Technical Report: EB-1A RFE Risk Analyzer System

---

1. Problem Framing and Approach

Problem Statement

EB-1A immigration petitions require demonstrating extraordinary ability through at least three of ten USCIS criteria. High RFE rates result from insufficient evidence, generic recommendation letters, and weak substantiation of claims. Legal practitioners need systematic tools to identify petition weaknesses before submission.

Approach

The Risk Analyzer leverages **Google Gemini AI** to automatically analyse petition documents, segment content into logical sections, and flag areas likely to trigger RFEs. The system provides structured risk assessments mapped to EB-1A criteria with actionable improvement recommendations.

2. Architecture and Tool Design

System Architecture

Input Documents → Document Processing → AI Analysis → Report Generation → Output

[DOCX/PDF/TXT] → [Text Extraction] → [Gemini AI] → [Risk Aggregation] → [DOCX Report]

Core Components

- **Document Processing Layer**: Handles multiple file formats using python-docx and PyMuPDF with rule-based segmentation for petition sections (personal statements, evidence, recommendation letters)
- **AI Analysis Engine**: Interfaces with Google Gemini AI using context-aware prompting that simulates USCIS adjudicator perspective with comprehensive EB-1A criteria knowledge
- **Report Generation System**: Produces professional DOCX reports with executive summaries, risk matrices, detailed findings, and visualization charts
- **Orchestration Layer**: Coordinates workflow execution with error handling and resource management

Key Components

- **`main.py`:** Orchestrates the workflow.
- **`utils.py`:** Handles file reading and text extraction.
- **`segmentation.py`:** Implements sectioning logic.
- **`analysis.py`:** Interfaces with Google Gemini AI for risk analysis.
- **`report.py`:** Formats and writes the final report.

Data Flow

1. **Input Validation**: Document format verification
2. **Text Extraction**: Content extraction with format preservation
3. **Intelligent Segmentation**: Section identification using pattern matching
4. **AI Analysis**: Concurrent processing through analysis pipeline
5. **Report Synthesis**: Automated comprehensive risk assessment generation

3. Edge Cases and Limitations

- **File Format Support:** Only DOCX, PDF, and TXT files are supported. Scanned PDFs (images) are not handled.
- **Section Detection:** Segmentation relies on clear section headers. Unconventional formatting may lead to missed or merged sections.
- **AI Model Limitations:** The Gemini model may not always interpret legal nuances or rare case-specific evidence correctly.
- **API Key and Model Access:** Requires a valid Google API key and access to the correct Gemini model. Model availability may vary by region or account.
- **Error Handling:** Minimal error handling for corrupt or missing files; improvements are needed for production use.

4. Future Extension Opportunities

Near-term Enhancements (3-6 months)

- **OCR Integration**: Tesseract implementation for scanned document support
- **Web Interface**: Browser-based application with real-time processing
- **Enhanced Segmentation**: Machine learning-based section identification

Medium-term Developments (6-12 months)

- **Predictive Modelling**: Historical RFE pattern analysis for improved accuracy
- **Integration Capabilities**: Legal practice management system connectivity
- **Batch Processing**: High-volume petition analysis capabilities

Long-term Vision (1-2 years)

- **Multi-category Support**: Extension to EB-2 NIW, EB-5, and other immigration categories
- **Custom Model Training**: Client-specific fine-tuning based on historical outcomes
- **Enterprise Solutions**: API services for legal technology platform integration