AI - Powered Legal Documentation Assistant

Vaishnavi C [1],
Department of Computer Science and Engineering,
Presidency University,
Bengaluru, India

Shruthi V [2],
Department of Computer Science and Engineering,
Presidency University,
Bengaluru, India

Ruthika S Shetty [3],
Department of Computer Science and Engineering,
Presidency University,
Bengaluru, India

Sreelatha PK [4]
Assistant Professor,
Department of Computer Science and Engineering,
Presidency University,
Bengaluru, India

Abstract: Legal documentation is a complex process that requires expert knowledge that makes it impossible for public/small businesses to access. The focus of this paper is on developing AI – Powered Legal Documentation Assistant, which simplifies the process of legal documentation. The assistant will be able to simplify legal terms and provide the text that can be understood by layman. The system will utilize NLP (Natural Language Processing) and machine learning (ML) algorithms to at minimum mistakes and confusion, extract qualitatively accurate documents from the system, which are legally valid. The proposed model seeks to address these inefficiencies by making existing legal services relatively inexpensive and improving the accuracy in their documentation. Users will be able to design documents according to their specifications enabling the solution to serve as a bridge by connecting as well as interfacing with legal data bases to check if the proposed documentation is within the parameters of local laws. The aim of this paper is to present the problem of the statement, the whole technology stack, predicted the outcome and what impact the system could have. Once the AI capable assistant is trained fully it would mostly serve small scale businesses and individuals at large in India where the need for legal documents is high, thus giving these people easier means to obtain such documents, consequently inciting more legal literacy and self-empowerment. Future enhancements may include expanding the range of supported documents and integrating expert legal consultations for complex cases.

Keywords—AI-powered legal assistant, legal documentation automation, Natural Language Processing (NLP), Machine Learning (ML), legal accessibility, document generation, legal compliance, small business legal support.

I. INTRODUCTION

Contracts, agreements, affidavits, and other legally binding documents are just a few examples of the legal documentation that is a crucial component of many business and personal transactions. Although creating these documents could be difficult and time consuming, and might require legal knowledge. In India, accessing legal services is difficult for small businesses due to the high costs and inexperience. These difficulties may cause legal problems and mistakes in understanding the legal context. These issues can be resolved by using Artificial Intelligence (AI), Natural Language Processing (NLP). Our paper focuses on reducing the dependency on legal associates or experts and producing precise legal texts in layman language.

The users will be able to enter their legal queries and our chatbot will processes the query and creates personalised legal documents that stick to the legal requirements and at the same time understood by common people. This paper aims in closing the gap between those who need legal services and those who can afford them, by utilising AI for legal documentation, which will make legal processes more accessible, economical, and efficient.

II. LITERATURE REVIEW

The study by Rithik Raj Pandey et al [1] uses Optical Character Recognition (OCR) along with a custom-trained GPT model to process and simplify legal data. Techniques like Natural Language Processing (NLP) and pattern recognition are used to improve the readability of the documents. The system involves a chatbot that allows users to get simplified legal documents or consult legal experts via virtual meetings. OCR is used to simplify the legal text into simple language. The chatbot allows the users to upload their legal documents to get assistance. Users will be able to consult legal experts through the platform. The system is integrated using publicly available legal databases to keep the data updated.

The study by Imogen Vimala et al [2] uses Natural Language Processing (NLP) and other Machine Learning (ML) techniques for drafting the contracts, text summarization etc. The system features chatbots, semantic analysis and document automation. The tech stack used includes HTML, CSS, JavaScript, PHP, MySQL and Collect.Chat. The system aims on improving the accessibility and assisting the users. The idea is to highlight the importance of technology advancements in the legal industry, which prioritizes user interaction and engagement. The system uses publicly available legal documents, research databases for training the model.

The study by Awez Shaikh et al [3] uses Machine Learning, Natural Language Processing (NLP) and Large Language Models (LLMs) for legal query handling, text summarization and legal document drafting. The system uses Optical Character Recognition (OCR) for extracting the text from PDFs and uses vector databases for storing the documents. The tech stack used includes a web application, but the exact implementation details were not provided in the paper. By implementing NLP and ML techniques, the platform aims to improve the accessibility to legal resources and provide the users with legal information more efficiently. The system integrates chatbot, which helps the users with their legal queries, and option to consult legal associates. The users would be able to handle legal matters confidently. The dataset has been derived from publicly available legal documents and legal databases, which makes sure that updated and relevant content is being used for training the model.

