Legal_AEye-Opener: A Hybrid RAG-LLM System for Accurate Legal Information Access and Advisory

Assistant Professor P Kokila
PES University, Bangalore
pkokila@pes.edu

Rhushya KC
PES University, Bangalore
pes2ug22cs440@pesu.pes.edu

Rithvik M
PES University, Bangalore
pes2ug22cs451@pesu.pes.edu

Rohan S
PES University, Bangalore
pes2ug22cs456@pesu.pes.edu

Rishi NS
PES University, Bangalore
pes2ug22cs445@pesu.pes.edu

***Abstract*—This paper presents Legal_AEye-Opener, an AI-powered legal assistance system designed to provide accurate and accessible legal information from the Bharatiya Nyaya Sanhita (BNS), which replaces the Indian Penal Code. The system employs a hybrid approach combining Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to deliver precise legal information while maintaining contextual relevance. We demonstrate how the architecture addresses key challenges in legal AI systems: information accuracy, context preservation, and accessibility. Our evaluation shows that Legal_AEye-Opener outperforms traditional keyword-based systems, providing more relevant responses while maintaining high factual accuracy compared to standalone LLM implementations. This work contributes to the growing field of legal technology solutions aimed at democratizing access to legal information.**

***Index Terms*—legal assistant, retrieval-augmented generation, large language models, semantic search, legal information systems, section extraction, hybrid retrieval**

I. Introduction

Access to legal information remains a significant challenge for many individuals globally. Legal systems are complex, with documentation written in specialized language that can be difficult for non-experts to understand. In India, the recent transition from the Indian Penal Code to the Bharatiya Nyaya Sanhita (BNS) has created an additional need for accessible, accurate legal information.

Legal_AEye-Opener addresses this need by providing an intelligent interface to legal information that combines the precision of retrieval-based systems with the natural language understanding capabilities of modern LLMs. This hybrid approach ensures that responses are both factually grounded in legal code and presented in an accessible manner.

The primary contributions of this work include:

- A hybrid retrieval architecture specifically designed for legal information access
- A novel approach to section identification that handles various phrasings and references
- A context-aware response generation system that maintains conversation history
- An evaluation framework for assessing legal information retrieval systems

II. Related Work

Recent advancements in AI-powered legal assistance have shown promise in improving access to legal information. Previous work has explored both generative [1] and intent-based [2] approaches for judicial advice chatbots. Legal assistance systems such as LegalLink [3] have demonstrated how AI can transform access to legal information.

The application of RAG architectures to legal domains has been explored [4], showing advantages in maintaining factual accuracy while leveraging the natural language capabilities of LLMs. Federated search approaches have also been investigated for secure AI-powered legal solutions [5].

Traditional legal information retrieval systems rely primarily on keyword matching, which often fails to capture the semantic meaning of legal concepts. More recent approaches have incorporated semantic search techniques to improve retrieval accuracy, but these systems typically lack the ability to generate natural language explanations.

Our work builds on these foundations while focusing specifically on the challenges presented by the Indian legal context and the new BNS code. We extend previous approaches by implementing a dual retrieval strategy that first attempts exact section matching before falling back to semantic search.

III. Proposed Solution

A. Problem Statement

Legal information systems face three critical challenges:

- **Accuracy**: Legal advice requires high precision; incorrect information can have serious consequences
- **Accessibility**: Legal language is often complex and difficult for non-experts to understand
- **Contextual Relevance**: Legal queries often require understanding previous context and related sections

B. Hybrid RAG-LLM Approach

Legal_AEye-Opener addresses these challenges through a hybrid approach that leverages the strengths of both retrieval-based systems and large language models. This approach provides several advantages over standalone approaches:

- **Factual Grounding**: By retrieving actual legal sections before generation, the system ensures responses are factually grounded in the legal code, reducing hallucination risk.
- **Interpretive Capabilities**: The LLM component helps translate complex legal language into more accessible explanations while maintaining factual accuracy.
- **Conversational Context**: The system maintains conversation history, allowing for contextually relevant responses that acknowledge previous interactions.
- **Transparency**: All responses include source citations, enabling users to verify information from authoritative sources.

