Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Know Your Law: A Conventional AI Chatbot for Legal Awareness Using Retrieval Augmented Generation

A Sri Pranav
Department Of CSE
R.M.K. Engineering College Kavarapettai, Thiruvallur

Balaji J
Department Of CSE
R.M.K. Engineering College Kavarapettai, Thiruvallur

Diwakar R
Department Of CSE
R.M.K. Engineering College Kavarapettai, Thiruvallur

Dr. PaulRaj D
Professor, Department Of CSE
R.M.K. Engineering College Kavarapettai, Thiruvallur

Abstract **- Access to legal information remains challenging for ordinary citizens due to the complexity and technical nature of statutory documents such as the Indian Penal Code (IPC). Traditional search systems often require precise legal terminology and return lengthy, difficult-to-understand texts. This paper presents Know Your Law, a Retrieval-Augmented Generation (RAG) based Conversational AI chatbot designed to simplify and improve access to legal knowledge. The system integrates hybrid retrieval using dense embeddings (Sentence Transformers) and sparse retrieval (BM25), supported by Pinecone vector storage, while a Large Language Model (LLaMA via Groq API) generates simplified, context-grounded responses. To ensure accuracy and ethical compliance, the system incorporates structured response validation, query filtering, and mandatory legal disclaimers. Experimental results demonstrate improved retrieval precision, reduced hallucination risk, and effective simplification of complex legal provisions, highlighting the potential of hybrid RAG systems in democratizing legal awareness.**

Index Terms **- Legal Informatics, Retrieval-Augmented Generation (RAG), Conversational AI, Hybrid Information Retrieval, Natural Language Processing (NLP), Legal Awareness Systems, Large Language Models (LLMs), Vector Databases, Pinecone, AI-based Legal Assistance.**

I. INTRODUCTION

Legal systems are designed to maintain order and justice, yet they remain inaccessible to the average citizen. Statutory documents are lengthy, highly structured, and filled with technical terminology. Understanding rights, penalties, and legal definitions often requires professional assistance.

Artificial Intelligence, particularly Natural Language Processing (NLP), offers an opportunity to transform legal texts into understandable knowledge. Conversational AI systems allow users to interact in natural language, removing the need for technical search syntax.

The Know Your Law system aims to:

- Convert complex legal text into simplified explanations.
- Allow natural language queries.
- Ground responses strictly in authoritative legal documents.
- Prevent misuse through strict legal query filtering.

A. Problem Statement

Although legal documents such as the Indian Penal Code (IPC) are publicly available, they are written in complex language and structured in a way that is difficult for ordinary citizens to interpret. Traditional search systems depend heavily on exact keywords and legal terminology, making them ineffective for users who express queries in simple, everyday language. Furthermore, AI systems that generate responses without grounding in authoritative sources may produce inaccurate or misleading legal information. Hence, there is a need for a reliable, context-grounded conversational system that can accurately retrieve relevant legal provisions, simplify them for better understanding, and operate strictly within the limits of legal awareness without offering personalized legal advice.

B. Scope of the Project

The scope of the project is to design and implement a Retrieval-Augmented Generation (RAG) based Conversational AI chatbot that delivers simplified legal information to enhance public legal awareness. The system includes document processing, hybrid retrieval using dense and sparse embeddings, structured response generation through a Large Language Model, and safety mechanisms such as legal query filtering and mandatory disclaimers. It supports natural language interaction and ensures responses are grounded in authoritative legal documents. The project is limited to providing general legal information and does not include case-specific guidance, legal representation, or professional legal advisory services.

II. LITERATURE SURVEY

[1] Analyzed the limitations of traditional keyword-based information retrieval systems in legal databases, highlighting the inefficiency of Boolean search mechanisms in handling complex statutory queries. The study emphasized the need for intelligent semantic retrieval methods to improve

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

access to legal documents and enhance user understanding in legal informatics systems.

