Retrieval-Augmented Generation for Indian Legal Corpus

Intelligent Legal Assistant for Indian Legislations using Retrieval-Augmented Generation (RAG)

Objective

The project aims to build a semantic legal search and Q&A system over Indian legislative documents (e.g., Constitution, GST Acts, Criminal Laws). It enables users to ask natural language questions and receive accurate, concise, and reference-grounded answers using Retrieval-Augmented Generation (RAG) powered by OpenAI and ChromaDB.

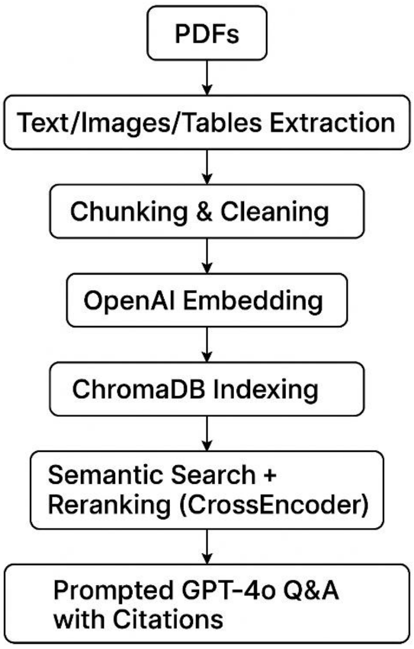
Problem Statement

The Indian legal system comprises thousands of statutes, amendments, and constitutional provisions spanning multiple domains such as criminal law, taxation, civil procedure, and more. Legal researchers, policymakers, and citizens often face significant challenges in accessing relevant provisions, understanding their context, and identifying applicable laws efficiently.

Key Challenges:

- **Dispersed Information:** Legal documents are spread across lengthy and often complex PDFs without structured indexing.
- **Low Searchability:** Keyword-based PDF search is inadequate for answering contextual queries.
- **Lack of Intelligent Tools:** No publicly available AI-powered legal assistant exists that can perform retrieval-augmented reasoning grounded in official legal texts.
- **Repetition and Noise in Raw PDFs:** Legal PDFs often contain repeated content in both text and tables, along with embedded images and formatting issues.

SYSTEM DESIGN


PDFs
Text/Images/Tables Extraction
Chunking & Cleaning
OpenAI Embedding
ChromaDB Indexing
Semantic Search + Reranking (CrossEncoder)
Prompted GPT-4o Q&A with Citations

How It Works

1. Data Processing Pipeline

Load PDF Documents

- `pdfplumber` is used to extract:
    - Text
    - Tables
    - Images

Preprocessing & Chunking

- Text chunks grouped using font size and line gaps
- Image and table bounding boxes excluded from OCR text
- Metadata is captured per paragraph/table/image (e.g., document name, page number)

2. Embedding & Vector Storage

Vector Embedding

- Model: `text-embedding-ada-002` (OpenAI)
- Embedding Input: Cleaned paragraphs
- Output: Vector + Metadata

Vector Storage: ChromaDB

- Stored as persistent local DB
- Collections:
    - `RAG_on_Indian_Legislations` for primary lookup
    - `Legislations_Cache` for fast repeated queries

3. Semantic Search & Re-ranking

Semantic Retrieval

- Top 10 documents retrieved by embedding cosine similarity

CrossEncoder Reranking

- Model: `cross-encoder/ms-marco-MiniLM-L6-v2`
- Improves relevance by scoring (query, chunk) pairs
- Only top 3 used in final generation

4. Response Generation

GPT-4o Prompting

- Role: Legal assistant
- Prompt: Inject top 3 retrieved references + user query
- Answer must be:
    - Clear and concise
    - Cited with legal references (sections/articles)

Technology Stack

Layer	Tools / Libraries
NLP Embedding	OpenAI `text-embedding-ada-002`
Vector Store	ChromaDB (chromadb)
OCR & Parsing	pdfplumber
Semantic Search	OpenAI Embeddings + ChromaDB
Reranking	Sentence Transformers CrossEncoder
Q&A Generation	OpenAI gpt-4o
Environment	Python 3.11, Jupyter Notebook

