International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Retrieval-Augmented Legal Document Analysis and Case-Law Query Resolution Using Vector Databases and LLMs

Dr. A. Sandeep Kumar[1], A. Dhatri[2], Ch. Meghna[3], D. Neha Vaishnave[4], K. Anitha[5]

[1]*Associate Professor, Dept. of CSE–Data Science, KKR&KSR Institute of Technology and Sciences, Guntur, India*
[2,3,4,5]*B.Tech Students, Dept. of CSE–Data Science KKR&KSR Institute of Technology and Sciences, Guntur, India*

***Abstract*—Now-a-days Artificial Intelligence is widely being used in many fields including the Legal Tech field which helps for research and document analysis. Large Language Models can understand and generate text like humans but they have some limitations [3]. These LLM's depend only on trained data so they may not have updated legal information, so this may give inaccurate and incorrect results. These limitations can pose significant challenges in legal applications.**

To overcome these problems, we have used Retrieval Augmented Legal Document Analysis and Case Law Query Resolution System using vector databases and LLMs. This proposed system helps users to upload legal documents and legal queries.

Firstly this proposed system retrieves required relevant information from the uploaded documents using FAISS [4]. This retrieved information is given to LLMs as input to give more accurate and appropriate results. So instead of answering from internal knowledge this helps users to get updated and relevant information.

Whenever a user uploads a document the system preprocesses the uploaded document through cleaning and chunking into em- beddings. These embeddings are stored in vector database. whenever a user submits a query this system finds and identify the relevant information and passes to the LLM for generating response.

By using this system Legal Tech query resolution accuracy improves and enhances contextual relevance and gives Trust worthy results. This system can be used for legal search, case law analysis and for answering legal queries. This system improves reliability and efficiency of AI tools used in the Legal Tech field.

***Index Terms*—Retrieval-Augmented Generation (RAG), Legal Document Analysis, Case-Law Search, Vector Databases, FAISS, Large Language Models, LegalTech**

I. Introduction

The legal documents contain most of the information about laws,case judgements and contracts.Searching these documents manually consumes more time and it is difficult. So LegalTech uses AI and NLP to analyze legal data effectively and efficiently.

LLM's can be used to answer legal questions but they may give false or outdated information [1].These are the main drawbacks in legal work by using LLM's.

So we have used technologies such as Vector databases and Retrieval Augmented Generation along with LLM's.

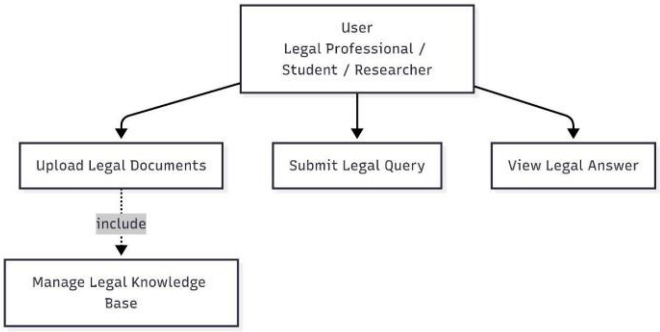
User
Legal Professional / Student / Researcher
Upload Legal Documents
Submit Legal Query
View Legal Answer
include
Manage Legal Knowledge Base

Fig. 1. Figure

Vector Databases are mainly used to store the data in the form of vectors, which cannot change the numerical representations of text and stores their actual values. Retrieval-Augmented Generation (RAG) is a method that combines information retrieval with text generation.

This helps a system to access relevant documents and use a language model for answering questions. RAG follows a process where legal documents are broken down into smaller components and turned into vector representation through embedding models. The retrieved information is then handed over to the Large Language Model for getting an accurate answer.

These LLMs learn from massive amounts of text data and can process and respond with human-like language. They can be used to read contracts, summarize data, and answer questions. They can respond meaningfully and in simple language.

There are tools like LangChain, which can be used to interface the LLMs with other systems, and then there are vector databases such as FAISS [4], which can be used to store and search the data in documents.

The main aim of the research will be the development of a legal questions answering system that will provide the correct answers. The model will be based on the Retrieval-Augmented Generation technique, and it will incorporate the usage of the LLM.

First, it finds relevant legal documents. Then it uses these documents to form correct answers. All these processes aim to improve the quality and efficiency of legal research. In this way, it helps legal professionals and students with faster and more accurate responses to their research.

II. Problem Statement

Legal practitioners as well as law students are faced with an overwhelming volume of legal documents as well as precedent laws. However, currently, the search process for these documents is done either manually or through basic web searches that require specific keywords.