[2] Explored the use of conversational AI systems in legal consultation scenarios using Recurrent Neural Networks (RNNs). The research demonstrated improvements in main-taining contextual continuity during multi-turn conversations but identified challenges related to factual accuracy and relia-bility, which are critical requirements in legal applications.

[3] Introduced the Retrieval-Augmented Generation (RAG) framework for knowledge-intensive NLP tasks, separating retrieval and generation processes to reduce hallucinations. The study demonstrated that grounding responses in retrieved documents significantly improves factual consistency, making RAG architectures suitable for domain-specific applications such as legal information systems.

[4] Proposed Sentence-BERT (SBERT), a Siamese network-based approach for generating dense sentence embeddings to enable efficient semantic similarity search. Experimental results showed significant improvements in semantic retrieval performance compared to traditional embedding methods, making it highly applicable for legal document indexing and retrieval.

[5] Investigated hybrid retrieval strategies combining dense semantic embeddings and sparse keyword-based methods such as BM25. The study demonstrated that hybrid approaches outperform individual retrieval techniques in domain-specific question-answering systems by balancing conceptual relevance with exact keyword matching.

[6] Examined structured information extraction from un-structured legal PDF documents and emphasized the impor-tance of maintaining section boundaries during text process-ing. The research highlighted chunking strategies and prepro-cessing techniques to prevent knowledge fragmentation and improve retrieval accuracy in legal AI systems.

[7] Discussed the application of Artificial Intelligence in legal analytics and legal practice, outlining the potential of AI tools to enhance legal research, decision support, and public legal awareness. The study emphasized ethical considerations, transparency, and the importance of defining boundaries be-tween informational assistance and professional legal advice.

[8] Provided technical documentation for efficient inference of Large Language Models (LLaMA), highlighting high-speed response generation and optimized deployment for real-time applications. The study supports the feasibility of integrating high-performance LLM inference into conversational legal systems.

[9] Presented vector database architectures supporting hy-brid search mechanisms, enabling scalable storage and high-speed similarity search for large embedding datasets. The study demonstrated improved retrieval efficiency and scalabil- ity for domain-specific AI applications such as legal knowl- edge systems.

[10] Discussed normalization techniques in data preprocess-ing to enhance machine learning model stability and perfor-mance. The study highlighted the importance of consistent data transformation methods in embedding-based retrieval pipelines to ensure reliable similarity scoring.

III. OVERVIEW OF EXISTING SYSTEM

Existing legal information systems primarily rely on traditional search engines, static government portals, or rule-based expert systems to provide access to legal documents. These systems typically require users to enter specific keywords, legal citations, or section numbers to retrieve relevant statutes. The output is usually presented as full-length legal text in the form of PDF documents or structured web pages, without any simplification or contextual explanation. While these platforms provide authentic legal content, they do not assist users in understanding complex legal language or interpreting statutory provisions in practical terms.

Moreover, most existing systems lack conversational capabilities, meaning users cannot interact naturally or seek clarifications in simple language. The reliance on keyword-based retrieval limits semantic understanding, often leading to irrelevant results if the exact legal terminology is not used. Additionally, traditional systems do not support dynamic indexing of newly uploaded legal documents or provide structured, simplified summaries. As a result, although legal information is technically accessible, it remains practically difficult for common citizens to comprehend and utilize effectively.

IV. PROPOSED SOLUTION

The proposed solution is the development of Know Your Law, a Retrieval-Augmented Generation (RAG) based Conversational AI chatbot that provides simplified and context-grounded legal information through natural language interaction. The system utilizes a hybrid retrieval approach combining dense semantic embeddings (Sentence Transformers) and sparse keyword-based retrieval (BM25) to accurately identify relevant legal provisions from authoritative sources such as the Indian Penal Code (IPC). Retrieved content is then processed by a Large Language Model (LLaMA via Groq API) to generate structured and easy-to-understand explanations while strictly grounding responses in the retrieved context. The solution also incorporates PDF document processing, intelligent text chunking, vector indexing using Pinecone, and a Legal Question Filtering Module to ensure that only law-related queries are answered. Additionally, mandatory disclaimers and response validation mechanisms are implemented to maintain ethical boundaries, prevent hallucinations, and ensure the system functions solely as a legal awareness tool without providing personalized legal advice.