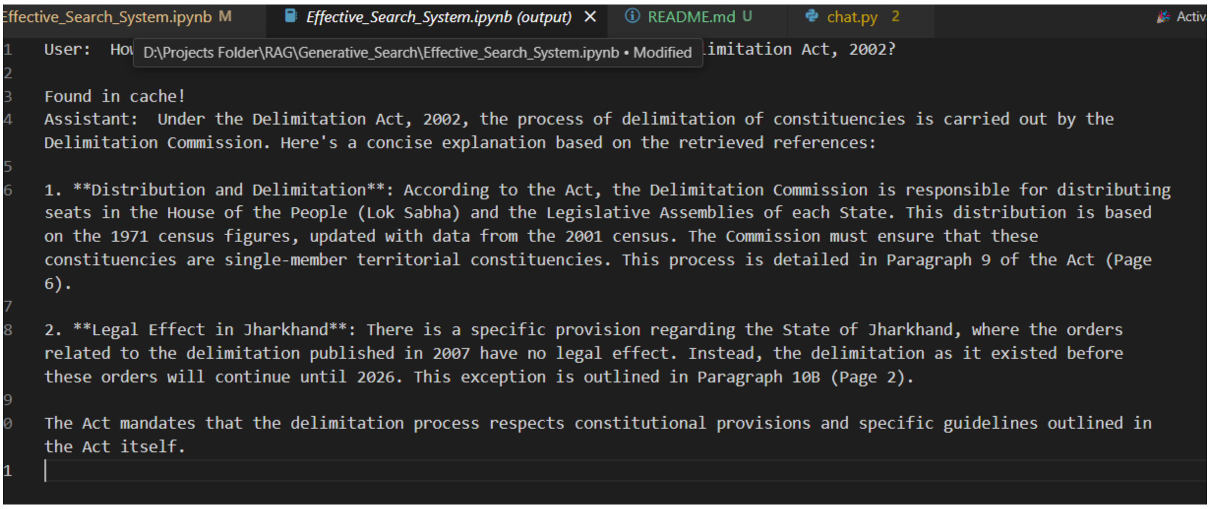
Achievements

- Parsed multiple PDFs (text, tables, images)
- Avoided duplicate content from tables/images
- Built ChromaDB vector store using OpenAI embeddings
- Implemented fast cache-based semantic search
- Reranked search results using sentence-transformers
- Generated GPT-4o responses with citations