The study by G. Kiran Kumar et al [4] uses Optical Character Recognition (OCR) and Natural Language Processing (NLP) to simplify the legal documents and generate documents understood by layman. The system features text summarization, legal document drafting and text simplification. The process includes repeatedly refining an AI model, usability testing and getting the user feedback, to improve the accuracy and increase the accessibility for small-scale businesses and individuals. The difficulties faced by the layman in handling the legal documents has been taken into consideration while developing the system.

The dataset used is from publicly available legal documents, cases, contracts and other legal databases, which makes sure that updated and relevant content is used for training the model.

The study by Lalita Panika et al [5] uses LangChain, Next.js, MongoDB, Prisma and Pinecone to develop an AI – powered legal documentation assistant. The system uses Natural Language Processing (NLP) for simplification and generation of legal documents, and Pinecone, a vector storage, for retrieving the documents efficiently. The system also contains a chatbot, which is developed using the functionalities of OpenAI's GPT models to enable conversation management. Swagger UI React is used for API documentation and Kinde Auth is used for authentication. SimpliLegal aims in making it possible for the common people without any legal knowledge to access the legal information. The dataset used to train the model includes case laws, statutes, and other legal databases.

The study by Sayash Kapoor et al [6] explores the role of AI in legal tasks such as information processing, or tasks that require creativity. The paper highlights the issues faced with the dataset like inaccurate data, irrelevant data or incomplete data. These issues lead to overlapping of the training data with the test data, which in turn affects the performance of the model. The AI model is trained for tasks like summarization, prediction and retrieval, using GPT-4 and other predictive models like COMPAS. The major issue faced by Sayash Kapoor et al is the lack of clean data, and this affects the evaluation of the AI model. The dataset used was derived from publicly accessible legal texts, case laws, archives, and other open-access legal databases.

The study by Drashti Shah et al [7] explores the application of Artificial Intelligence (AI) and Machine Learning (ML) in the legal industry. The paper has used Retrieval Augmented Generation (RAG) models, Natural Language Processing (NLP) techniques like BERT and GPT, and Optical Character Recognition (OCR) to extract and generate legal text. The chatbot allows the users to upload their legal documents and get any assistance or guidance required. The idea is to make legal advice more accessible. The study has faced issues in handling different document formats and understanding the legal context. The main goal of the paper is to make legal advice more accessible through the AI-powered chatbot, which connects the users with legal associates. The system also has its limitations like privacy concerns, OCR accuracy dependency, misunderstanding of legal language, and being unable to adapt to different legal systems of different countries. The study states the importance of having better document handling techniques and semantic interpretation to get results that are more accurate from these AI systems. The dataset consists of legal documents like case laws, contracts, lease agreements, loan agreements and other judicial records. However, the original source of dataset was not disclosed. The data used was semi-structure and unstructured, which requires extraction and processing to handle different formats of files like PDFs, Word, images etc.

The study by Jhanvi Aroraa et al [8] uses Information Retrieval and Natural Language Processing (NLP) to explore the field of AI-driven legal research. The relevant content from the legal documents and statutes is extracted using techniques like BM25, Top2Vec embeddings, Law2Vec embeddings, and BERT. The system classifies legal texts into rhetorical roles. The research has faced issues in processing lengthy legal texts, having limited context awareness, and imbalance in data for classification tasks. The primary outcome of the paper is an AI-powered legal assistant, which improves the efficiency of legal document retrieval. This legal assistant system ranks among the top 10 submissions at FIRE 2020. There were several issues faced like BM25's lack of deep contextual understanding, high costs of BERT, and inefficiencies in cosine similarities. While using topic-modelling techniques, the accuracy would be impacted due to the loss of case-specific information. The paper suggests using advanced abstraction techniques and hyper parameter tuning for better precision. The dataset used consists of 3,260 case sheets and 197 statutes, retrieved from Forum for Information Retrieval Evaluation (FIRE) 2020.