V. METHODOLOGY

The proposed system follows a **Retrieval-Augmented Generation (RAG) based architecture** designed to provide simplified legal information through a conversational AI

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

chatbot. The methodology integrates Natural Language Processing (NLP), information retrieval techniques, and Large Language Models (LLMs) to ensure accurate and context-grounded responses. The entire workflow consists of multiple stages including document acquisition, preprocessing, embedding generation, hybrid retrieval, and response generation.

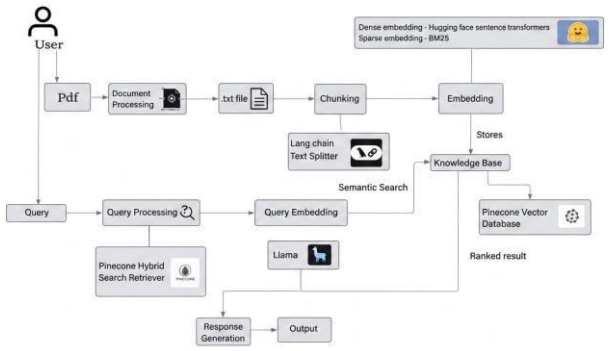

Fig. 1. Architecture Diagram

A. Data Acquisition

The first stage involves collecting authoritative legal documents that serve as the knowledge base for the system. Official legal sources such as the **Indian Penal Code (IPC)** and other legal documents are obtained and uploaded through the user interface. The system allows users to upload legal PDF files using a file uploader integrated into the web interface. These documents are temporarily stored and passed to the processing pipeline for further analysis.

B. PDF Text Extraction

Once the document is uploaded, the system extracts textual content from the PDF file. The **PyPDF2 library** is used to read the PDF file page by page and extract the textual information embedded within the document. The extracted text from each page is concatenated to form a single continuous text string representing the entire legal document. This stage converts unstructured PDF data into machine-readable text that can be further processed by the system.

C. Text Preprocessing

The extracted text often contains unwanted formatting such as page numbers, headers, special characters, and excessive whitespace. Therefore, a preprocessing stage is applied to clean the data. Regular expressions and Python string processing techniques are used to remove unnecessary symbols and normalize spacing. This step ensures that the textual data becomes structured and suitable for semantic analysis and indexing.

D. Text Chunking

Legal documents are typically lengthy and complex; therefore, the cleaned text is divided into smaller, semantically meaningful segments. The system employs a chunking strategy using the **LangChain text splitter**, where the document is divided into chunks of approximately **500 characters with an overlap of 100 characters** between consecutive chunks. This overlap helps maintain contextual continuity and prevents information loss during retrieval.

E. Hybrid Embedding Generation

To enable efficient semantic search, the system generates vector representations of each text chunk using a hybrid embedding approach consisting of both dense and sparse embeddings.

1. Dense Embeddings

Dense embeddings capture the semantic meaning of the text. Each chunk is processed using a **Sentence Transformer model (all-MiniLM-L6-v2)**, which converts the text into a 384-dimensional numerical vector representing its semantic content. These embeddings allow the system to understand the contextual meaning of legal queries.

2. Sparse Embeddings

Sparse embeddings are generated using the **BM25 algorithm**, which focuses on keyword-based matching. This method calculates the importance of terms within documents using term frequency and inverse document frequency.
Sparse embeddings are particularly useful when users refer to specific legal terms or section numbers.

The combination of dense and sparse embeddings enables the system to perform both semantic and keyword-based retrieval effectively.

F. Vector Storage

The generated dense embeddings, along with associated metadata, are stored in the **Pinecone vector database**. Pinecone enables efficient indexing and similarity search over high-dimensional vector data. Each stored vector is associated with metadata such as the source document and chunk identifier, enabling accurate retrieval of legal content during query processing.