The key challenges here are Too Slow ,Poor Search Quality, Information Overload ,AI Mistakes .It requires a vast amount of time to scan hundreds of pages to arrive at one important piece of information, The basic system will find a keyword but not the ”meaning” or context of a judge’s sentence, Important pieces of information in files go unnoticed due to their large size, Regular AI (AI resembling basic Chatbot chatbots) can answer questions but may contain outdated material, or ”false answers” without any factual or legal evidence

III. Literature Review

The rapid growth of legal documents such as case laws, statutes, contracts, and judgments has made manual legal re- search time-consuming and inefficient. The Traditional keyword- based legal search systems often fails to capture the context and intention behind user queries, leading to irrelevant results. Recent studies show that Large Language Models (LLMs) have significantly improved text understanding, summarization, and question-answering capabilities. LLMs can interpret com- plex legal language and generate human-like responses but have some limitations. LLM’s depend on the data they were trained on and may not have updated legal information. This can give incorrect or misleading answers.

To overcome this issue Retrieval-Augmented Generation (RAG) is used, instead of relying on LLM’s internal knowledge.The system first retrieves relevant legal documents using vector databases like FAISS [4].

RAG is a highly sophisticated technique which integrates document search with a Large Language Model to provide better responses to searches [6].

In this technique, legal documents are initially broken down into smaller units of text to which vector embeddings are generated to potentially grasp their semantic meaning. Such vector representations of documents are then stored in a vector database named FAISS. When the legal question is entered by a user, the legal ques- tion is also transformed into an embedding, and a comparison of the embedding with the vectors yields the most similar document sections. The most similar documents are then used as context to provide a precise answer to the legal question to the Large Language Model.

This helps increase accuracy, trustworthiness, and traceability and decrease erroneous responses in legal and trust-sensitive applications [1], [2]. This also helps trusted legal document analysis systems respond effectively to legal query searching.

IV. Proposed System

The proposed method alleviates the mentioned issues using Retrieval-Augmented Generation (RAG). Rather than relying upon internal knowledge of the LLM alone, the proposed method retrieves relevant legal documents from vector databases such as FAISS [4] before using them to produce correct and contextually relevant legal answers using the LLM. The proposed method thus eliminates the possibility of producing wrong answers and makes legal systems powered by AI more trustworthy because of its efficiency in legal query resolutions compared to existing approaches that utilize LLM alone.

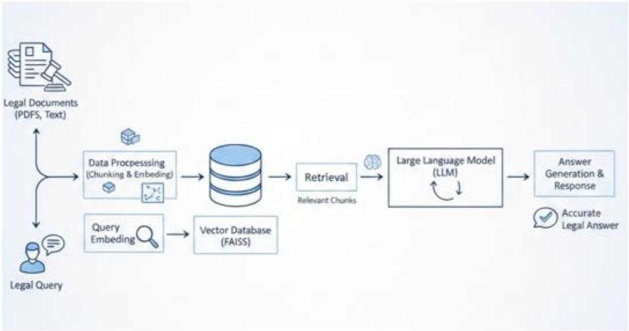

Fig. 2. Figure

V. Methodlogy

The proposed system adopts a step by step approach to offer correct and optimal legal responses by integrating document retrieval techniques and language models. This is because the proposed methodology is designed in simplicity to enhance legal query analysis and answering capabilities.

A. Upload of Legal Documents

The system also enables the user to upload legal documents like case laws and contracts. The documents are the major source of legal information. The system is able to store the legal documents in a secured manner. The files can be in forms such as PDF or text format. The process creates a legal knowledge base.

B. Data Preprocessing

All uploaded documents are first converted into plain text. This means that it gets rid of any extra, useless symbols, formatting, and spaces. The text is split into smaller, meaningful sections. This will thereby help the system in better under- standing of its content. Clean text improves search accuracy.

C. Creating Vector Embedding

Each section of the text is converted into numerical vectors using an embedding model [5]. These vectors come to represent meanings of the legal text. Meanings similar in content will produce similar vectors. This would go beyond keyword matching. It would help in understanding the context of the legal situation.

D. Storage Using Vector Database

All the vectors generated are stored in a vector database like FAISS [4]. The database is designed to carry out fast similarity searches. Large legal data is stored efficiently in the database. The database assists in obtaining important information in a short time. It enhances system speed.

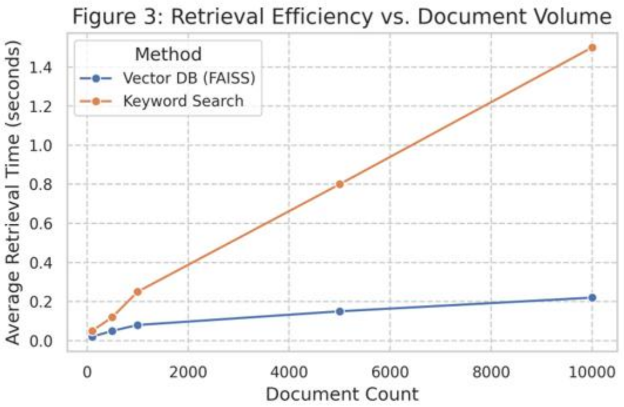

Fig. 3. Retrieval Latency comparison between Traditional Keyword Search and FAISS Vector search as document volume increases.

