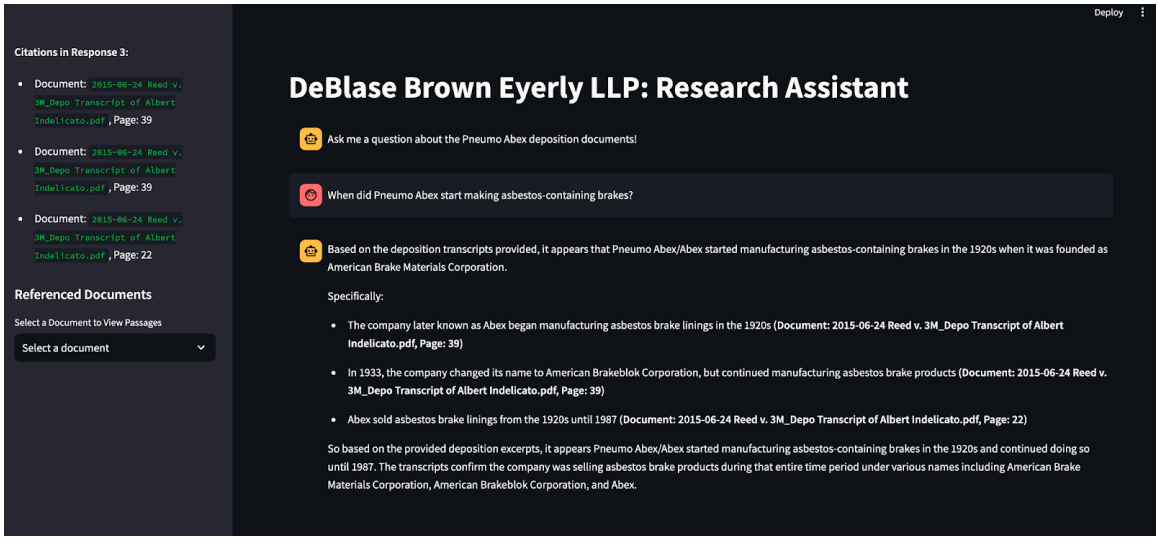
# Data Science Report: Transformer-Based Model for Legal Document QA




## Project Overview:

This project developed a sophisticated QA system using transformer models, designed to handle the complex legal documents associated with asbestos litigation. Legal documents, which are dense and filled with industry jargon, were processed through a combination of transformers and Retrieval-Augmented Generation (RAG). The system extracts precise, evidence-based answers, citing specific document locations to support legal research.

## Key Technical Components:

### 1. Data Preprocessing:

• Document Ingestion: PDFs containing legal depositions and exhibits are parsed using PyPDFDirectoryLoader, splitting them into manageable chunks (~800 characters) with an 80-character overlap. This overlap ensures the preservation of important context across chunks.

• Metadata Annotation: Each chunk retains its source file and page number, ensuring traceability when answering legal questions. This allows the model to provide evidence-backed answers.

### 2. Embedding & Retrieval:

• Embedding Generation: Using HuggingFace MPNet, each chunk of text is converted into embeddings and stored in a Chroma vector store. These embeddings enable efficient retrieval of relevant document sections.

• Maximal Marginal Relevance (MMR): The retriever balances between returning highly relevant and diverse document chunks. This ensures that the retrieved data is both precise and representative of broader contexts within the documents.

• Contextual Compression: To enhance retrieval speed, the system compresses retrieved documents using a transformer model, while preserving essential content for QA.

### 3. Question Answering System:

• Model: Claude-2, a transformer-based language model, is fine-tuned for legal text processing. The model handles complex legal language and generates evidence-based answers by analyzing retrieved chunks.

• Conversational Flow: The system supports follow-up questions and maintains the context across conversations using a Conversational Buffer Memory. This is critical for legal cases where researchers build questions on top of previous answers.

• Citation Management: Each answer is linked to the original documents (e.g., "Document: filename.pdf, Page: X"), ensuring that the legal professionals can trace every piece of information to its source.

### 4. User Interface & Interaction:

• Streamlit Interface: The front-end is built with Streamlit, providing an intuitive platform where users input questions and receive answers. The interface highlights citations and allows for PDF downloads, streamlining verification.

• Custom Prompt Templates: The model is guided by a custom-built prompt template to ensure responses are concise, accurate, and grounded in the context provided by the legal documents.

## Dataset & Use Case:

The system was trained and tested on legal deposition transcripts from asbestos litigation cases, such as Davidson v. Burns (2004) and Reed v. 3M (2015). These cases involve complex legal discussions about asbestos exposure, company liability, and health hazards.

• Document Structure: Each document includes metadata (case name, court, date, participants) and detailed transcripts structured in a Q&A format. Exhibits like internal memos and legal correspondence are also included for reference during depositions.

• Example Question: "When did [company] first become aware of the hazards of asbestos?" The model retrieves relevant excerpts and cites specific page numbers for the user to verify.

## Preprocessing & Model Tuning:

• Text Splitting: Documents are split into chunks for retrieval, ensuring critical legal context is not lost.

• Fine-tuning: The model is specifically optimized to handle legal jargon and nuanced language, crucial for producing accurate answers in legal contexts.

## Why Transformer & RAG Are Ideal:

• Contextual Understanding: Transformers can maintain context over long documents, ensuring accuracy in both retrieval and answer generation.

• Evidence-Based Results: RAG ensures that answers are not only generated but are based on document excerpts that are retrieved and cited, a critical requirement in legal settings.