Project Title: LegalDoc-RAG: Search Less, Win More

Project Context

Project Summary:

Lawyers waste hours searching through legal documents to find the right law or case that supports their argument. This slows down case preparation and increases costs. We aim to build a simple, fast, and trustworthy system that finds relevant laws and cases instantly with proof.

Problem statement:

The process of finding the right legal text with accurate citations is still slow and prone to errors. Manual searches across multiple documents consume significant time, often requiring lawyers or researchers to sift through large volumes of statutes and case law. Inaccuracies, such as missing or incorrect precedents, can have serious consequences, potentially weakening legal arguments or leading to unfavorable case outcomes. This inefficiency not only increases workload and costs but also creates a risk of overlooking critical legal information.

Project goals:

1. *Enable fast, accurate legal search that returns only the most relevant law or case excerpts.*
2. *Provide every answer with verified citations from authoritative legal sources.*
3. *Offer a simple, lawyer-friendly interface for quick adoption and use.*

Application areas:

1. Legal research in law firms
2. Contract and compliance checks
3. Government policy and regulation review

Project justifications

1. *Legal research inefficiency costs billions annually* (Josten, 2025)

2. *Reduces time-to-answer from hours to seconds, without compromising accuracy* (*Efficiency Without Compromise Using AI in Legal Research*, 2025)
3. *Improves legal service quality, reduces costs, and mitigates risk of errors.* (*Using AI for Legal Research: How to Boost Accuracy & Efficiency*, 2025)

Limitations & Constraints:

1. The MVP will support only one jurisdiction (e.g., US) due to dataset and time constraints.
2. The bundled dataset will not auto-update; new laws or cases will require a manual refresh.
3. The dataset will be limited to available public legal documents and may not include all historical records.
4. Optimized for small to medium datasets; search speed and accuracy may drop if scaled to millions of documents without further optimization.

Assumptions:

1. Publicly available legal datasets for the chosen jurisdiction are sufficient for MVP functionality.
2. All bundled legal documents are in machine-readable format.

Team Member Details:

Name	Email ID	Github ID
RISHIKESH P…	**rishikesh0523@gmail.com**	**rishikesh0523**
SANDESH K…	**sandeshkuikel07@gmail.com**	**sandeshkuikel07**
SHREYA UPR…	**shreyyauprety@gmail.com**	**shreyaupretyy**
PRAFULLA R…	**prafullapokhrel421@gmail.com**	**Prafulla45**

Requirements

Expected Inputs:

1. A plain-text legal question or keyword phrase (e.g., *"termination of contract for*

non-delivery", *"cases citing Section 23 of the Contract Act"*).

2. Jurisdiction Selection – User chooses the legal system/dataset to search within (limited to one for MVP).

Expected outputs:

1. Top Relevant Excerpts: Exact paragraph or clause from a statute or case law that matches the query.
2. Verified Citation Links: Clickable references to the original authoritative document (law, judgment, or official source).
3. Brief Context Summary: Short explanation of how the retrieved excerpt answers the query, generated only from the retrieved text.
4. Confidence/Relevance Score: Numerical indicator showing how relevant the retrieved excerpt is to the query.

End-User & Stakeholders Requirements

What task does end-user do with the product

Requirements	Details	Priority	Success Criteria
As a lawyer, I want to type a legal question and instantly get the most relevant law or case excerpt so that I can prepare arguments faster and reduce research time.	-	HIGH	The system consistently delivers top-ranked results that are directly applicable and trustworthy for preparing legal arguments, as validated by legal experts.
As a lawyer, I want to type a legal question and instantly get the most relevant law or case excerpt so that I can prepare arguments faster and reduce research time.	-	MEDIUM	Retrieved excerpts are consistently relevant to the user's query, and their accompanying citations are accurately formatted and verifiable.
As a legal researcher, I want to filter	-	HIGH	The jurisdiction filter

results by jurisdiction so that I only see laws and cases applicable to my work.			reliably and exclusively displays laws and cases from the selected legal system, ensuring the researcher works only with applicable materials.

Functional Requirements

What is the behavior of the system, what should the system do or support

Requirements	Details	Priority	Success Criteria
As a lawyer, I want to type a legal question and instantly get the most relevant law or case excerpt so that I can prepare arguments faster and reduce research time.	Use semantic search (dense retrieval) to understand query intent.	HIGH	The system **consistently ranks the most pertinent legal texts** within the top search results, as validated by legal experts.
As a lawyer, I want to type a legal question and instantly get the most relevant law or case excerpt so that I can prepare arguments faster and reduce research time.	Display the exact law or case excerpt with its reference link.	HIGH	Every citation is **fully verifiable and links directly** to the correct paragraph or clause in the source document without error.

