Legal-BERT QA System Documentation

Overview

The Legal-BERT QA System is an advanced AI-powered application that enables users to extract information from legal documents through natural language questions. The system leverages a Legal-BERT model fine-tuned on the CUAD (Contract Understanding Atticus Dataset) to provide accurate answers to complex legal queries.

Key Features

Multi-format Document Support: TXT, DOCX, PDF, and images (PNG, JPG, JPEG) with OCR capabilities

Legal Domain Specialization: Fine-tuned on CUAD dataset for legal terminology understanding

Advanced NLP Processing: Context-aware answer extraction with confidence scoring

User-Friendly Interface: Streamlit-based web interface with document preview and progress indicators

Model Architecture

Legal-BERT Base Model

The system uses BERT (Bidirectional Encoder Representations from Transformers) as its foundation. BERT is a transformer-based model that excels at understanding context in text through its bidirectional training approach.

Model Type	**bert-base-uncased**
Parameters	110 million
Layers	12
Attention Heads	12
Hidden Size	768
Max Sequence Length	512 tokens

Fine-Tuning on CUAD Dataset

- The model has been fine-tuned on the Contract Understanding Atticus Dataset (CUAD), which contains:
  - 510 legal contracts
  - 13,000+ expert annotations
  - 41 categories of legal questions
  - Focus on important clauses and provisions
- Fine-Tuning Process:

- Pre-trained BERT model initialized
- Trained on CUAD's question-answer pairs
- Optimized for extractive question answering
- Specialized for legal domain terminology

Technical Components

Text Extraction

File Type	Library Used
TXT	Python built-in IO
DOCX	docx2txt
PDF	PyPDF2
Images	pytesseract (Tesseract OCR)