Draft Assignment for AI Intern

Assignment Title:
AI-Powered Document Structuring & Data Extraction Task

Timeline: 2 days

## Objective

You are required to design and implement an AI-backed solution that transforms an unstructured document into a structured Excel output.

## Assignment Requirements

### 1. Input → Output Transformation

Using the file “Data Input.pdf”, build code (Python preferred, but any language is acceptable) that automatically converts the content into the structured format demonstrated in “Expected Output.xlsx.”

Your solution should:

- Accurately extract all information from the PDF
- Identify logical relationships between elements
- Format the extracted data into tabular Excel output
- Do not pre-define the keys in the code, let LLM determine the key.

### 2. Key:Value Relationship Detection

The code must:

- Determine key:value pairs within unorganized or semi-structured text
- Place each extracted pair into appropriate Excel columns
- Add additional contextual notes (pulled from the PDF content) into a “Comments” column
- Maintain clarity and logical coherence in determining what constitutes "context"

### 3. Complete Data Capture

You must ensure:

- 100% of the content in “Data Input.pdf” is captured in the Excel output
- Nothing is lost, nothing is summarized, or nothing is omitted
- Multi-line or complex textual structures are faithfully represented
- We will test your submission against other page text readable documents (no need of OCR)

### 4. Preserve Original Language

Where possible:

- Retain the exact original wording, sentence structure, and phrasing from the PDF
- Avoid paraphrasing unless required to form a clean key:value pair
- Do not introduce new information
- Feel free to use LLM of your choice

### 5. Deliverables

At the end of the assignment, please provide:

1. GitHub repository containing:

- Source code
- README with usage instructions
- Any model files or dependencies

2. A live demo hosted on a temporary domain or we can do a screenshare call and run through the code output live.

3. Final generated “Output.xlsx” for evaluation

### 6. Communication Requirement

Before starting the task:

> Please reach out with any questions or clarifications to ensure the process remains effective and organized.