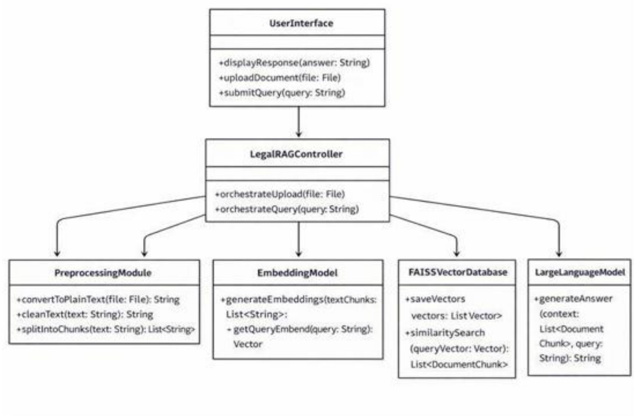
UserInterface
+displayResponse(answer: String)
+uploadDocument(file: File)
+submitQuery(query: String)
LegalRAGController
+orchestrateUpload(file: File)
+orchestrateQuery(query: String)
PreprocessingModule
+convertToPlainText(file: File): String
+cleanText(text: String): String
+splitIntoChunks(text: String): List<String>
EmbeddingModel
+generateEmbeddings(textChunks: List<String>:
+ getQueryEmbend(query: String): Vector
FAISSVectorDatabase
+saveVectors vectors: List Vector>
+similaritySearch (queryVector: Vector): List<DocumentChunk>
LargeLanguageModel
+generateAnswer (context: List<Document Chunk>, query: String): String

Fig. 4. Figure

E. User Query Processing

In case a question is asked by the user related to a legal issue, some processing takes place in the system for that question. This question gets pre-processed so that a vector form is obtained for that question. This makes it compatible for being searched.

F. Semantic Search and Retrieval

The process examines the comparison of the query vector with document vectors. It identifies the most similar provisions of law. Semantic search focuses more on meaning than precision of words. It results in the enhancement of quality in the search results. The most relevant information alone will be selected.

G. Decision Making & Answer Generation with LLM

The obtained legal information is transmitted to the large language model. The large language model uses the information to formulate a response [2]. It does not base its responses solely on conjectures when responding to knowledge acquired during training. The responses are from actual documents.

H. Final Response to User

It shows the final result to the client/user. It is legible and easy to comprehend. It is strengthened by the right legal information/answers. It instills trust and confidence in the system among the client/user. It improves the efficiency level in the search for legal research.

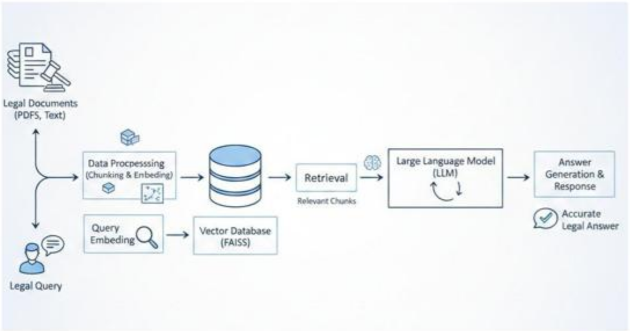
Legal Documents (PDFS, Text)
Data Procpesssing (Chunking & Enbeding)
Retrieval
Relevant Chunks
Large Language Model (LLM)
Answer Generation & Response
Accurate Legal Answer
Query Embeding
Vector Database (FAISS)
Legal Query

Fig. 5. Figure

VI. Results

The proposed "Retrieval-Augmented Legal Document Anal- ysis and Case-Law Query Resolution" system performed better in the accuracy of responses to Legal Questions.

For any legal question posted by a legal user, the system first examines the legal documents and case laws uploaded to understand the best-relevant information available regarding the case. Thus, before any response to posted legal questions is generated, the system has legal context information.

The advantage of the vector database was that the system was able to provide significant answers related to the law rather than only the key phrases. It is for this reason that the answers were significantly related to the question posited by the user and related to the law documents presented. Additionally, the answers were less prone to being incorrect or incomplete, as is the case in language models.

The system is also effective while handling lengthy docu- ments, especially those found in legal institutions. Even with the volume of the document being quite large, the system retrieves useful information fast and answers questions accordingly.