Jinqi Lai et al. [9] researched into the adoption of Large Language Models (LLMs) in the legal domain is found in this study by Jinqi Lai and others. It describes how artificial intelligence (AI) can facilitate judges, provide automated legal document generation, or improve productivity in legal research. The manuscript emphasizes on the fact that although legal LLMs may see training on court rulings, statutes, and case records, of the serious issue even now concerning access to data for privacy. The algorithms like BERT, GPT, and the specific legal models such as ChatLaw and LawGPT are useful approaches towards the text processing and decision making. The study raised some research gaps, namely biased AI results, inconsistent datasets, and non-interpretability of court rulings. Ethical dilemmas arise when human rights are potentially jeopardized as AI is used to influence a judge's ruling in court and make police predictions. This is one of the major disadvantages that legal LLMs would have: taking in the Edata found in past legal data, which could eventually lead to unfair decisions. This also warns because a judge's right of judgement may be restricted when judiciary independence is compromised by excessive usage of AI. In addition, these models can hardly be verified for true performance due to absence of possible benchmarking and real-world testing. The paper gives possibilities of improving legal AI while stressing on the need to make responsible utilization through open, fair, and better governance of datasets.

The study by Nguyen Ha Thanh [10] on LawGPT 1-0, an AI-capable legal assistant powered by and optimized using GPT-3 for the legal domain. Without any need for manual input, LawGPT 1.0 produces legal documents or answers to any legal question, and it provides legal advice by using the transformer architecture with attention mechanisms and fine-tuning techniques. However, even with all the good points it has, this paper mentions some of its disadvantages, among which is the inability to explain itself, raising doubts about the credibility and accountability of AI's legal judgments. In addition, since the model lacks Reinforcement Learning from Human Feedback (RLHF), its ability to improve responses based on interaction with the user is limited. , ethical and legal issues remain regarding privacy, accountability, and bias in AI-generated legal advice. It also has another important limitation since the current version of LawGPT 1.0 only understands and uses English, which renders it unusable in multilingual legal systems. The study advises future improvements to include support of more languages and provide better explainability features, which, however, have not yet been implemented. Its practical reliability becomes more questionable due to lack of discussion about real-world deployment and the opaqueness of the sources of its datasets. However, notwithstanding these drawbacks, LawGPT 1.0 can still use in a future of increased access to legal services through its provision of AI-powered legal aid services 24 hours a day.

III. METHODOLOGY

A. Overview

The whole development cycle of AI-driven Legal Documentation Assistant employs structured processes that ensure accuracy, reliability, and efficiency of operation of the system under given constraints. It is built upon AI models programmed in Python for the purpose of analysing and retrieving legal documents. In the methodology, five major stages are identified: data gathering, pre-processing, model development, document retrieval, and validation. With the exception of data gathering, each of the above is critical for bestowing upon the AI system the ability to cope with complex legal texts, extract valuable legal nomenclature, and deliver relevant legal documents as responses to user inquiries.

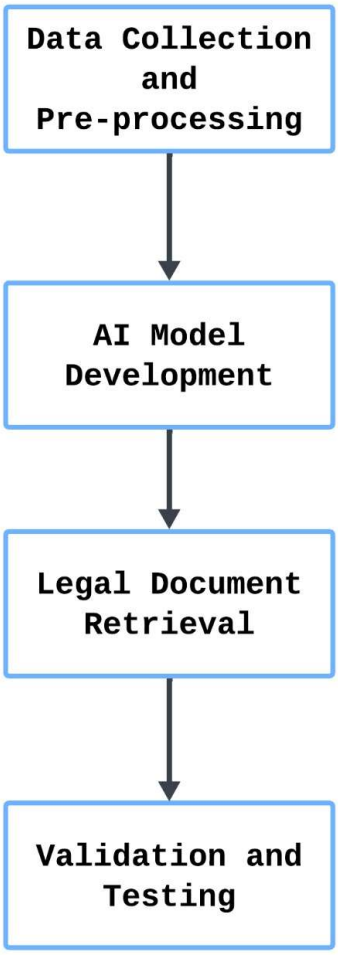

Fig. 1 - Phases of AI-Powered Legal Document Assistant

Fig. 1 shows the working process of the legal assistant. The first step is the Data Collection and Pre-processing, where the relevant data is collected from the given data and this collected data is processed for training the model. The next step is the AI model development, which involves training the model based on our requirement. The next step is the Legal Document Retrieval, where similarity search is performed to find the best answer for the legal query entered by the user. The last step is the Validation and Testing stage, where the system undergoes evaluation of its performance in real-world scenarios.

