

An AI-Powered Legal Case Retrieval and Risk Analysis Platform Using Retrieval-Augmented Generation (RAG)

Borra Malika¹ and Burra Sathwika²

^{1,2} Students, Department of Emerging Technologies (CSE-AI&ML),
Mahatma Gandhi Institute of Technology, Hyderabad, Telangana — 500 075

bmalika_csm226606@mgit.ac.in

bsathwika_csm226607@mgit.ac.in

Guide: Ms. Swetha Sharma, Assistant Professor, Department of Emerging Technologies,
Mahatma Gandhi Institute of Technology, Hyderabad, Telangana — 500 075

Abstract - Legal research is increasingly challenged by the rapid growth of judicial data and the limitations of traditional keyword-based systems, which often fail to capture contextual meaning. LegalEase AI addresses this issue through an intelligent platform that leverages Retrieval-Augmented Generation (RAG) and Natural Language Processing (NLP) to enhance case discovery and reasoning. It enables users to input natural language queries and retrieves semantically relevant case laws while providing contextual summaries and predictive risk insights, simplifying research for practitioners and students. The platform features a user-friendly interface with search, analytics, and chatbot features.

Results show improved accuracy and efficiency in case retrieval, with strong semantic relevance. The system follows a structured pipeline including data ingestion, preprocessing, semantic embedding, retrieval, and AI-driven reasoning. Legal data is stored in PostgreSQL and converted into vector embeddings using SentenceTransformer models, which are indexed with FAISS for efficient similarity search. A FastAPI backend manages query processing and integrates LLaMA-based models for explanation and risk evaluation, while a Next.js Progressive Web Application provides an interactive and consistent risk analysis. The system demonstrates how RAG and transformer models enable contextual understanding and scalable, user-friendly legal research solutions.

Keywords: Retrieval-Augmented Generation (RAG), Legal AI, FAISS Vector Database, Natural Language Processing (NLP), Legal Case Retrieval, Risk Analysis, Semantic Search.

I. INTRODUCTION

Legal research has become increasingly complex due to the rapid growth of judicial data and the limitations of traditional keyword-based search systems, which often fail to capture contextual meaning, forcing users to manually sift through large volumes of case law and reducing efficiency. To address this, LegalEase AI is introduced as an intelligent platform that leverages Retrieval-Augmented Generation (RAG) and Natural Language Processing (NLP) to enable semantic case retrieval and analysis. The system allows users to input natural language queries and retrieves contextually relevant cases while providing summaries, risk analysis, and predictive insights, significantly improving accessibility and productivity. Technically, the platform consists of three core components: a Progressive Web Application frontend for interactive user experience, a backend for data processing and API management, and an AI layer that integrates semantic embeddings, FAISS-based retrieval, and large language models to deliver accurate and context-aware legal insights.

Regarding the technical details, there are three main components of the LegalAI : frontend, backend, and AI model.

A. Motivation

Legal research is a time-intensive process that requires analyzing large volumes of case laws and legal documents. Traditional systems often fail to provide context-aware results, making it difficult for users to quickly find relevant information. There is a strong need for an intelligent system that can simplify legal research, reduce manual effort, and provide meaningful insights. LegalEase AI is motivated by the goal of improving efficiency,

accessibility, and decision-making in the legal domain using advanced AI techniques.

B. Problem Definition

Legal practitioners and students face significant challenges in retrieving relevant case laws due to the limitations of keyword-based search systems. These systems do not understand the semantic meaning of queries, leading to irrelevant or excessive results. Additionally, users must manually interpret legal texts and predict outcomes without intelligent assistance. The problem is to develop a system that can understand natural language queries, retrieve contextually relevant cases, and provide automated analysis and insights.

C. Existing System

Current legal research platforms such as online legal databases primarily rely on Boolean keyword searches and citation-based filtering. While they provide access to vast repositories of case laws, they lack semantic understanding and AI-driven insights. Users must manually refine queries and interpret results, which is time-consuming and inefficient. Moreover, these systems do not offer features like predictive analysis, contextual summaries, or interactive query handling through conversational interfaces.

