# Technical Design Document for O-1A Insight

## Introduction

O-1A Insight is a sophisticated application designed to evaluate qualifications for the O-1A visa using **Natural Language Processing (NLP)** and **Machine Learning (ML)** technologies. The system automatically evaluates CVs against the O-1A visa's rigorous criteria.

## System Architecture

### 1. Application Framework

- **FastAPI**: Selected for its asynchronous capabilities and built-in data validation. FastAPI provides advanced routing, dependency injection, and interactive API docs with Swagger UI, crucial for real-time debugging and testing of endpoints.

### 2. Text Processing and Extraction

- **PyPDF2**: Utilized for its simplicity in reading from and controlling PDFs, which is essential for text extraction processes.
- **spaCy**: Chosen for its efficient and robust NLP capabilities, spaCy is used to split text into sentences, tokenize, and lemmatize text, preparing it for further analysis.

### 3. Entailment and Text Classification

- **Transformers** and **PyTorch**: Applied for using pretrained models from Hugging Face's library, specifically `distilbert-base-uncased-mnli` for entailment tasks related to text, capturing subtle aspects of language essential for assessing visa eligibility.

### 4. Model Deployment

- **Uvicorn**: A fast, lightweight ASGI server supporting asynchronous tasks, useful for handling I/O operations like file uploads and NLP processing.

## Data Flow

- **File Upload**: Users upload a CV via the FastAPI endpoint, supporting PDF and text files.

- **Text Extraction**: The `cv_processor.py` module reads text content from the CV, handling various formats and encodings as necessary.
- **Criteria Mapping and Evaluation**: Text is forwarded to `nlp_extractor.py`, which evaluates the text using spaCy and analyzes each sentence with the Transformers model based on preconfigured criteria templates.
- **Evidence Aggregation and Scoring**: Sentences meeting criteria above a threshold are logged as evidence. Scores are calculated based on match strength and keyword occurrence specific to each criterion.
- **Evaluation**: The `evaluator.py` module calculates score aggregates, using a heuristic to determine an overall qualification rating from the number of criteria passed and the strength of the evidence.

# Design Choices

## Why FastAPI instead of Flask or Django?

- **FastAPI**'s architecture is suited for asynchronous request handling, critical for I/O-bound activities in our application, such as file uploads and external API calls to ML models.

## Why spaCy instead of NLTK or other NLP libraries?

- **spaCy** provides industrial-strength performance in processing speed and supports vector-based text processing, vital for semantic analysis in this application.

## Why Transformers?

- The use of a cutting-edge model like `distilbert-base-uncased-mnli` is justified by its performance on natural language inference tasks, central to O1A Insight's criteria assessment mechanism.

# Conclusion

O1A Insight leverages advanced technologies in NLP and web frameworks to deliver a fast, efficient, and accurate solution for determining O-1A visa eligibility. Its performance-driven, scalable, and precise architecture ensures quick and reliable visa status determinations for users.