B. Data Collection and Preprocessing

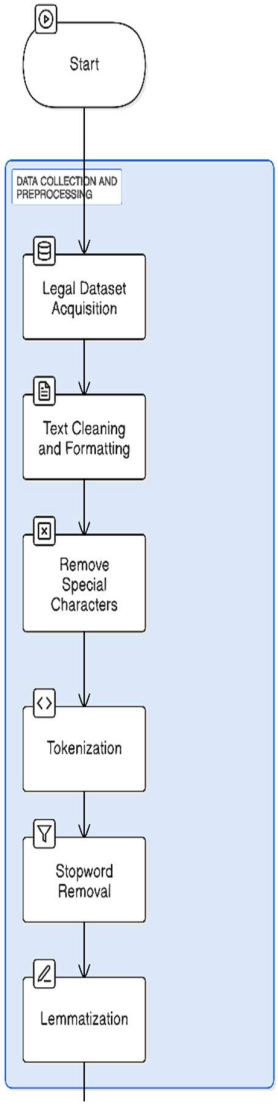

Fig.2 - Phases of Data Collection and Pre-processing

Fig.2 shows the phases of a legally automated document retrieval system: "Data Collection and Pre-processing". Legal datasets acquisition, cleaning, or formatting, special character removal, tokenization, stopword removal, and lemmatization are some processes that prepare data for their effective retrieval and analysis.

a. Legal Dataset Acquisition

It takes much more data to be considered adequate to train and run a nice AI system. The information collected from different sources being publicly available includes case law collections, open-source legal documents, government legal repositories, and databases of law firms. Contracts, agreements, policies, and legal notices are just a few types of legal document data available in the prescriptive dataset. Legal experts examine the collected resource materials for data quality and relevance, removing obsolete or jurisdiction-bound materials that could limit its generalization capacity.

b. Text Cleaning and Formatting

Once raw legal data is collected and curated, it is subjected to rigorous pre-processing to bring it up to a level fit for AI training and usability.

This includes:

- **Removing Special Characters and Formatting Artifact**s: unnecessary symbols, extra spaces, and formatting errors are purged, emphasizing the text's relative simplicity.
- **Tokenization**: Decomposing legal text into sentences and words for further structured analysis.
- **Stopword Removal**: Common, yet uninformative, words (e.g. the, is, an) are omitted, leaving meaningful legal content.

C. AI Model Development

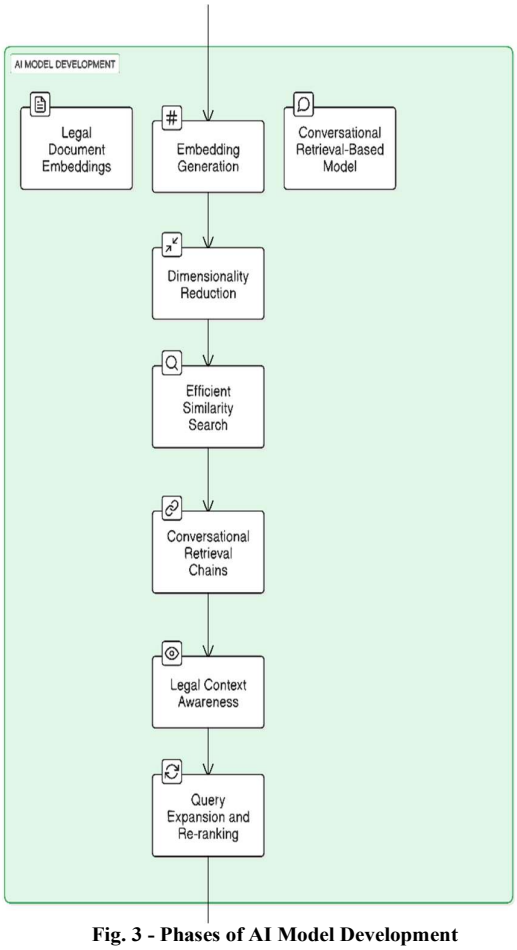

Fig. 3 - Phases of AI Model Development

Fig.3 shows the process of AI model development for legal document retrieval, which incorporates embedding generation, dimensionality reduction, similarity search, and conversational retrieval chains to efficiently, contextually retrieve and rank legal documents for enhanced user experience.