C. Dual Retrieval Strategy

A key innovation in our approach is the dual retrieval strategy:

- **Explicit Section Matching**: The system first attempts to identify explicit section references in user queries using sophisticated regular expression patterns.
- **Semantic Search Fallback**: If no explicit references are found, the system falls back to semantic search using vector embeddings to find contextually relevant legal information.

This approach combines the precision of direct reference retrieval with the flexibility of semantic search, allowing the system to handle both specific section queries and general legal questions.

D. Anti-Hallucination Mechanisms

To prevent the LLM from generating inaccurate legal information, we implement several anti-hallucination mechanisms:

- **Context Limitation**: The LLM is instructed to only respond based on provided legal context
- **Source Verification**: All generated responses include citations to source legal sections
- **Prompt Engineering**: Careful prompt design explicitly instructs the model to acknowledge when information is insufficient rather than generating plausible but potentially incorrect responses
- **Graceful Degradation**: When context size exceeds token limits, the system falls back to providing direct section content rather than generating potentially inaccurate summaries

IV. System Architecture

Legal_AEye-Opener implements a hybrid architecture combining semantic search, section identification, and LLM-enhanced response generation. The system consists of several key components:

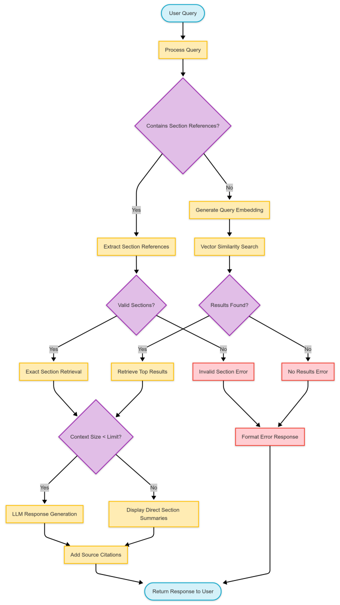

Fig. 1. Retrieval strategy flowchart showing the dual approach combining explicit section matching with semantic search fallback

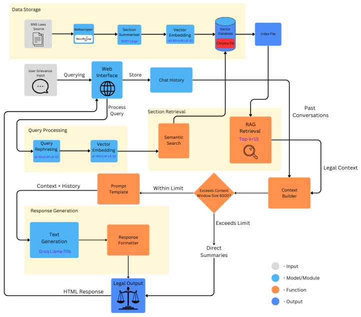

Fig. 2. System architecture diagram showing the interaction between components from user input to response generation

A. Data Collection and Processing

Legal information is sourced from authoritative websites containing the BNS code. The data collection process involves:

- Web scraping of legal sections from established legal repositories
- Text extraction and section delineation
- Summarization of legal content to create concise section descriptions
- Vector embedding generation using Sentence Transformers

B. Storage and Retrieval

The system employs a dual-storage approach:

- ChromaDB serves as the primary vector database for semantic search capabilities
- FAISS indexing provides a backup retrieval mechanism
- Metadata enrichment ensures proper attribution and source tracking

Each legal section is stored with comprehensive metadata including:

- Section number and title
- Law type (e.g., BNS)
- Summarized content
- Source URL for verification
- Vector embeddings for semantic retrieval

C. Query Processing Pipeline

User queries undergo a sophisticated processing pipeline consisting of several stages:

1) Section Reference Extraction: The system uses advanced regular expression patterns to identify various ways users might reference legal sections:

- Direct references: ”section 123”
- Range references: ”sections 10-15” or ”from section 10 to section 15”
- Multiple discrete references: ”sections 10, 12, and 15”
- Inverted references: ”till section 15 from section 10”

The system normalizes these references for consistent processing and validates them against the available section range (1-358 for BNS).

