**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

Automatization of tasks in the absence of complex AI systems in legal documentation is difficult because of the pre-existing knowledge schema block, which requires specialized skills from trained individuals or small businesses without infrastructure and resourced to defend themselves legally. The goal of this project is to create a Legal Documentation Assistant AI, which simplifies the process through automation and speaks in the plainest of languages that is understandable to as many users as possible without using legal terms. 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 project 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. However, creating these documents can be a difficult and time-consuming process that frequently calls for specific legal knowledge. Accessing legal services is extremely difficult for individuals and small businesses, especially in India, because of the high costs, inexperience, and complexity of legal terminology. These difficulties may result in mistakes, misunderstandings, and even legal issues.   
Automating and streamlining legal documentation is possible thanks to developments in Artificial Intelligence (AI), especially Natural Language Processing (NLP) and Machine Learning (ML). In order to lessen reliance on legal experts, this research aims to create an AI-powered legal documentation assistant that can produce precise legal documents in plain language.

By letting users enter appropriate information, the system seeks to improve legal accessibility, efficiency, and accuracy. AI then processes the data and creates personalised legal documents that adhere to legal requirements. The problem statement, technology stack, anticipated results, and impact of the suggested solution are all covered in this paper. It is anticipated that the AI-powered assistant will improve access to legal resources for both individuals and small businesses, reduce errors, and expedite the legal documentation process. The study also identifies possible obstacles to putting such a system into place, such as issues with data privacy, legal compliance, and ethical considerations.   
This project intends to close the gap between those who need legal services and those who can afford them by utilising AI for legal documentation, which will make legal procedures more accessible, economical, and efficient.

**II. LITERATURE REVIEW**

In order to process and streamline legal documents, Rithik Raj Pandey et al.'s study [1] combines optical character recognition (OCR) technology with a custom-trained GPT model. To improve document readability, the AI model makes use of Natural Language Processing (NLP) and pattern recognition techniques. With the help of the platform's chatbot, users can draft or streamline legal documents and even hold virtual consultations with legal experts. The solution simplifies legalese and makes document creation easier with OCR technology. Users can communicate with a chatbot for assistance or upload legal documents for processing through the system. Through the platform, users can have direct consultations with legal professionals, which greatly enhances the documentation process.

To maintain the generated content up-to-date and relevant, the system incorporates legal databases. The AI model was trained using publicly accessible legal data.

Imogen Vimala et al.'s study [2] uses machine learning and natural language processing (NLP) methods for document retrieval, contract draughting, and summarising legal texts. To increase efficiency, the system includes document automation, semantic analysis, and chatbots driven by AI. HTML, CSS, JavaScript (frontend), PHP (server-side), MySQL (database), and CollectChat (AI chatbot development) are all components of the tech stack. By offering real-time support and editable legal templates, the system seeks to increase accessibility. By focussing on user engagement and customisation to meet a range of legal needs, this initiative not only seeks to democratise legal access but also emphasises the significance of technological advancements in legal practices.

Legal research databases, publicly accessible legal texts, and legal document templates provide the dataset used to train the AI model.

Awez Shaikh et al.'s study [3] uses machine learning, Natural Language Processing (NLP), and Large Language Models (LLMs) for query handling, summarisation, and legal document draughting. The system incorporates a secure vector database for document storage and uses optical character recognition (OCR) to extract text from PDFs. Although specific implementation details are not given, the tech stack consists of a web-based platform with customisable templates. Utilising cutting-edge technologies like machine learning and natural language processing, the platform aims to improve access to legal information and provide users the confidence they need to handle legal issues, thereby promoting a more inclusive legal system.

To ensure accurate and effective document generation, the dataset used for model training is collected from external legal databases, publicly accessible legal documents, and legal resources.

G. Kiran Kumar et al.'s study [4] uses optical character recognition (OCR) and natural language processing (NLP) to create and streamline legal documents. The system has real-time integration with legal databases, a simplification tool, and a document draughting engine. Data security and privacy are also given top priority. To increase document accuracy and accessibility for individuals and small businesses, the methodology incorporates usability testing, user feedback integration, and iterative AI model refinement. The project tackles the challenges non-experts encounter when navigating India's complicated legal documentation. Accuracy and compliance in document generation are guaranteed by real-time integration with legal databases.

To ensure adherence to the most recent legal standards, the AI models are trained using publicly accessible legal datasets, agreements, and case laws.

LangChain, Pinecone, Next.js, Prisma, and MongoDB are used in the study by Lalita Panika et al. [5] to create an AI-powered platform for legal documentation. The system uses vector storage (Pinecone) for effective legal document retrieval and incorporates Natural Language Processing (NLP) for document generation and simplification. Conversational engagement with legal documents is made possible by chatbot functionality driven by OpenAI's GPT models. Additionally, the platform incorporates Kinde Auth for secure authentication and Swagger UI React for API documentation. SimpliLegal is a groundbreaking invention that makes it possible for more people to access rights and legal information by reducing mistakes and democratising legal services. Statutes, case laws, and legal databases provide the dataset used to train the AI models.