a. The System Employs Textual Embeddings For Search Document Retrieval

The legal texts are converted into numerical vector representations through the following methodology using Hugging Face transformer models:

- **Embedding Generation**: The legal texts are transformed into numerical vector representations.
- **PCA Dimensionality Reduction**: PCA, or Principal Component Analysis, facilitates reducing the length of vector keeping the important semantic substance intact.
- **Fast and Efficient Similarity Search**: The system incorporates FAISS (Facebook AI Similarity Search) and HNSW (Hierarchical Navigable Small World) graphs for fast and scalable retrieval of related legal texts.

b. Conversational Retrieval-Based Model

The system implements the following features for better retrieval of the legal information:

- **Conversational Retrieval Chains**: The system iteratively performs query refinements to ensure that the application provides documents most relevant to the case at hand.
- The train recording of legal context awareness. The model comprehends the user's query within the context of efficient, relevant retrieval of legal information.
- **Expansion and re-ranking of queries**: The queries are supplemented with pertinent legal terms, and the documents retrieved are ranked semantically.

D. Legal Document Retrieval

a. User Query Processing

This indicates that the system analyzes a user's query to retrieve the most relevant legal documents. Significant steps include:

- **Semantically embolden query**: User input would then be translated into an embedding vector aligned with the stored embeddings for legal documents.
- **Similarity matching**: The effective ways of looking for documents similar using FAISS and HNSW.
- **Ranking & Filtering**: Prioritize the most relevant legal documents and eliminate irrelevant returns.

b. Dynamic Legal Text Retrieval

The legal texts retrieved are:

- **Context-Based**: To mean the documents are reflective of the intent of the user.
- **Relevance-based ranking**: Those that are more competent represent the relevant legal information.
- **Adaptive to User Queries**: The system enhances evaluation as it interacts with users.

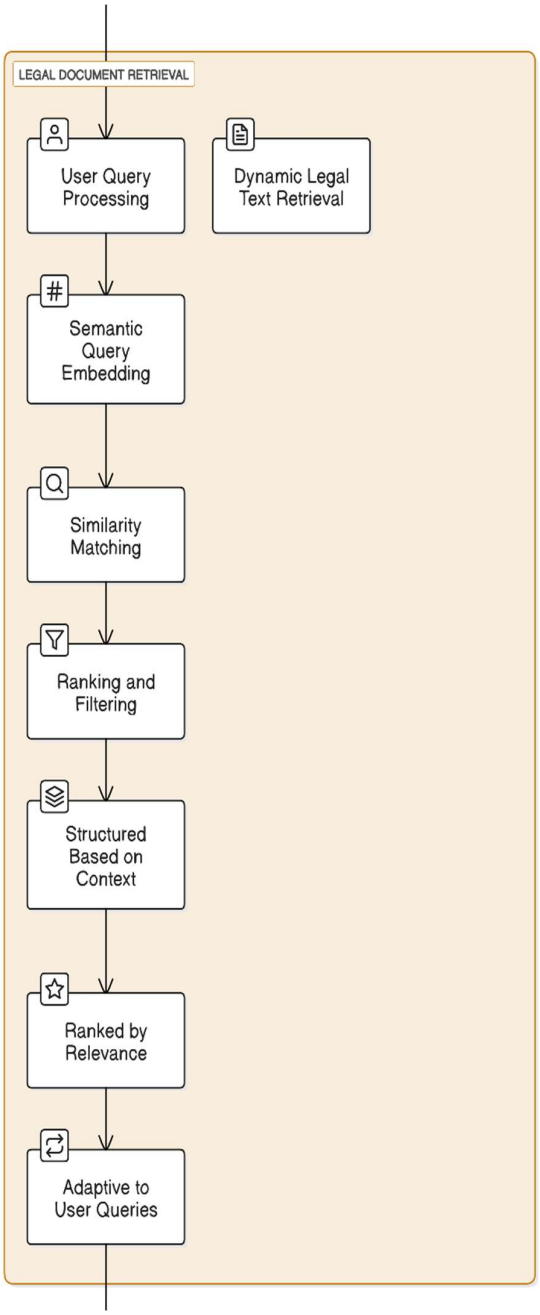

Fig. 4 - Phases of Legal Document Retrieval

The system flowchart entitled "Legal Document Retrieval" is shown in Fig. 4. The processes are described starting from user query processing, semantic embedding, matching for similarity and ranking and filtering of the results by adaptation based on user queries. The aforementioned guarantees that the retrieval of legal texts is accurate and contextually relevant.

