**Design Choices and Evaluation**

**System Design**

The system is designed as a FastAPI application with a single endpoint that accepts a CV or resume text as input. The core of the system is a large language model (LLM) trained on the eight O-1A visa criteria and various examples of CVs and resumes.

The decision to use an LLM was made for several reasons:

1. **Natural Language Understanding**: LLMs excel at understanding and processing natural language input, making them well-suited for analyzing the unstructured text found in CVs and resumes.
2. **Knowledge Representation**: LLMs can effectively represent and reason about the complex and nuanced criteria required for the O-1A visa assessment.
3. **Adaptability**: LLMs can be fine-tuned on domain-specific data, allowing for better performance on the O-1A visa assessment task.
4. **Scalability**: LLMs can handle a wide range of input lengths and formats, making them suitable for processing diverse CVs and resumes.

The system architecture consists of the following components:

1. **Data Preprocessing**: The input CV or resume text is cleaned and formatted for compatibility with the LLM.
2. **LLM Inference**: The preprocessed text is fed into the LLM, which generates scores for each of the eight O-1A criteria and an overall assessment.
3. **API Endpoint**: The FastAPI endpoint handles incoming requests, passes the input to the LLM, and returns the generated scores and assessment.

**Evaluating the Output**

The output of the O-1A Visa Qualification Assessment AI can be evaluated on two levels:

1. **Criteria Scores**: The individual scores for each of the eight O-1A criteria should be analyzed to understand the strengths and weaknesses of the applicant's qualifications. Higher scores indicate stronger evidence for meeting that particular criterion.
2. **Overall Assessment**: The overall assessment provides a high-level summary of the applicant's qualifications based on the combined scores for all criteria. This assessment should be used to gauge the overall strength of the application and identify areas that may require additional documentation or evidence.

See the README.md for information about how to run the service & evaluate the output.