D. Proposed System

The proposed system, LegalEase AI, utilizes Retrieval-Augmented Generation (RAG) and Natural Language Processing (NLP) to enable intelligent legal case retrieval and analysis. It converts legal documents into semantic embeddings and uses FAISS for efficient similarity search. The system also integrates large language models to generate summaries, risk assessments, and contextual explanations. A Progressive Web Application frontend provides an interactive interface, including search, analytics, and chatbot features, making legal research more efficient and user-friendly.

II. LITERATURE SURVEY

Research in intelligent legal information systems has been rapidly expanding with the increasing digitization of judicial data. Most existing legal research platforms primarily focus on providing access to large repositories of case laws rather than enabling meaningful interaction and contextual understanding. However, recent advancements in artificial intelligence, natural language processing, and retrieval-augmented generation have opened new possibilities for semantic case retrieval and automated legal reasoning. This paper examines how these technologies can be effectively integrated into legal research systems to enhance efficiency, accuracy, and user interaction. Table I highlights selected literature related to AI-driven legal analytics and semantic retrieval systems.

S.no.	Year	Author	Title	Methodology
1	2024	J. Zhang & L. Wang	AI-Enhanced Legal Case Retrieval Systems	Uses LLMs with transformer embeddings for semantic retrieval and summarization
2	2024	P. Shah & L. Nair	CaseLawRAG System	Combines FAISS vector search with GPT-based RAG pipeline
3	2023	R. Krishna & M. Iyer	AI-Powered Legal Research Using BERT	Uses BERT embeddings for semantic similarity matching
4	2022	A. Bhattacharjee et al.	LegalBERT	Domain-specific transformer trained on legal corpus
5	2020	T. Kim & D. Park	PWA-Based Legal Analytics Platform	Uses PWA architecture for cross-platform deployment

III. METHODOLOGY

A. Dataset Collection

The dataset for LegalEase AI consists of structured and unstructured legal case data collected from publicly available legal repositories and synthetic datasets. Each case includes metadata such as case title, court, year, legal sections, judgment summary, and verdict. The data is stored in JSON format and later ingested into a PostgreSQL database. Additionally, textual content from judgments is extracted and prepared for semantic processing, forming the foundation for training and retrieval operations.

B. Data Preprocessing

Data preprocessing involves cleaning and transforming legal text into a structured format suitable for semantic analysis. This includes removing special characters, normalizing text, tokenization, and lemmatization. The processed text is then converted into vector representations using embedding models. User queries are also preprocessed similarly to ensure consistency in comparison with stored case embeddings, enabling accurate semantic matching.

C. Model Selection

The system uses SentenceTransformer (all-MiniLM-L6-v2) to generate embeddings for legal text, as it provides efficient and meaningful semantic representations. For generating explanations, summaries, and risk insights, a LLaMA-based large language model is used. This combination ensures both fast retrieval and context-aware reasoning.

D. Backend Development

The backend is implemented using FastAPI and is responsible for handling API requests such as case search, analysis, and chatbot queries. It integrates PostgreSQL for storing case data and FAISS for performing fast vector similarity search. The backend ensures smooth communication between the database, retrieval system, and AI models.

E. Frontend Development

The frontend is developed using Next.js as a Web Application. It provides a simple interface where users can enter legal queries, view retrieved case results, analyze risk insights, and interact with a chatbot. The focus is on clarity, responsiveness, and ease of use.

F. AI Integration

The system uses a Retrieval-Augmented Generation (RAG) approach. FAISS retrieves the most relevant cases based on embeddings, and the LLaMA model generates summaries, explanations, and risk analysis based on the retrieved content and user query. This ensures accurate and context-aware outputs.

G. Deployment

The application is containerized using Docker and deployed using Docker Compose. It includes services for the FastAPI backend, PostgreSQL database, and frontend application. The system can be deployed on cloud platforms and is designed to handle queries efficiently with low latency.

