**Fit.Fest 2025: Hackathon**

Problem Title: **Engineering using Gen AI: Agentic Data Extraction & processing**

Problem Description: PDFs like manuals, research papers, and legal files hold a lot of valuable info—but it's hard for AI systems to extract it well. Documents like grid manuals, maintenance logs, and compliance records often exist in the form of such complex PDFs. These files are not just text-based they contain:

* Scanned images (requiring OCR)
* Tables embedded as images or diagrams
* Design diagrams (e.g. Sequence diagrams, block diagrams, flowcharts)
* Bold/italicized highlights, notes, and handwritten annotations
* Multi-column layouts and rotated pages
* Visual cues that indicate importance (e.g., warnings, captions)

**Overview:**

The goal is to intelligently extract and structure content from complex PDFs including scanned text, diagrams, tables, and multi-column layouts into meaningful chunks that can later be embedded and searched semantically.

**UniChunk – Stage 1**

In this first stage, goal was to:

* Extract clean text from each page of the PDF
* Perform OCR on embedded/scanned images
* Group the content into layout-aware 'chunks' like:
  + Headings
  + Paragraphs
  + Lists
* Save the results as a structured .json file (one per page)"

**Technologies Used**

|  |  |
| --- | --- |
|  | **Tool** |
| PDF Parsing | PyMuPDF (fitz) |
| OCR | Tesseract |
| Chunking | Custom logic using Python |
| Export | JSON |

**Stage-1 Architecture:**

