns incomplete results

**6. Security & Privacy**

* No data storage — all parsing is done in memory
* Can be integrated with AES-256 encryption pipeline from previous project

**✅ Achieved Features**

| **Feature** | **Status** |
| --- | --- |
| Multi-format document input | ✅ Done |
| OCR for scanned PDFs | ✅ Done |
| LLM-based entity extraction | ✅ Done |
| Fallback regex-based parsing | ✅ Done |
| JSON/CSV/TXT export | ✅ Done |
| In-memory secure processing | ✅ Done |
| Modular notebook structure | ✅ Done |

**🧠 Example Prompt Used**

text

CopyEdit

You are an intelligent assistant that extracts contact information from business documents. Return results in a structured JSON format with the following keys: name, email, phone, company, role, billing\_address, shipping\_address, message\_context.

**🧾 Final Notes**

* Fully functional prototype notebook for document understanding
* Can be integrated into MSME or legal CRM tools
* Compatible with the existing encryption + summarization modules