AI Legal Assistant — Document Analysis & Clause Summarization

Overview

The AI Legal Assistant is a Streamlit-based intelligent system designed to analyze legal PDF documents and explore clause datasets. It helps users — including legal professionals, students, and auditors — automatically:

- Extract key legal clauses
- Analyze risks and compliance
- Summarize lengthy documents
- Visualize clause types
- Download structured PDF summaries

Technologies & Libraries Used

Category	Library/Tool	Purpose
Web App	Streamlit	Interactive frontend for user interface
PDF Reading	PyMuPDF (fitz)	Parsing and reading PDF documents
Text Analysis	re	Regular expression for extracting data
Summarization	transformers (Hugging Face)	Using pre-trained NLP models
Model Used	t5-base	Pre-trained transformer model for summarization
Charting	plotly.express, plotly.graph_objects	Interactive pie charts, gauge meters

PDF Output	fpdf	To generate downloadable summary PDFs
Data	pandas	CSV reading and tabular data manipulation

Dataset Used

We use the CUAD Dataset (Contract Understanding Atticus Dataset):

- File: cuad_descriptions.csv
- Purpose: Helps users explore different clause types (e.g., "Confidentiality", "Termination", "Arbitration")
- Structure: CSV format with legal clause descriptions
- Application: Used in the "Explore Clause Dataset" feature to allow interactive summarization

Model Used for Summarization

- Model: t5-base
- Source: Hugging Face Transformers
- Type: Pre-trained Text-to-Text Transformer
- Purpose: Converts large legal clause text into short summaries
- Why T5? General-purpose, efficient on short prompts, and quick inference in lightweight application.

Features Implemented

PDF Upload & Processing

Clause Detection & Classification

Risk & Compliance Analyzer

Clause Summarization

Structured Info Extraction

Clause Category Visualization

CUAD Dataset Clause Explore

Downloadable PDF Summary

How It Works (Backend Logic)

1. PDF Upload → fitz extracts full text
2. Clause Count → Regex + keywords to classify clauses
3. Clause Checker → Matches expected legal keywords
4. Summarization → pipeline("summarization", model="t5-base")
5. Risk Score → Rule-based score + findings + visual chart
6. Export → fpdf builds downloadable report

Project Strengths

- Fast, lightweight, and no database dependency
- Uses pre-trained NLP model — no retraining needed
- Suitable for legal document screening & summarization
- Downloadable reports make it easy to share with teams or lawyers