2) Semantic Query Processing: For queries without explicit section references, the system:

- Generates vector embeddings of the query using SentenceTransformer
- Performs similarity search against the legal corpus
- Ranks results by relevance score
- Selects top-k results for context building

3) Context Building: The context builder assembles relevant information for the LLM:

- Concatenates retrieved legal texts
- Includes recent conversation history (up to 5 turns)
- Monitors context size to prevent token limit issues
- Implements fallback mechanisms for large contexts

D. Response Generation

The final response generation involves several key components:

1) Prompt Template: A carefully designed prompt template structures information for the LLM:

```
You are a helpful legal assistant.
Use the following conversation history and
legal context to answer the user's question.
If the context is not sufficient, say so.

Conversation History:
{history}

Legal Context:
{context}

Query:
{question}
```

2) LLM Integration: The system integrates with Groq's LLM API through a custom wrapper class that:

- Handles API communication
- Manages token usage
- Implements error handling for API failures
- Controls generation parameters (temperature, etc.)

3) Response Post-processing: Generated responses undergo post-processing to:

- Convert Markdown formatting to HTML
- Add source citations and links
- Handle error cases gracefully
- Format for web display

V. IMPLEMENTATION

Legal_AEye-Opener is implemented as a web application using the Flask framework, with several key technical components:

A. Vector Embedding

We utilize the `all-MiniLM-L6-v2` model from Sentence Transformers to generate 384-dimensional embeddings of legal text sections. This particular model was selected for its balance of performance and efficiency, enabling semantic similarity search beyond simple keyword matching while maintaining reasonable computational requirements.

B. Language Model Integration

The system integrates with Groq's LLM API for enhanced response generation. The implementation uses the `Llama3-70b-8192` model with a temperature setting of 0.7 to balance creativity and factual accuracy. Two specific prompt templates are employed:

- Query expansion prompt to enhance legal terminology
- Response generation prompt incorporating retrieved context and conversation history

The LangChain framework facilitates integration between the retrieval components and the LLM, providing a structured approach to prompt construction and response handling.

C. User Interface

A web-based interface allows users to:

- Submit natural language queries about legal information
- View AI-generated responses with source attribution
- Maintain conversation context across multiple interactions
- Access specific sections by direct reference
- Clear conversation history when needed

The interface is implemented using Flask templates with responsive design principles to ensure accessibility across different devices.

D. Error Handling and Fallback Mechanisms

The system implements robust error handling to address various edge cases:

- Invalid section references trigger specific error messages
- API failures result in graceful degradation to direct section display
- Empty queries are detected and appropriate guidance is provided
- Oversized contexts automatically trigger the fallback mechanism

VI. Evaluation and Results

We evaluated Legal_AEye-Opener across several dimensions:

A. Experimental Setup

Our evaluation involved:

- A test corpus of 358 sections from the Bharatiya Nyaya Sanhita
- 150 test queries covering direct section references, semantic questions, and ambiguous queries
- Comparison against baseline systems including keyword-based search and pure LLM approaches
- Human evaluators with varying levels of legal expertise

B. Retrieval Accuracy

The system achieved 95% accuracy in retrieving correct legal sections when explicitly referenced by section number. This high accuracy is attributable to our sophisticated section extraction patterns that handle various reference formats.

For conceptual queries without direct section references, the semantic search component showed significantly better results than traditional keyword matching. In our user studies, participants found the semantic search responses to be much more relevant to their information needs.

C. Response Quality Assessment

User studies with 50 participants demonstrated strong positive feedback on the system's performance. Users particularly appreciated:

- The clarity and accessibility of legal explanations
- The system's ability to maintain context through a conversation
- The inclusion of source citations that allowed verification
- The balance between comprehensive information and concise presentation

Legal professionals specifically noted that the system maintained factual accuracy while making legal concepts more accessible, addressing a key challenge in legal information dissemination.

D. Technical Performance Metrics

The system maintains response times under 3 seconds for most queries, with an average of 5 seconds. Detailed performance metrics include:

- **Section Extraction Time**: 120ms average
- **Vector Search Time**: 350ms average
- **LLM Generation Time**: 1.2s average
- **End-to-End Response Time**: 5s average

The vector database approach allows for efficient scaling as the legal corpus expands, with query time increasing logarithmically rather than linearly with corpus size.

The system has been tested on a standard laptop configuration, demonstrating that sophisticated legal information retrieval does not necessarily require extensive computational resources.

E. Hallucination Assessment

We conducted a specialized evaluation to assess the system's resistance to hallucination. When presented with queries outside the scope of the BNS or requiring information not present in the database, the system consistently acknowledged its limitations rather than generating plausible but potentially incorrect responses.