G. Query Processing

*When a user submits a query through the chatbot interface, the query first passes through a **legal query filtering module** that ensures the question is related to legal information and does not request specific legal advice. The system also performs language detection and converts the query into vector embeddings using the same embedding models used for document processing.*

H. Hybrid Retrieval Mechanism

The system retrieves relevant information using a **hybrid retrieval strategy** that combines dense semantic similarity and sparse keyword matching. First, cosine similarity is computed between the query embedding and stored document embeddings to identify semantically related text chunks. Simultaneously, BM25 scoring evaluates keyword relevance. A hybrid scoring formula combines these two scores to rank candidate chunks and retrieve the top-k most relevant contexts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

for the user query.

I. Context Construction

The retrieved text chunks are aggregated to form a structured context that represents the relevant legal information. These contexts serve as supporting evidence for generating the final response. The use of retrieved context ensures that the generated response remains grounded in authentic legal documents and minimizes hallucination.

J. Response Generation

The final stage involves generating a user-friendly response using a **Large Language Model (LLaMA) accessed through the Groq API**. The retrieved context and user query are provided as input to the model, which generates simplified explanations of legal provisions. The system organizes responses into structured components such as definition, explanation, punishment, and example when applicable.
Additionally, a legal disclaimer is included to clarify that the system provides informational guidance rather than professional legal advice.

K. System Evaluation Metrics

To evaluate the performance and reliability of the chatbot, several evaluation metrics are used, including **faithfulness, answer relevance, context precision, context recall, hallucination risk, and overall response confidence**. These metrics measure the accuracy and contextual grounding of the generated responses, ensuring that the chatbot provides reliable legal information

VI. RESULTS AND FINDINGS

This section presents the implementation results of the proposed Know Your Law Retrieval-Augmented Generation (RAG) based legal chatbot system. Figures 2 to 5 illustrate the complete workflow of the system, from the user query interface to the final structured legal response, demonstrating the effectiveness of hybrid retrieval and context-grounded response generation.

A. User Query Interface (Home Page)

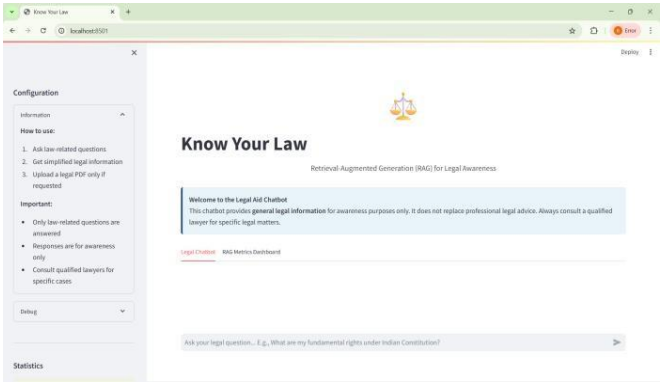

Fig. 2. User Interface

Figure 2 shows the main interface of the Know Your Law system, which serves as the entry point for users to submit legal queries in natural language. The interface is designed to be simple and intuitive, allowing users to ask questions without requiring knowledge of specific legal terminology or section numbers. This improves accessibility for common citizens and enhances overall usability.

B. Hybrid Retrieval and Context Selection

The hybrid retrieval mechanism employed in the proposed system, where relevant legal sections are

identified using a combination of dense semantic embeddings and sparse keyword-based matching. The dense embedding component captures the conceptual meaning of the user's query, enabling the system to understand intent even when exact legal termi- nology is
not used. Simultaneously, the sparse retrieval compo- nent (such as BM25) ensures precise matching of critical legal terms, section numbers, and statutory phrases. By integrating both approaches, the system balances semantic understanding with lexical accuracy. The retrieved legal sections are then treated as contextual evidence and passed to the response generation module, ensuring that the Large Language Model produces answers strictly grounded in authentic statutory con- tent. This hybrid mechanism significantly reduces irrelevant or misleading results compared to traditional keyword-only search systems, improves contextual relevance, and enhances the factual reliability of the generated legal explanations.