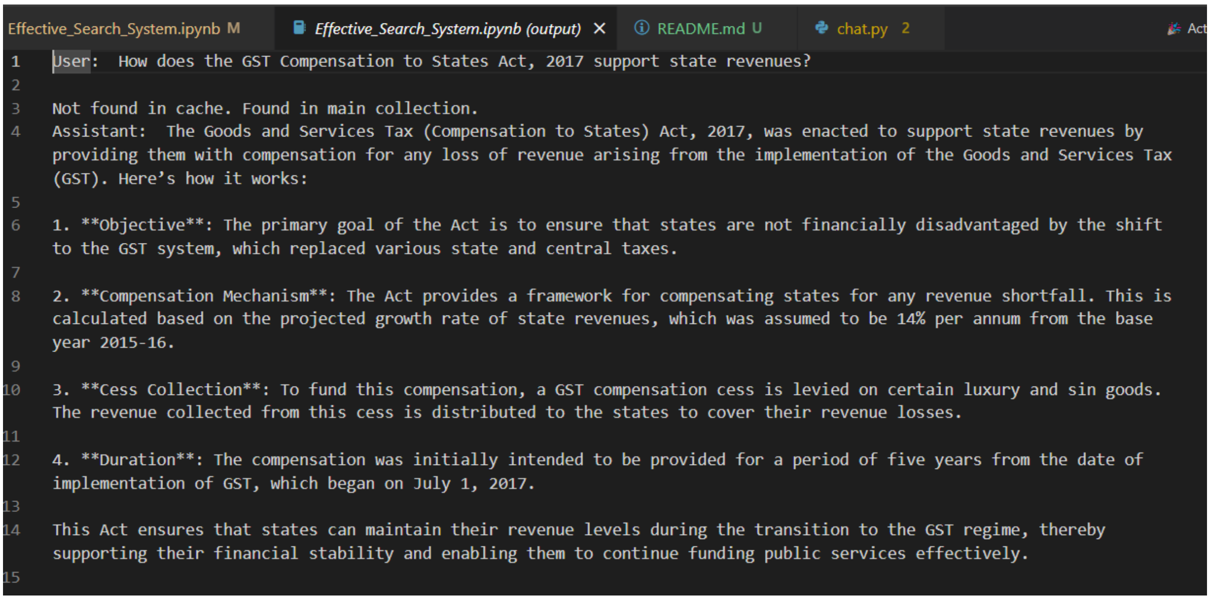
Sample User Queries

- What are the key features of the Bharatiya Nagarik Suraksha Sanhita, 2023?
- What changes does the Bharatiya Nyaya Sanhita, 2023 introduce in criminal law?
- How is electronic evidence treated under the Bharatiya Sakshya Adhiniyam, 2023?
- What is the procedure for filing a civil suit under the Code of Civil Procedure, 1908?
- What are the fundamental rights guaranteed by the Constitution of India?
- How is income from house property taxed under the Income Tax Act, 1961?
- What is the objective of the Direct Tax Vivad se Vishwas Act, 2020?
- What are the key provisions of the Central Goods and Services Tax Act, 2017?
- How does the GST Compensation to States Act, 2017 support state revenues?
- What are the rules for inter-state supply under the Integrated GST Act, 2017?

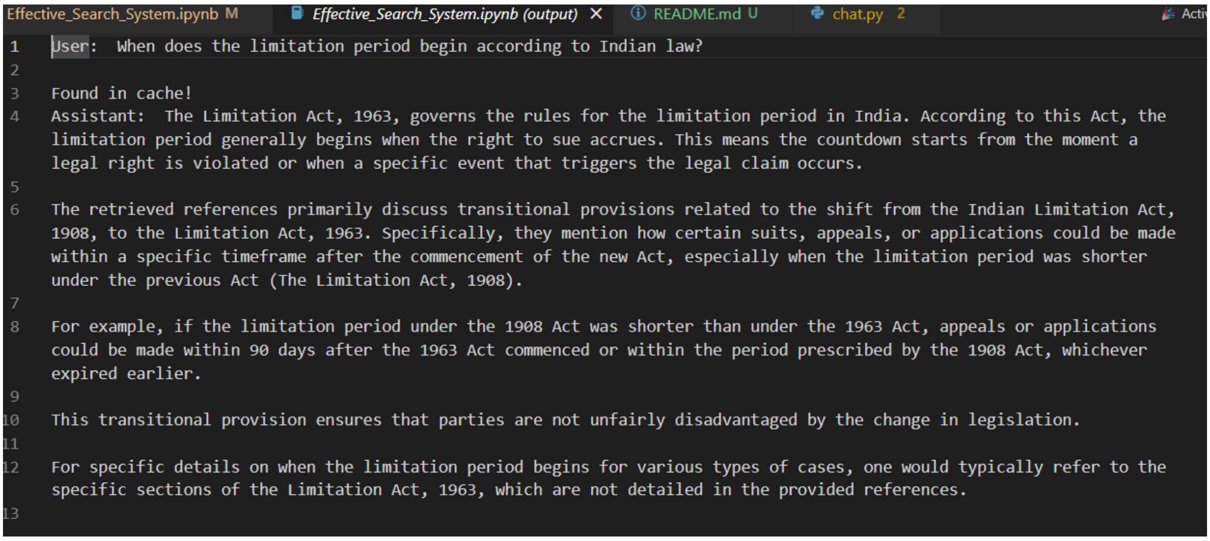
SAMPLE CONVERSATIONS (RAG)


