

LLM Document Analyzer

Time Limit: 3 Hours

Problem Statement

Build an intelligent document analyzer that identifies document types and detects missing critical information in business documents.

What You'll Build

A web application that processes PDF documents, identifies their type (contract/invoice), and uses LLM to find missing required fields.

Resources Provided

- **Document Templates:**
https://drive.google.com/drive/folders/1RtkEUgxGIjgX__bVrfaRIBr52X54iNy3?usp=sharing
 - Complete contract with all required fields (for positive testing)
 - Incomplete invoice with missing critical fields (for negative testing)
- **Required Fields Reference:**
 - Contract: party_1, party_2, signature, date, payment_terms
 - Invoice: invoice_number, amount, due_date, tax, bill_to, bill_from
- **LLM Access:** Use any LLM API (OpenAI, Anthropic, Google, Hugging Face, or local models)

Core Features (Implement 3 out of 4)

1. Document Upload & Processing

- Upload PDF documents through web interface
- Extract and display text content from PDFs
- Store document content in database

2. LLM Document Classification

- Use LLM to automatically identify document type
- Support contracts, invoices, and reports
- Display detected document type with confidence

3. Missing Fields Analysis

- LLM-powered detection of missing required fields
- Compare document content against field requirements
- Generate detailed missing fields report

4. Improvement Recommendations

- Provide specific suggestions for completing documents
- Highlight critical vs optional missing information
- Generate actionable improvement checklist

Technical Requirements

- **Frontend:** Document upload interface, results display
- **Backend:** PDF processing, LLM integration, analysis APIs

- **Database:** SQLite with tables for documents and analysis results
- **LLM Integration:** Any LLM API for document classification and field detection

Success Criteria

- Successfully upload and extract text from PDF documents
- LLM correctly identifies document types (contract vs invoice)
- Accurately detects missing required fields
- Provides clear, actionable recommendations

Time Management Tips

- **Hour 1:** Project setup, PDF upload, text extraction
- **Hour 2:** LLM document type detection, database integration
- **Hour 3:** Missing field analysis, results display, testing

Prioritize accuracy over speed - reliable field detection is more valuable than processing many documents!