# LEGAL DOCUMENT QUERY SYSTEM

**Your law Questions Answered**


Presented by

- Shivangi Singh
- Sahil Shenoy
- Sachit Lele

# CONTENT



# Introduction

The Question Answering System for Legal Documents of India is designed to simplify the process of understanding and accessing legal documents in India. This project focuses on developing a state-of-the-art Natural Language Processing (NLP) system that can answer questions related to legal documents in PDF format. The objective is to make legal information more accessible to users by allowing them to ask questions about legal PDFs, ultimately enhancing legal literacy and enabling easier navigation through the complexities of Indian law

# Problem Statement.

Our goal is to develop a model that can accurately answer questions based on the content of legal documents especially indian legal docs, making it easier for users to find and utilize the information they need.

Models such as roBERTa,LLAMA and naive approaches are applied to get the result

Preprocessing of PDF'S

01

Applying Naive approach
Cosine similarity

02

roBERTa

03

Llama

# Preprocessing

- Remove Header and Footer Notes of all pages of Documents
- Combine the pages into one single document
- Extract the raw text from the document
- Remove Punctuation marks and symbols  like(*,!#% |  - ^*- etc.)
- Extract context from the data by splitting the raw text on subsections demarcated by roman numerical like ( i, ii , iv)

# Cosine Similarity

Preprocessing legal documents and creating a dataset with passages, questions, answers and starting index.

Create a word-index dictionary from the document.

Split the document into sentences and store them in a list.

Convert each sentence into a binary vector representation based on the presence of words.

Preprocess the user's question by removing stop words.

Convert the processed question into a binary vector.

Calculate cosine similarity between the question vector and sentence vectors.

Identify the sentence with the highest similarity score as the answer.and Extract and present the answer to the user.

# roBERTa.

Preprocessing legal documents and creating a dataset with passages, questions, answers and starting index.

Importing the necessary libraries and seting up logging for the model and defining the training and evaluation data,

Creating model configuration and specifing parameters for training.  We are using the RoBERTa model with "roberta-base" and configuring batch size and evaluation settings.

Initializing and training the RoBERTa model and also validating on validation dataset

After training, evaluating the model's performance on the evaluation data.

Using the trained model to make predictions on new legal documents and question

Depending on the evaluation results,  fine-tuning the model further, adjusting hyperparameters,

# Llama.

Preprocessing legal documents and creating a dataset with passages, questions, answers and starting index.

Convert data into word embeddings and Store embeddings in a database.

Accept user queries and then Calculate similarity scores between user queries and stored embeddings.

Retrieve embeddings with the highest similarity scores.

Generate refined answers based on retrieved embeddings.

Present answers to users in a user-friendly interface.

Ensure scalability and optimization and Continuously monitor and update the system for performance improvements.

# Results and Discussion.

In our quest to create a question-answering model for legal documents in PDF format, we employed various methods, including Cosine Similarity, Word2Vec, RoBERTa, and LAMA. The findings can be succinctly summarized as follows:

Cosine Similarity: This basic method produced poor results due to its inability to grasp the intricate language and context inherent in legal documents.

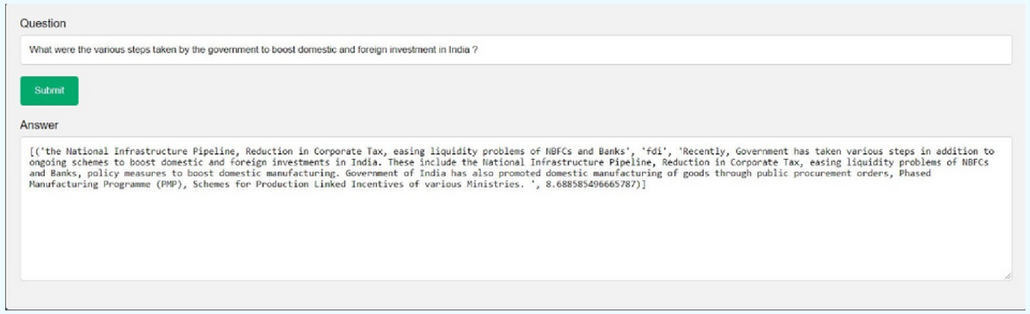
RoBERTa: RoBERTa, a transformer-based model, outperformed Cosine Similarity, demonstrating its suitability for understanding legal text, albeit with room for improvement.

Evaluation loss: 0.471

LAMA: LAMA, a specialized language model , emerged as the top performer, offering the most accurate and contextually relevant answers. Its domain-specific focus proved invaluable for legal question answering.

The choice of model significantly impacts the accuracy and effectiveness of our question-answering system. LAMA's domain-specific expertise makes it the preferred choice for handling legal documents, ensuring precise and valuable results.

# Deployment.




All three of our approaches, including Cosine Similarity, RoBERTa, and LAMA are deployed on the Streamlit platform. This deployment strategy provides a unified and user-friendly interface for our legal document question-answering system, allowing legal professionals and researchers to choose the most suitable approach for their specific tasks. Streamlit's versatility empowers users to seamlessly interact with the models, ensuring efficient information and answer extraction from legal documents in a convenient and accessible manner.

# Limitations.

- **Data Quality and Quantity**: The performance of machine learning models, including LLaMa, relies on the quality and quantity of training data. Inaccurate representation in legal documents can hinder the model's effectiveness.
- **Fine Tuning**: Fine-tuning quality is vital. Poor optimization, limited or biased data, and overfitting can all impact a model's performance and generalizability.
- **Hardware and Resources:** Transformer models like RoBERTa and LLaMa are resource-intensive, making their training and fine-tuning costly and time-consuming.
- **Interpretable Output :** While LLaMa may be accurate, the interpretability of its results is crucial for legal professionals to trust the model's decisions.

# Conclusion.

By employing various approaches, including Cosine Similarity, RoBERTa, and LAMA , in the context of question-answering for legal documents, we are addressing a substantial challenge in the legal domain. Legal documents often contain intricate and highly specialized language and require a deep understanding of context and semantics. These models, especially domain-specific ones like LAMA, offer the potential to significantly improve the accuracy and efficiency of extracting pertinent information and answers from legal texts, facilitating better legal research, analysis, and decision-making.