E. Validation and Testing

a. Comparison with Existing Legal Documents

The obtained outputs are compared to ensure that they comply with pre-existing templates for legal documents.

- **Structural Consistency**: Complementing the frameworks of actual legal documents.
- **Completeness**: Verifying that all required clauses are included and formatted correctly

b. Performance Metrics

The system holds that it is more retrieval-oriented rather than generation oriented. Hence, the performance evaluation metrics include:

- **Precision and recall**: measures effectiveness in retrieval of legal documents.
- **Mean reciprocal rank**: The relevant document is being identified within the ranked results.
- **Embedding similarity scores**: measures how similar retrieved legal documents are from the input query.

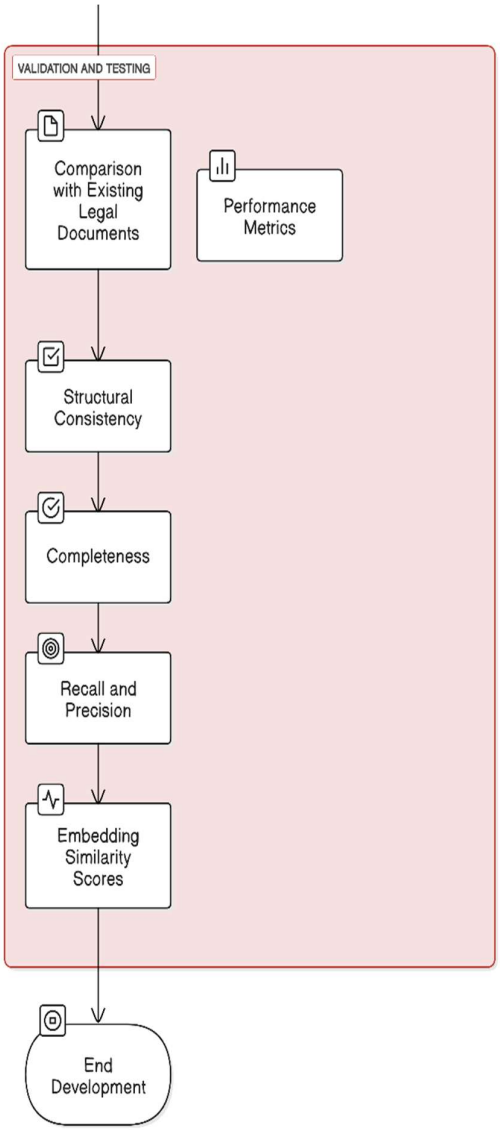

Fig. 5 - Phases of Validation and Testing

Fig.5 depicts the AI Model Development Process for Legal Document Retrieval, which incorporates embedding generation, dimensionality reduction, similarity search, and conversational retrieval chains. It ensures retrieval and ranking of legal documents in a manner that incorporates context through efficient and effective user interaction with the models.

F. Ethical Considerations and Future Enhancements

a. Data Privacy and Security

Security measures implemented to protect the information:

- **Anonymizing Sensitive Information**: There is no personal information within the actual data that could violate the privacy concerns of the users.
- **Strong Encryption Mechanism**: Protects legal data being transmitted and stored.
- **Privacy Regulation compliance**: General Data Protection Regulation (GDPR) and any other rules of data protection.

b. Future Enhancements

To expand functionality, the system aims to:

- **Support for Multiple Languages- enhancing non-English legal documents´ accessibility.**
- **Real-time legal consultation -AI-powered assistance for legal queries.**

It will prove highly effective in collecting and analyzing legal texts with great accuracy, all through the AI-based Legal Documentation Assistant. By such methodological approaches, the task of legal documentation would be simplified, and legal professionals and organizations would benefit from the text embeddings, semantic search, and conversational retrieval chain.

IV. RESULTS

Fig. 6 - Response to a query on Fundamental Law by the Assistant

Fig. 6 shows the output interface of the AI Legal Word Document Assistant. The interface is displayed to the users with a message at the top and instructs them to type "exit" to end the session. The primary section includes a query input field, where users can enter their legal queries In the image shown, the user asked about the most fundamental law in the Indian Constitution, and the chatbot responded by explaining that Fundamental Rights, enshrined in Part III of the Constitution.

