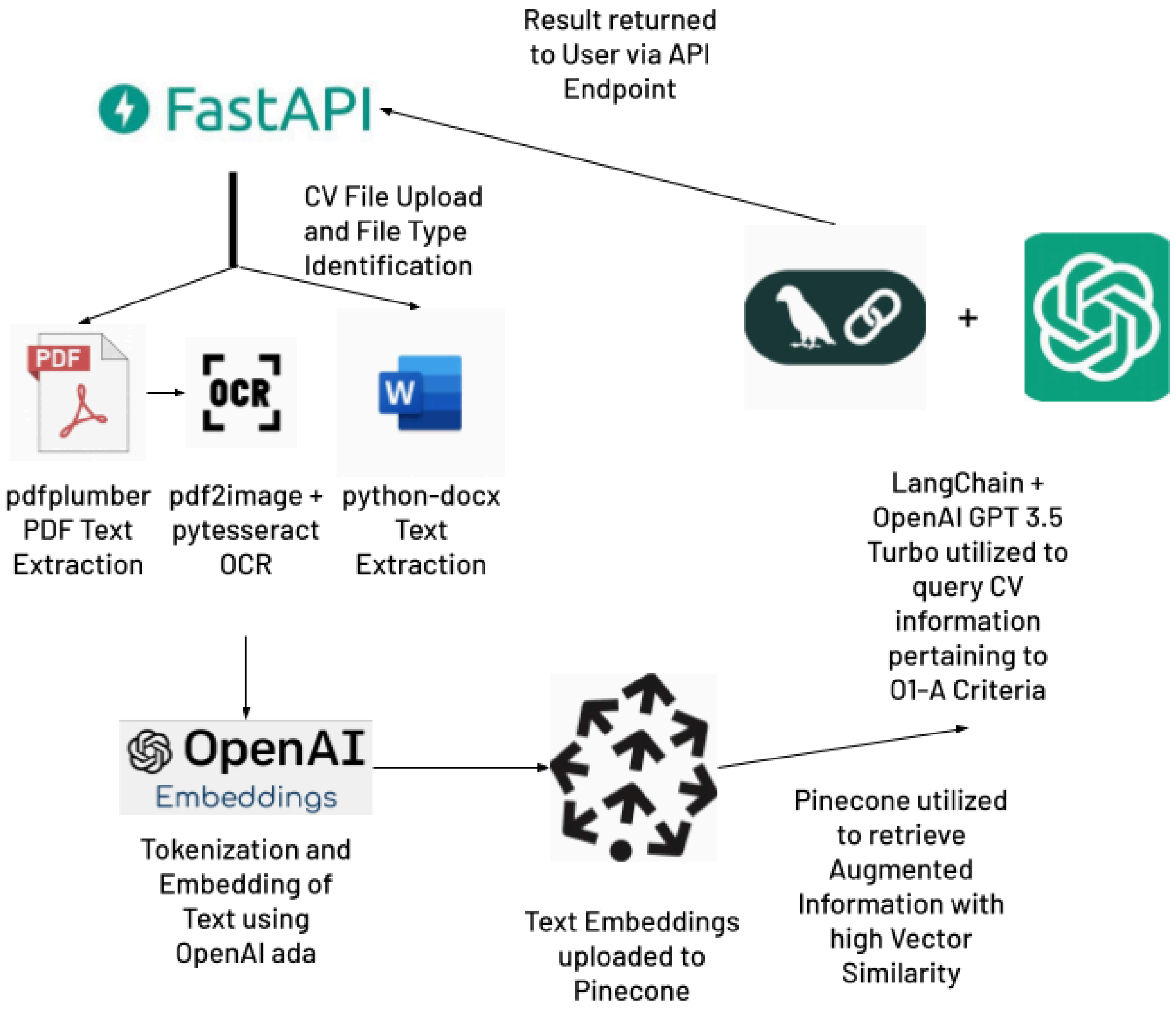
Alma System Design Document

Overall System Design Diagram:


Result returned to User via API Endpoint
FastAPI
CV File Upload and File Type Identification
PDF
OCR
W
pdfplumber PDF Text Extraction
pdf2image + pytesseract OCR
python-docx Text Extraction
+
LangChain + OpenAI GPT 3.5 Turbo utilized to query CV information pertaining to 01-A Criteria
OpenAI Embeddings
Tokenization and Embedding of Text using OpenAI ada
Text Embeddings uploaded to Pinecone
Pinecone utilized to retrieve Augmented Information with high Vector Similarity

Design Choices according to Task Components:

1. CV Parsing

- CV Input can be of three types, PDFs with extractable text, PDFs containing scanned Images, and DocX Files
- Priority here is to obtain the plain text from CV without errors, optimizing for time complexity
    - PDFs with Extractable Text - pdfplumber is reliable and provides output within O(N) time contingent on number of pages in PDF
    - DOCx with Extractable Text - pythondocx is reliable and provides output within O(N) time contingent on number of paragraphs in DOCx
    - PDFs with Scanned Image - EasyOCR is a powerful but slow library but only useful for complex images. Considering CVs are relatively more structured and standardized documents, pdf2image+pytesseract is a much faster combination for text extraction.
- Chose multiple document type support for wider accessibility, considering that this will ideally be a low barrier of cost service for a large number of users
- Chose multiple page scanning instead of single page scanning for OCR option because professionals applying for O-1A are more likely to have multiple page CVs (publications, conferences)

2. CV Assessment and Result Computation

- Within the constraints of this assessment, I do not have access to a large dataset of publicly available resumes for the specific group of professionals that likely apply to the O-1A Visa. Hence, it is very difficult for me to train a classification model of degree of criterion fulfillment for each criteria.
- The best-suited model for this project is a pre-trained transformer model. Considering that we are providing this model with additional private CV data, a Retrieval Augmented Generation pipeline is aptly suited.
- The design choices within this pipeline architecture are hence the following
    - Choice of LLM-interfacing Framework - LangChain has highest compatibility and support for different LLMs, also has an inbuilt RAG Query Method
    - Choice of database - Pinecone as a vector database is better than traditional relational databases because the task at hand is keyword based search instead of semantic search, vector similarity based searches are quicker and better suited for large CVs

- Choice of embeddings - OpenAI's ada embeddings are well suited and compatible with Pinecone, plus cheap to use
- Choice of Model - This is unfortunately dictated by my personal cost considerations. For high quality of output and low cost, Open AI GPT-3.5 Turbo is best considering rate limiting for free accounts

- Utilized a single query thoroughly practically tested through normal ChatGPT access instead of multiple queries due to rate limiting. Additionally, ChatGPT has effective access to information to determine merit within each criteria except pinpointing high remuneration. Considering this is a rough estimate and not an application meant to provide an exact outcome, the parameters assigned via prompt engineering for determining high remuneration are great.

Evaluation Criteria for Pipeline

This is divided into two parts: detection accuracy for all resume items meeting the criteria conditions for each criteria, and overall accuracy in prediction O1-A application acceptance/rejection chances

1. <u>Detection Accuracy</u>
   - This is pretty straightforward to evaluate. A human reader can categorize resume items into the O1-A Criteria, and this can the examined against the Model Output to create a Confusion Matrix
   - $Detection\ Innaccuracy\ Score\ = (\Sigma\,[(0.8\ *\ FP)\ +\ (0.2\ *\ FN)]\ for\ all\ i\ \in [Criterias]$

2. <u>Overall Accuracy</u>
   - This is a simple label match percentage for Predicted Outcomes as referenced against actual client data on O1-A Visa Acceptance/Rejection