Three main areas have been examined in the study by Sayash Kapoor et al. [6], which looks at AI's role in legal tasks: information processing, tasks that require creativity or judgement, and predictive analytics. The study does, however, highlight serious problems with these datasets, such as biases, errors, and data contamination, which occurs when test and training data overlap and inflates performance estimates. These datasets are used to train AI models for tasks like document summarisation, case prediction, and legal information retrieval. The quality of the datasets used is still a major concern, even though generative AI systems like GPT-4 and predictive models like COMPAS have been applied to legal tasks. The study highlights that one of the biggest obstacles to properly assessing AI in legal contexts is the absence of clear, objective, and thorough datasets.

Notwithstanding these problems, the study indicates that while AI may be helpful in automating repetitive legal tasks, it will not be able to completely replace human judgement in more complicated legal situations.

The study examines datasets that are frequently used in legal AI applications, such as court rulings, case law, publicly available legal documents, and court filings. These datasets are frequently obtained from law-specific archives, court records, and open-access legal databases.

Drashti Shah et al.'s study [7] analyses the application of machine learning (ML) and artificial intelligence (AI) in legal aid, particularly for the analysis of loan and employment contracts. To extract and interpret legal information, it uses Natural Language Processing (NLP) methods like BERT and GPT, Optical Character Recognition (OCR), and Retrieval-Augmented Generation (RAG) models. The suggested system makes legal aid more accessible by enabling users to upload legal documents and communicate with an AI-powered chatbot for legal advice. Nonetheless, the study points out important gaps, such as a lack of contextual awareness, trouble managing various document formats, and difficulties with semantic inference. The primary result is a community-based platform for legal advice that offers AI-generated legal insights and links users with legal experts.

In spite of its improvements, the system still has drawbacks, including reliance on OCR accuracy, misunderstandings of legalese, and privacy issues. Additionally, it has trouble adapting to various legal systems, which restricts its worldwide applicability. For more precise, legal AI systems, the study highlights the need for enhanced semantic interpretation and better document handling strategies. Overall, the paper helps automate legal procedures, but it still needs to be improved to get past its issues.

Although the precise retrieval source is not specified, the dataset used includes legal documents such as loan agreements, employment contracts, and court case records. In order to handle various formats, such as Word files, PDFs, and scanned images, these semi-structured and unstructured documents need text extraction and processing techniques.

Jhanvi Aroraa et al.'s study [8] uses information retrieval and Natural Language Processing (NLP) techniques to investigate AI-driven legal research. It retrieves essential statutes and legal precedents using BERT-based classification, BM25, Topic Embeddings (Top2Vec), and Law2Vec embeddings. The system efficiently categorises legal texts into rhetorical roles and automates the retrieval of legal precedents. Nonetheless, the study points out important shortcomings, including poor context awareness, difficulties processing long documents, and an imbalance in data when it comes to classification tasks. The paper's primary result is an AI-based legal research assistant that ranks in the top 10 submissions at FIRE 2020 and increases the effectiveness of legal document retrieval. In spite of its improvements, the system has drawbacks, such as BM25's limited contextual awareness and BERT's high computational costs. Furthermore, retrieval accuracy may be impacted by topic modelling techniques losing case-specific information. In order to improve precision, the paper emphasises the necessity of improved abstraction techniques and hyperparameter tuning. Although the research helps automate legal research overall, more advancements are needed for increased precision and effectiveness.

The Forum for Information Retrieval Evaluation (FIRE) 2020 provided the dataset, which consists of 197 statutes and 3,260 case documents.

Jinqi Lai et al. [9] have investigated the use of Large Language Models (LLMs) in the legal domain in the study. It talks about how artificial intelligence (AI) can help judges, automate the creation of legal documents, and boost legal research productivity. The study emphasises that although legal LLMs receive training on court rulings, statutes, and case records, data accessibility is still difficult because of privacy concerns. For text processing and decision-making, algorithms like BERT, GPT, and specialised legal models like ChatLaw and LawGPT are employed. The study does point out certain research gaps, such as biassed AI results, inconsistent datasets, and the difficulty of interpreting court rulings. There are also ethical questions about how AI-driven court rulings and predictive policing might compromise human rights. A significant drawback of legal LLMs is that they tend to maintain prejudices derived from prior legal data, which can result in unjust verdicts. The study also cautions that a judge's freedom of judgement may be limited if judicial independence is compromised by an excessive reliance on AI. Furthermore, it is challenging to evaluate the actual performance of these models due to the absence of benchmarking and real-world testing. The paper offers suggestions for enhancing legal AI, but in order to guarantee responsible adoption, it highlights the necessity of openness, equity, and improved dataset governance.