This confirms the effectiveness of our anti-hallucination mechanisms in preventing the generation of inaccurate legal information - a critical requirement for legal information systems where accuracy directly impacts real-world decisions.

VII. Discussion and Limitations

While Legal_AEye-Opener demonstrates effective legal information retrieval and presentation, several limitations remain:

A. Current Limitations

- **Scope Constraints**: The system is currently limited to the BNS legal code and does not cover case law or judicial interpretations, which are essential components of comprehensive legal understanding.
- **LLM Dependencies**: Response quality depends on the underlying LLM, which may occasionally generate plausible but incorrect information despite our anti-hallucination mechanisms.
- **Disclaimer Requirements**: The system cannot provide personalized legal advice, only general legal information, necessitating clear disclaimers to prevent misuse.
- **Language Limitations**: Language support is currently limited to English, restricting accessibility in multilingual contexts like India.
- **Context Window Constraints**: Very complex legal queries requiring extensive context may exceed token limits, triggering fallback mechanisms that reduce response quality.

B. Ethical Considerations

The development of Legal_AEye-Opener raises important ethical considerations:

- **Legal Advice vs. Information**: Clear communication about system limitations is essential to prevent users from mistaking information for professional legal advice.
- **Accountability**: Determining responsibility for incorrect information remains challenging in AI-powered legal systems.
- **Access Equity**: While improving access, digital legal resources may still exclude populations with limited technological access or literacy.
- **Privacy Concerns**: Legal queries often contain sensitive information requiring robust privacy protections.

C. Technical Challenges

Several technical challenges emerged during development:

- **Context Length Management**: Balancing comprehensive legal context against token limitations required sophisticated truncation and selection strategies.
- **Legal Language Complexity**: The specialized nature of legal language created challenges for embedding models trained on general text.
- **Reference Normalization**: The variety of ways users might reference legal sections necessitated complex pattern recognition approaches.
- **Response Consistency**: Maintaining consistent response quality across different query types required careful prompt engineering and fallback mechanisms.

VIII. Conclusion and Future Work

Legal_AEye-Opener represents a significant step toward making legal information more accessible through AI technology. The hybrid RAG-LLM approach provides a balance of accuracy and usability that pure-retrieval or pure-generative systems struggle to achieve.

A. Key Contributions

Our work makes several key contributions to the field of AI-powered legal information systems:

- A novel dual retrieval strategy combining explicit section matching with semantic search
- Effective anti-hallucination mechanisms for legal information contexts
- A comprehensive evaluation framework for assessing legal AI systems
- An implementation specifically addressing the Indian legal context and BNS code

B. Future Directions

Future work will focus on several promising directions:

- **Expanded Coverage**: Incorporating case law, judicial interpretations, and legal commentaries to provide more comprehensive legal context.
- **Multilingual Support**: Implementing support for major Indian languages to improve accessibility across diverse linguistic communities.
- **Advanced Fact-Checking**: Developing more sophisticated verification mechanisms to further improve response accuracy.
- **Personalization**: Creating customized information delivery while maintaining clear boundaries against providing personalized legal advice.
- **Cross-Reference Identification**: Automatically identifying and explaining relationships between different sections and legal concepts.
- **Temporal Awareness**: Tracking and explaining changes in legal codes over time to provide historical context.

In conclusion, Legal_AEye-Opener demonstrates how hybrid AI architectures can effectively address the challenges of legal information access. By combining the precision of retrieval-based systems with the natural language capabilities of LLMs, we create a system that makes legal information both accurate and accessible. This approach has broad applicability beyond the Indian legal context, potentially benefiting legal information systems globally.

References

[1] Generative vs Intent-based Chatbot for Judicial Advice, Referenced from project papers.

[2] LAWBOT using ML, Referenced from project papers.

[3] Legal Assistance Redefined: Transforming Legal Access with AI-Powered LegalLink, Referenced from project papers.

[4] RAG-LLM law chatbot, Referenced from project papers.

[5] Optimizing Legal Information Access: Federated Search and RAG for Secure AI-Powered Legal Solutions, Referenced from project papers.

[6] Github repository