The extraction of responses is performed using document retrieval systems and in this case is performed with the help of a chatbot. During the user query, the first step the system does is applies PyMuPDF (fitz) to extract and preprocess legal documents saved in PDF format. Within the headings and the associated contents structures, the system is capable of identifying the required data with ease. The next step is to parse the query and identify the relevant terms for comparison against the retrieved legal text. The user query assumptions are resolved mostly through text matching and keyword searching.

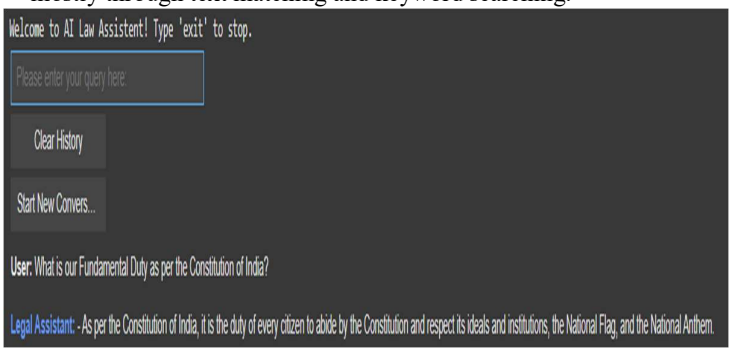

Fig. 7 - Chatbot responding to a legal query on fundamental rights

In Fig. 7, the user inquired about the fundamental duty within the Constitution of India, and received an answer from the chatbot explaining provisions of the Constitution. The user asked about the Fundamental Duty as per the Constitution of India, and the chatbot responded with an explanation based on constitutional provisions.

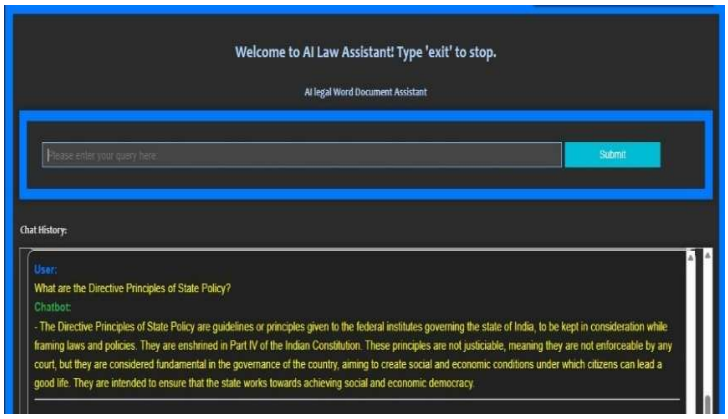

Fig. 8 - Response to a query on Directive Principles by the Assistant

In Fig. 8, the user asked about the Directive Principles of State Policy. The chatbot answered this question comprehensively citing that Part IV of the Constitution of India lays down the Directives Principles of State Policy (DPSP), which are meant to be followed for governance in India.

V. FUTURE ENHANCEMENT

This system has shown great promise throughout this evaluation period, but efficiency, scale, and user experience could still use some improvements. One key aspect that needs improvement is the application of sophisticated machine learning methodologies that optimize performance and predictive capabilities. Through artificial intelligence, efficiency can be enhanced, and human intervention can be lowered, thus automating some procedures to yield results that are more accurate. In addition, the incorporation of real-time analytics and data visualization features will enable more insightful and actionable data representations.

Another major enhancement to the platform is the improved multi-platform/device integration. Integrating seamlessly with mobile apps and cloud services will boost user engagement and accessibility. Cross platform, compatibility may be added, allowing users to access the system from different operating systems. In addition, working in the future we can really increase scalability and needs optimization performance. Our micro services-based architecture allows

the system to handle higher loads seamlessly and to provide more seamless interactions as additional users enter.

Moreover, implementing distributed computing techniques will lead to an increase in speed, reliability, and even user friendliness, which is necessary to support larger scale deployments. Soliciting user input will be important for guiding the subsequent versions of the system in addition. We intend to constantly evaluate and adjust the system through usability testing and feedback and ensure relevance of the system in the near future.

