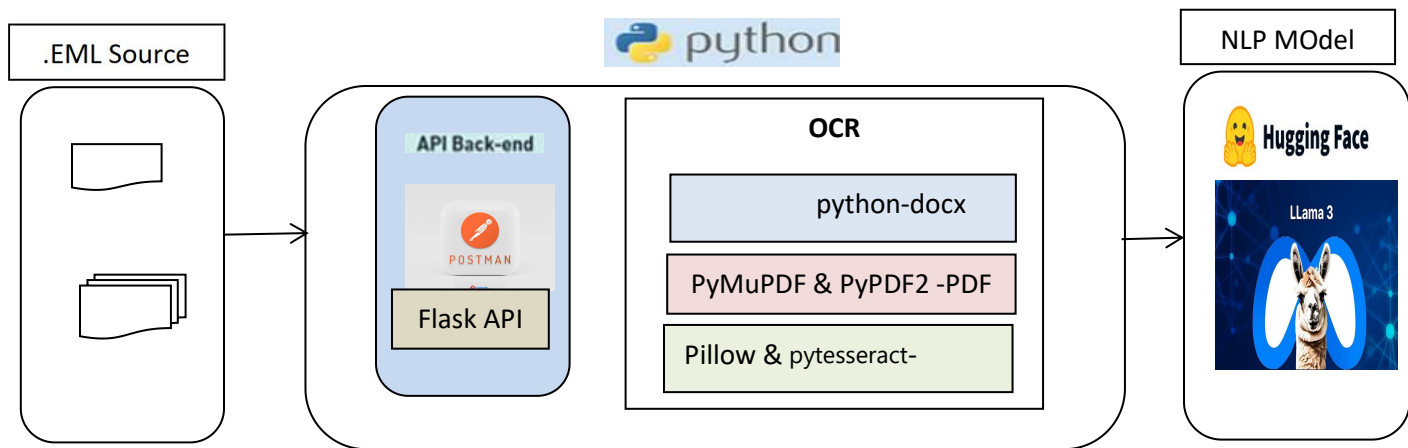


## Technology Architecture



### Flask API : Used to Processing .eml Files and Batch File Processing

**This Flask API provides two endpoints:**

- 1 Single .eml File Processing: Accepts an .eml file as input, extracts attachments, and processes them based on file type.

**End Point : /process/file - Single .eml File Processing**

- 2 Batch Processing from a Folder: Allows users to select a folder path and process multiple files at once.

**End Point 2: /process/path - Batch Processing**

**Method: POST**

**Description:**

Accepts a single .eml file as input or .eml files from folder path

Extracts attachments from the email if any .docx, .pdf, .jpeg, and .gif files.

Returns extracted text from the attachments as well email body.



dataextract.txt

The screenshot shows a REST client interface. At the top, there's a dropdown menu set to 'POST' and a text input field containing the URL 'http://127.0.0.1:5000/process/file?'. To the right of the URL is a blue 'Send' button. Below the URL bar, there are tabs for 'Params', 'Authorization', 'Headers (9)', 'Body', 'Scripts', 'Tests', and 'Settings'. The 'Body' tab is selected. Under the 'Body' tab, there are radio buttons for 'none', 'form-data' (which is selected), 'x-www-form-urlencoded', 'raw', 'binary', and 'GraphQL'. Below these options is a table with columns 'Key', 'Value', 'Description', and 'Bulk Edit'. The table has one row with a checked checkbox in the 'Key' column, the key 'file', a dropdown menu set to 'File', the value 'AU Transfer Request.eml', and a file upload icon in the 'Value' column. The 'Description' column is empty.

### Why Use OCR(Optical Character Recognition) Approach?

1. Handles Multiple File Formats (.docx, .pdf, .jpeg, .gif)
2. Uses Optimal Libraries (PyMuPDF for PDFs, pytesseract for images)
3. Supports OCR for Image-based Text Extraction
4. Stores Data in a Structured Array for Further Processing to LLM Model

This script **parses a .eml email file**, extracts **attachments**, and reads their content based on their file type: The extracted content is stored in an **array of dictionaries** for further processing.

### Use Cases :

1. **Automated Data Entry**
2. **Extracting Text from Files & Handwritten Notes**
3. **Processing PDF , JPEG , Docx Documents** (OCR for text extraction)

### Limitations:

1. **Handwriting Recognition** → Lower accuracy for messy handwriting
2. **Complex Backgrounds** → May cause OCR errors
3. **Low-Resolution Images** → Blurry images reduce accuracy
4. **Languages & Fonts** → Some languages require additional training

### LLaMA 3 Model :

Meta is working on **next-gen Llama models**

Expected to be **larger, more powerful, and efficient**

Focused on **better reasoning, coding, and multimodal capabilities**

---

### Key Features of LLaMA Models

- ✓ **Optimized for Efficiency:** Smaller models with performance comparable to larger AI models
- ✓ **Open-Source:** LLaMA 2 is freely available for commercial and research use
- ✓ **Fine-Tuning Support:** Customization for chatbots, coding, and other NLP applications
- ✓ **Transformer-Based Architecture:** Uses an improved version of **decoder-only transformers**

### OutPut From .Eml File:

```
{  
  "request_type": "***Summary:** The purpose of this document is to request an AU Transfer for account"  
}
```

Challenges :

### Achievements:

1. Successfully implemented reading of **single and batch .eml files** as input.
2. Extracted and processed **attachments** from .eml files, including .jpeg, .docx, and other formats.
3. Developed a **model that analyzes request types** and provides structured output.
4. Implemented **content extraction** from email bodies, images (.jpeg), and documents .docx).

---

### Challenges & Next Steps:

1. **Integrating an Offline LLM Model** – Downloading and running an **offline** LLM model from **Hugging Face** efficiently.
2. **Generating Quality Test Data** – Creating **diverse and high-quality datasets** for model training and validation
3. **Expanding Use Case Coverage** – Enhancing functionality to support **broader scenarios** and meet all requirements.