ffective_Search_System.ipynb M
Effective_Search_System.ipynb (output)
README.md U
chat.py 2
Activ
D:\Projects Folder\RAG\Generative_Search\Effective_Search_System.ipynb • Modified

```
User:  Ho imitation Act, 2002?

Found in cache!
Assistant:  Under the Delimitation Act, 2002, the process of delimitation of constituencies is carried out by the
Delimitation Commission. Here's a concise explanation based on the retrieved references:

1. **Distribution and Delimitation**: According to the Act, the Delimitation Commission is responsible for distributing
seats in the House of the People (Lok Sabha) and the Legislative Assemblies of each State. This distribution is based
on the 1971 census figures, updated with data from the 2001 census. The Commission must ensure that these
constituencies are single-member territorial constituencies. This process is detailed in Paragraph 9 of the Act (Page
6).

2. **Legal Effect in Jharkhand**: There is a specific provision regarding the State of Jharkhand, where the orders
related to the delimitation published in 2007 have no legal effect. Instead, the delimitation as it existed before
these orders will continue until 2026. This exception is outlined in Paragraph 10B (Page 2).

The Act mandates that the delimitation process respects constitutional provisions and specific guidelines outlined in
the Act itself.
```


Effective_Search_System.ipynb M
Effective_Search_System.ipynb (output)
README.md U
chat.py 2
Act

```
User:  How does the GST Compensation to States Act, 2017 support state revenues?

Not found in cache. Found in main collection.
Assistant:  The Goods and Services Tax (Compensation to States) Act, 2017, was enacted to support state revenues by
providing them with compensation for any loss of revenue arising from the implementation of the Goods and Services Tax
(GST). Here’s how it works:

1. **Objective**: The primary goal of the Act is to ensure that states are not financially disadvantaged by the shift
to the GST system, which replaced various state and central taxes.

2. **Compensation Mechanism**: The Act provides a framework for compensating states for any revenue shortfall. This is
calculated based on the projected growth rate of state revenues, which was assumed to be 14% per annum from the base
year 2015-16.

3. **Cess Collection**: To fund this compensation, a GST compensation cess is levied on certain luxury and sin goods.
The revenue collected from this cess is distributed to the states to cover their revenue losses.

4. **Duration**: The compensation was initially intended to be provided for a period of five years from the date of
implementation of GST, which began on July 1, 2017.

This Act ensures that states can maintain their revenue levels during the transition to the GST regime, thereby
supporting their financial stability and enabling them to continue funding public services effectively.
```


Effective_Search_System.ipynb M
Effective_Search_System.ipynb (output)
README.md U
chat.py 2
Acti

```
User:  When does the limitation period begin according to Indian law?

Found in cache!
Assistant:  The Limitation Act, 1963, governs the rules for the limitation period in India. According to this Act, the
limitation period generally begins when the right to sue accrues. This means the countdown starts from the moment a
legal right is violated or when a specific event that triggers the legal claim occurs.

The retrieved references primarily discuss transitional provisions related to the shift from the Indian Limitation Act,
1908, to the Limitation Act, 1963. Specifically, they mention how certain suits, appeals, or applications could be made
within a specific timeframe after the commencement of the new Act, especially when the limitation period was shorter
under the previous Act (The Limitation Act, 1908).

For example, if the limitation period under the 1908 Act was shorter than under the 1963 Act, appeals or applications
could be made within 90 days after the 1963 Act commenced or within the period prescribed by the 1908 Act, whichever
expired earlier.

This transitional provision ensures that parties are not unfairly disadvantaged by the change in legislation.

For specific details on when the limitation period begins for various types of cases, one would typically refer to the
specific sections of the Limitation Act, 1963, which are not detailed in the provided references.
```