IV. SYSTEM DESIGN

A. Dataflow Diagram

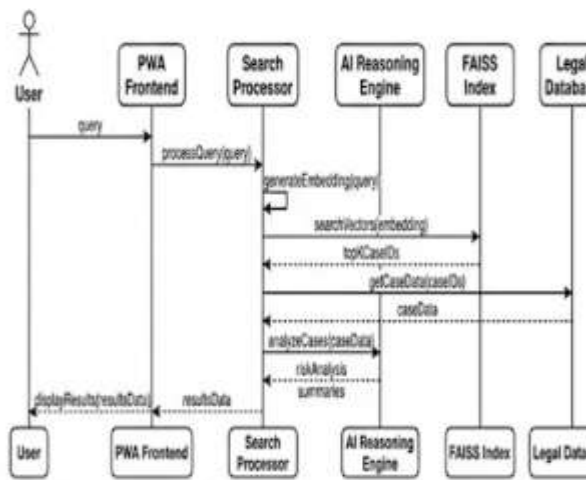
Figure 1 illustrates the overall data flow of the LegalEase AI system. The process begins with the Legal Case Database, where structured case records (including summaries, judgments, and metadata) are stored in PostgreSQL. These cases are preprocessed and converted into vector embeddings using a SentenceTransformer model, which are then indexed in the FAISS vector database. When a user submits a legal query through the PWA frontend, the query is sent to the backend API, where it is processed and converted into an embedding. The FAISS engine retrieves the top-k most semantically similar cases, which are then fetched from PostgreSQL. These results are passed to the AI engine (LLaMA), which generates summaries, explanations, and risk analysis before sending the final response back to the frontend for display.



[Fig. 1: Dataflow Diagram]

B. Sequence Diagram

Figure 2 depicts the interaction between the User, Frontend (PWA), Backend API, FAISS Engine, Database, and AI Engine. The sequence starts when the user enters a legal query. The frontend sends the query to the backend, which processes it and generates an embedding. The backend then queries the FAISS index to retrieve similar case vectors. Corresponding case data is fetched from PostgreSQL and passed along with the user query to the AI engine. The AI model generates contextual insights and risk analysis, which are returned to the backend and then displayed to the user via the frontend interface.



[Fig. 2:Sequence Diagram]

C. Workflow Diagram

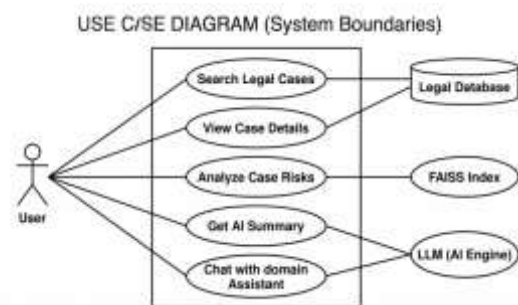
Figure 3 represents the end-to-end workflow from the user's perspective. The process begins when the user opens the LegalEase AI application and enters a legal query. The system processes the query, retrieves relevant cases, and performs AI-based analysis. The results, including case summaries and risk insights, are displayed on the dashboard. The user can further interact with the chatbot to ask follow-up questions, enabling continuous and interactive legal exploration.



[Fig. 3:Workflow Diagram]

D. Use Case Diagram

Figure 4 shows the primary interactions between the User and the LegalEase AI system. The user can perform actions such as entering legal queries, viewing retrieved case results, analyzing risk insights, and interacting with the AI chatbot. The system components (FAISS Engine, AI Model, and Database) support these actions in the background. The diagram highlights the simplicity, interactivity, and efficiency of the platform in enabling intelligent legal research.



[Fig. 4:Usecase Diagram]

V. TESTING AND RESULTS

A. Testing

A comprehensive testing process was carried out to evaluate different components of the LegalEase AI system:

- **Functional Testing:** Verified that core functionalities such as legal query input, semantic case retrieval, risk analysis, and chatbot responses work correctly.
- **Unit Testing:** Individual modules including query processing, embedding generation, FAISS retrieval, and AI response generation were tested independently.
- **Integration Testing:** Ensured smooth interaction between the frontend (PWA), backend (FastAPI), PostgreSQL database, FAISS engine, and LLM-based AI services.
- **UI Testing:** Evaluated the frontend interface for usability, responsiveness, and ease of navigation across different devices.
- **Data Validation Testing:** Checked the integrity and correctness of legal case data and validated user queries before processing.
- **Response Accuracy Assessment:** Assessed the relevance and correctness of retrieved cases and AI-generated summaries, explanations, and risk insights.
- **Performance Testing:** Measured system response time for query processing, semantic retrieval using FAISS, and AI-based reasoning to ensure efficient and low-latency performance.