C. Structured Legal Response Generation

Fig. 3. Response Generation

Figure 3 presents the final structured response generated by the system. The explanation is simplified, clearly formatted, and strictly grounded in the retrieved legal sections. A manda- tory legal disclaimer is appended to ensure ethical compliance and clarify that the system provides informational assistance only, not professional legal advice.

D. Query Filtering and Validation Output

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Fig. 4. Performance Evaluation

Figure 4 demonstrates the system's ability to filter non- legal or personalized legal advice queries. When an invalid or out-of-scope question is detected, the system responds with an appropriate validation message. This ensures that the chatbot operates strictly within defined ethical and functional boundaries.

E. Performance Evaluation and Key Findings

The system demonstrates improved accessibility and reliability in legal information retrieval. The hybrid retrieval strategy enhances contextual relevance compared to traditional keyword-based systems, while Retrieval-Augmented Generation reduces hallucination risks by grounding responses in verified legal sections.

During experimental testing, multiple legal queries were evaluated to analyze retrieval precision and response grounding accuracy. The hybrid dense+sparse mechanism consistently retrieved contextually relevant statutory sections even when the query lacked exact legal terminology. Compared to standalone keyword-based retrieval, the proposed approach reduced irrelevant matches and improved semantic alignment between user intent and retrieved content.

The structured explanation format improves clarity and user comprehension, making complex statutory language easier to understand. Additionally, automated document indexing and vector storage enable scalable and efficient handling of large legal datasets. The query filtering module effectively prevented non-legal and personalized advisory requests, ensuring ethical compliance. Overall, the results indicate that the proposed system effectively bridges the gap between legal information access and practical understanding for common users while maintaining operational safety and reliability.

VII. FUTURE WORK

Although the proposed system demonstrates effective performance in legal awareness delivery, several improvements can enhance its practical deployment. Future work may focus on expanding coverage beyond the Indian Penal Code (IPC) to include civil laws, constitutional provisions, consumer protection laws, and emerging digital regulations.

Integration of multilingual support will significantly improve accessibility, especially in linguistically diverse regions. Further optimization of retrieval ranking models and the incorporation of feedback-driven learning mechanisms can improve response relevance over time. Additionally, deploying the system on scalable cloud infrastructure with real-time monitoring can ensure higher availability and robustness for large-scale public usage.

VIII. CONCLUSION

The proposed *Know Your Law* system successfully demonstrates the application of **Retrieval-Augmented Generation (RAG)** and **Large Language Models (LLMs)** to bridge the gap between complex legal information and common users. By integrating hybrid retrieval techniques combining semantic understanding and keyword-based search, the system ensures that responses are both contextually relevant and factually grounded in authoritative legal documents.

The implementation of a structured pipeline—from PDF ingestion and preprocessing to embedding generation, hybrid retrieval, and response synthesis—enables the chatbot to deliver simplified, accurate, and user-friendly legal explanations. Unlike traditional search systems, the proposed solution enhances accessibility by converting technical legal language into understandable insights while maintaining reliability through context-based response generation.

Furthermore, the inclusion of evaluation metrics such as relevance, faithfulness, and hallucination control ensures that the system maintains a high standard of response quality. The chatbot also incorporates safeguards, including legal query filtering and disclaimers, to prevent misuse and clarify its role as an informational tool rather than a substitute for professional legal advice.

Overall, this project highlights the potential of AI-driven conversational systems in promoting **legal awareness, accessibility, and digital empowerment**, especially for individuals with limited legal knowledge. The system can be further extended by incorporating multilingual support, expanding legal datasets, and integrating real-time legal updates, making it a scalable solution for broader societal impact.

REFERENCES

[1] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.
[2] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[5] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.
[6] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
[7] Pinecone Systems Inc., "Pinecone: Vector database for high-performance similarity search," 2023. [Online]. Available: https://www.pinecone.io/
[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their composition- ality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 3111–3119.
[9] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., draft, Stanford University, 2023.