VI. CONCLUSION

AI-powered Legal Documentation Assistant will mobilize the expertise in modern NLP techniques including text embeddings, FAISS-based similarity search and other conventions retrieval to expedite the legal document processing. Accurate and contextual retrieval of legal clauses from the proprietary knowledge database in seconds.

Using vector-based similarity matching and sentence embeddings jumpstarts effective clause extraction and question answering, making the assistant a substantial improvement over traditional legal document retrieval. FAISS indexing method makes it possible to achieve real-time responses to legal inquiries and maximizes search efficiency. The output is evaluated using precision, recall, and relevance scores to ensure that the system will provide solid and trustworthy legal aid.

Moreover, legal text pre-processing techniques like named entity recognition (NER) along with lemmatisation and stop word removal improve system understanding of complex legal terminology. Security measures like data encryption and anonymization make sure that legal and ethical standards are observed. Future development attempts to improve the system's capabilities by adding multi-language interfaces, real-time validation by legal specialists, and more extensive legal field coverage. This multi lingual support allows for the easy and fast retrieval of structured law documents, which solves the problem many of the legal industry professionals, organizations, and citizens currently face.

VII. REFERENCES

[1]. Rithik Raj Pandey, Sarthak Khandelwal, Satyam Srivastava, Yash Triyar and Mrs. Muquitha Almas, "LegalSeva: AI - Powered Legal Documentation Assistant", International Research Journal of Modernization in Engineering Technology and Science, vol. 06/Issue: 03, March 2024.

[2]. Imogen Vimala, Sreenidhi J. and Nivedha V, "AI - Powered Legal Documentation Assistant", Journal of Artificial Intelligence and Capsule Networks. 6. 210-226. 10.36548/jaicn.2024.2.007.

[3]. Awez Shaikh, Rizvi Mohd Farhan, Zahid Zakir Hussain and Shaikh Azlaan, "AI - Powered Legal Documentation Assistant", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN: 2349-5162, Vol.11, Issue 4, page no. k526-k530, April-2024.

[4]. G. Kiran Kumar, A. Shreyan, G. Harini, M. Balaram, (2024), "AI - Powered Legal Documentation Assistant", International Journal of Engineering Innovations and Management Strategies 1 (1):1-13.

[5]. Lalita Panika, Aastha Gracy, Abhishek Khare, Sanket Mathur and S. Hariharan Reddy, "SimpliLegal: An AI - Powered Legal Document Assistant", International Research Journal of Modernization in Engineering Technology and Science, vol. 06/Issue: 04, April 2024.

[6]. M. E. Kauffman and M. N. Soares, "AI in legal services: New trends in AI-enabled legal services," Service Oriented Computing and Applications, vol. 14, pp. 223–226, Oct. 2020, doi: 10.1007/s11761-020-00305-x.

[7]. S. Kapoor, P. Henderson, and A. Narayanan, "Promises and pitfalls of artificial intelligence for legal applications," arXiv, Feb. 6, 2024.

[8]. L. B. Eliot, "AI and Legal Argumentation: Aligning the Autonomous Levels of AI Legal Reasoning," arXiv preprint arXiv: 2009.11180, 2020.

[9]. J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan, "Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model," arXiv preprint arXiv:2306.16092, May 2024.

[10]. Q. Steenhuis, D. Colarusso, and B. Willey, "Weaving Pathways for Justice with GPT: LLM-driven Automated Drafting of Interactive Legal Applications," arXiv preprint arXiv: 2312.09198, Dec. 2023.

[11]. D. Shah, J. Vasi, T. Gandhi, and K. Dabre, "AI & ML Based Legal Assistant," International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 07, pp. 706-708, Jul. 2024.

[12]. J. Aroraa, T. Patankara, A. Shaha, and S. Joshia, "Artificial Intelligence as Legal Research Assistant," in Forum for Information Retrieval Evaluation (FIRE), Hyderabad, India, Dec. 2020.

[13]. P. N. Devaraj, R. T. P. V, M. K. R, and A. Gangrade, "Development of a Legal Document AI-Chatbot," School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India.

[14]. J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large Language Models in Law: A Survey," arXiv preprint, arXiv: 2312.03718, Nov. 2023.

[15]. Nguyen, H. T., "A Brief Report on LawGPT 1.0: A Virtual Legal Assistant Based on GPT-3," arXiv preprint arXiv: 2302.05729v2, 2023.