B. Results

As shown in Fig. 5, the interface of the *LegalEase AI* platform presents a clean and user-friendly landing page for semantic legal search. The interface includes a central search bar where users can describe their legal scenario in natural language. Suggested query prompts such as property disputes, cyber fraud, and cheque bounce cases are provided to guide users. On the right side, a Legal RAG Assistant chat panel is embedded, enabling interactive communication. The interface is designed to simplify legal query input and enhance accessibility for users without legal expertise.



[Fig. 5: Application Interface]

Fig. 6 illustrates the analytics dashboard generated after a user submits a query. The system provides

visual insights such as verdict distribution using a donut chart and case frequency over the years through a bar graph. These visualizations help users quickly understand trends in similar legal cases. The layout ensures that complex legal data is presented in an intuitive and interpretable manner, supporting better decision-making.

[Fig. 6: Case Law Analytics Dashboard]

Fig. 7 depicts the detailed results section of the platform. It displays relevant case precedents along with similarity scores, court details, and concise summaries. Each case includes an option to view further explanations. On the right side, a Risk Assessment panel provides predictive insights such as possible legal consequences and suggested next steps. This integration of retrieval and analysis enables users to not only explore past cases but also understand potential outcomes and actions..



[Fig. 7: Case Retrieval and Risk Assessment Interface]

VI. CONCLUSION

The development of LegalEase AI represents a significant advancement in the domain of intelligent legal research systems. The proposed platform effectively addresses the limitations of traditional keyword-based legal search tools, which



often lack contextual understanding and require extensive manual effort. By integrating Retrieval-Augmented Generation (RAG) with Natural Language Processing, the system enables users to input natural language queries and receive semantically relevant case laws along with contextual summaries and risk analysis. This approach enhances efficiency, reduces research time, and improves the overall quality of legal decision-making.

The implementation of a full-stack architecture ensures seamless interaction between the frontend, backend, database, and AI components. The use of a Progressive Web Application (PWA) enables cross-platform accessibility and a user-friendly interface, while the integration of FAISS and transformer-based models ensures fast and accurate semantic retrieval. Additionally, the incorporation of large language models provides explainable insights, making the system both practical and intuitive for users. Overall, the project demonstrates how AI-driven technologies can transform legal research into a more accessible, efficient, and intelligent process.

REFERENCES

- [1] J. Zhang and L. Wang, "AI-Enhanced Legal Case Retrieval Systems," *Journal of Computational Law and AI*, 2024.
- [2] P. Shah and L. Nair, "CaseLawRAG: AI-Driven Legal Precedent Retrieval Using RAG," *IEEE Transactions on AI in Legal Systems*, 2023.
- [3] R. Krishna and M. Iyer, "Semantic Legal Search Using BERT-Based Models," *International Conference on Legal Informatics*, 2023.
- [4] A. Bhattacharjee et al., "LegalBERT: Pre-Trained Transformer Models for Legal NLP," *ACL Proceedings*, 2022.
- [5] R. Das and V. Iyer, "Vector-Based Legal Text Retrieval Using FAISS," *ACM Symposium on Information Retrieval*, 2022.
- [6] T. Kim and D. Park, "Progressive Web Applications for Legal Analytics Platforms," *International Journal of Web Technologies*, 2020.
- [7] S. Brown and J. Miller, "Explainable AI in Legal Decision Support Systems," *IEEE Access*, 2021.
- [8] M. Gupta and K. Rao, "Natural Language Processing for Legal Document Analysis," *Journal of AI Research*, 2021.