Non-Functional requirements

How well the product must perform in terms of Security, Accuracy, Speed, Aesthetics, Usability, Efficiency, Performance, Maintainability, Error-Checking, …

Requirements	Details	Priority	Success Criteria
As a lawyer, I want to type a legal question and instantly		HIGH	Search results are delivered **without any**

get the most relevant law or case excerpt so that I can prepare arguments faster and reduce research time.			**noticeable delay**, ensuring a fluid and efficient user workflow.
As a user, I want the system to provide highly accurate and relevant results so that I can trust the information for legal work.		HIGH	The top-ranked results are consistently judged by legal experts to be **authoritative, trustworthy, and directly applicable** to the query.
As a user, I want the system to be secure so that my searches and any uploaded data are protected.		MEDIUM	All user data is handled securely with **end-to-end encryption**, and the system successfully passes a standard security audit.

Data collection: Required data /source of the data

Data Identification:

- **Type:** Text-based legal documents.
- **Content:** Statutes, case law, and official legal provisions for a single jurisdiction (MVP scope).
- **Format:** Machine-readable text files (no OCR required).

Data Sources:

- **Primary:** COLIEE dataset (statutes + case law).

- **Alternative/Local:** Nepal Law Commission Acts & Regulations (if focusing on Nepal).
- **Optional Expansion:** CUAD dataset for contract clauses (future scope).

Data Acquisition Methods:

- Direct download from public repositories (COLIEE, open government legal portals).
- Manual curation to remove duplicates and irrelevant files.

Data Collection Challenges:

1. **Formatting inconsistencies**: different documents may have varied layouts; solved by standardizing to plain text before indexing.
2. **Incomplete coverage**: public datasets may not contain all historical cases; MVP will clearly state coverage limitations.

Benchmarking the Model

Establishing a Baseline Model:

- **Baseline Approach:** Use **BM25 keyword search** over the legal corpus to retrieve relevant documents. This provides a simple, fast, and transparent starting point for measuring improvements.
- **Baseline Metrics:**
    - **Relevance Accuracy:** Percentage of top-5 search results judged relevant by a legal expert.
    - **Response Time:** Time taken to return results from the static dataset.

Proposed Model for Improvement:

- **Retriever:** Dense semantic search model (e.g., **bge-m3**) to capture meaning beyond exact keyword matches.
- **Evaluation Criteria:**
    - **Retrieval Accuracy:** The proposed semantic search model must demonstrably outperform the baseline BM25 model by consistently ranking more contextually relevant documents higher for a standard set

of legal queries

- **Citation Integrity:** The system must exhibit unwavering reliability in its citations. Every generated reference must accurately point to the specific source paragraph. Any instance of a hallucinated or incorrect citation is considered a critical failure.
- **Response Time:** The system must process queries and return results with minimal latency, providing an interactive experience that does not disrupt a legal professional's research workflow.

Success Definition: The proposed dense retrieval + reranking system is considered a success if it outperforms BM25 in relevance accuracy while maintaining equal or faster search speeds, and ensures zero hallucinated citations.

Resources required:

Hardware:

- **Development Machine:**
  - Minimum: 16 GB RAM, 4-core CPU, 10 GB free storage.
  - Recommended: GPU-enabled environment (e.g., NVIDIA T4 or RTX 3060+) for faster embedding generation.

- **Hosting:**
  - Local deployment for MVP testing.
  - Optional: Cloud hosting (AWS EC2 / Google Cloud / Azure) for multi-user access during demos.

Software & Tools:

- **Programming Language:** Python 3.7+
- **Libraries/Frameworks:**
  - PyTorch (model handling)
  - SentenceTransformers (embedding generation)
  - FAISS or Weaviate (vector search index)

    - Streamlit or Gradio (web interface)
- **Version Control:** Git + GitHub for collaboration
- **Dataset Storage:** Local JSON or CSV files (bundled in application)

Datasets:

- **Primary:** COLIEE dataset (statutes + case law)

Human Resources:

- **Model Developer:** Implement and fine-tune retriever and search pipeline.
- **Frontend Developer:** Build user-friendly search UI.
- **Legal Domain Expert:** Validate retrieved results for accuracy and relevance.

Other:

- Internet access for dataset download and library installation.
- Shared workspace (GitHub project, Notion/Trello board for task tracking).