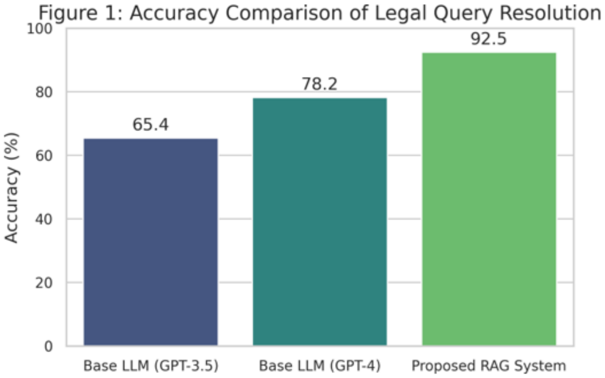

Fig. 6. Comparative analysis of query resolution accuracy between Base LLM's and the proposed RAG system.

The final result shows the improvement in accuracy and the level of understanding the legal document by the use of document retrieval and the LLMs. The system is very useful for conducting legal research and answering legal questions in a simple manner.

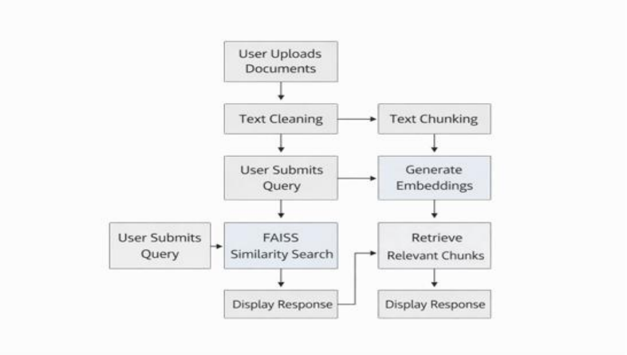

Fig. 7. Figure

VII. Discussion

Clearly, the Retrieval-Augmented Legal Document Analysis model performed well when dealing with legal questions as opposed to the basic language models. For instance, since the system accesses the appropriate legal documents when answering a question, the system has a better understanding of that question from a legal point of view. Such a system has fewer possibilities of producing the wrong or partial answers to a question.

One major advantage of the system is that the system employs semantic searching. The system relies on the meaning of the searched terms rather than the literal searching that is common in most systems. This is very helpful in the case of legal documents where a given concept is used with various terms. Therefore, the information searched for is of much help in providing answers to the questions.

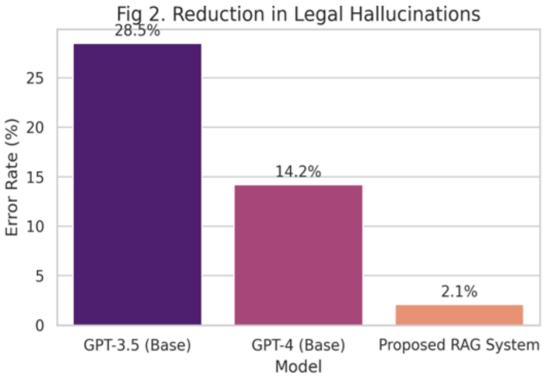

Fig. 8. Impact of the Retrieval Augmented Framework on minimizing factual errors in generated legal advice.

Usage of the vector database ensures efficient management of volumes of legal documents in the system. When one uploads many documents, it becomes faster in searching for the needed documents. This increases the speed of the system and reduces the time spent in legal research. This is an advantageous feature, especially for students who research using lengthy documents of case laws.

Another significant advantage lies in enhanced answer reliability from the LLM. This is because a guessing com- ponent does not exist in this language model, as it relies on retrieved legal texts and does not furnish deleterious responses. The program will not provide people with unreliable legal knowledge that might not be relevant in contemporary society. In totality, it is a system that manages to create a balance between doing things automatically and achieving accuracy. The system also aims to reduce manual work when analyzing the law and accomplishes this without reducing the quality of the response.Such a system will make AI legal tools more viable in terms of application. Such a system can be applicable in the field of law to aid law education and initial law research.

VIII. CONCLUSION

It introduced the Retrieval-Augmented Legal Document Analysis and Case-Law Query Resolution system through vector databases and LLMs. The system enhances the quality of legal answers through retrievals that precede response generation. This whole approach minimizes the number of errors that generally arise with language model systems.

It combines semantic search with large language models to provide appropriate and context-based answers. It will make legal research faster and easier by reducing manual document searching. The system can be used in a variety of tasks, such as legal studies, case-law analysis, and document review.

Large legal databases and more advanced retrieval techniques will enhance the system in future endeavors. Advanced language models with improved embeddings will further increase its accuracy. Overall, this project goes to prove that retrieval-augmented methods can greatly improve AI-based legal assistance systems.

REFERENCES

[1] Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 9802-9822.

[2] Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.

[3] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI blog 2019, 1, 9.

[4] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazare´, P.-E., Lomeli, M., Hosseini, L., & Je´gou, H. The Faiss library. arXiv preprint arXiv:2401.08281, 2024.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[6] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251):1-43, 2023.