Nguyen Ha Thanh [10] discusses LawGPT 1.0, an AI-powered legal assistant optimised on GPT-3 for the legal domain, in the study. LawGPT 1.0 generates legal documents, responds to legal enquiries, and offers legal advice using the transformer architecture with attention mechanisms and fine-tuning techniques. irrespective of its strengths, the study points out a number of drawbacks, including its inability to be explained, which casts doubt on the reliability and accountability of legal judgements produced by AI. Furthermore, the model's capacity to improve responses in response to user interactions is diminished because it does not support Reinforcement Learning from Human Feedback (RLHF). Concerns about privacy, accountability, and possible bias in AI-generated legal recommendations are still unresolved from an ethical and legal standpoint. The fact that LawGPT 1.0 only supports English at this time, which restricts its use in multilingual legal systems, is another significant flaw.

Future enhancements, such as adding more language support and better explainability features, are suggested in the study but have not yet been put into use. Its practical reliability is further undermined by the lack of discussions about real-world deployment and the opaqueness of dataset sources. Even with these drawbacks, LawGPT 1.0 has the potential to increase accessibility to legal services by providing AI-powered legal aid around-the-clock.

A Non-Disclosure Agreement (NDA) prevents the precise dataset source from being revealed, but the model is trained on a sizable corpus of legal text.

**III. METHODOLOGY**

**I. Overview**

An organised approach is used in the creation of the AI-powered Legal Documentation Assistant to guarantee the precision, dependability, and effectiveness of the system. Python-based AI models are used to carry out the project, which focusses on the analysis and retrieval of legal documents. Data gathering, preprocessing, model development, document retrieval, and validation are the five main stages of the methodology. In order to improve the AI system's ability to handle complex legal texts, extract important legal terms, and retrieve pertinent legal documents in response to user queries, each phase is essential.

**II. Data Collection and Preprocessing**

**a. Legal Dataset Acquisition**

To train and improve the AI model, a large dataset is necessary. The information was collected from a variety of publicly accessible sources, such as case law collections, open-source legal documents, government legal repositories, and databases of law firms. Contracts, agreements, policies, and legal notices are among the various kinds of legal documents included in the dataset. Legal professionals examine the gathered materials to guarantee data quality and relevance, removing any out-of-date or jurisdiction-specific content that might limit the model's capacity for generalisation.

**b. Text Cleaning and Formatting**

Once collected, the raw legal data undergoes thorough pre-processing to improve its quality and ease of use for AI training. This includes:

* **Removing Special Characters and Formatting Artifacts** – Unnecessary symbols, extra spaces, and formatting errors are removed to maintain textual simplicity.
* **Tokenization** – The legal text is divided into sentences and words for structured analysis.
* **Stopword Removal** – Common but non-informative words (e.g., “the,” “is,” “an”) are filtered out to focus on meaningful legal content.

**III. AI Model Development**

1. **Legal Document Embeddings and Vector Search**

For effective document retrieval, the system makes use of text embeddings. Using Hugging Face transformer models, legal texts are transformed into numerical vector representations as part of the following methodology:

* Embedding Generation: Legal text is converted into numerical vector representation.
* Dimensionality Reduction with PCA: Principal Component Analysis (PCA) is used to minimise the size of the vector while maintaining crucial semantic data.
* Effective Similarity Search with FAISS & HNSW: The system uses FAISS (Facebook AI Similarity Search) and Hierarchical Navigable Small World (HNSW) graphs to facilitate the quick and scalable retrieval of related legal texts.

1. **Conversational Retrieval-Based Model**

The following features are included in the system to enhance the retrieval of legal information:

* Conversational Retrieval Chains: The system iteratively refines queries to deliver the most pertinent legal documents.
* Legal Context Awareness: The model efficiently retrieves pertinent legal information by comprehending user queries in context.
* Query Expansion & Re-ranking: Relevant legal terms are used to broaden queries, and the retrieved documents are ranked according to their semantic relevance.

**IV. Legal Document Retrieval**

1. **User Query Processing**

In order to obtain the most pertinent legal documents, the system analyses user queries. Important actions consist of:

* Semantic Query Embedding: This process matches stored legal document embeddings with user input by converting it into an embedding vector.
* Similarity Matching: An effective method for finding documents that are similar is to use FAISS and HNSW.
* **Ranking & Filtering:** Prioritising the most pertinent legal documents and eliminating results that aren't relevant is known as ranking and filtering.

1. **Dynamic Legal Text Retrieval**

The legal texts that were retrieved are:

* Context-Based: This ensures that the documents reflect the intent of the user.
* Relevance-based ranking: Documents with higher rankings have more relevant legal information.
* Adaptive to User Queries: In response to user input, the system improves search results.

**V. Validation and Testing**

**a. Comparison with Existing Legal Documents**

The outputs obtained are compared to pre-existing templates for legal documents to make sure:

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

**b. Performance Metrics**

Evaluation is done using the following metrics because the system is more concerned with retrieval than text generation:

* Precision and Recall: This measures how well the system retrieves pertinent legal documents.
* Mean Reciprocal Rank (MRR): Assessing the position of the pertinent document in ranked results.
* Embedding Similarity Scores: These evaluate how closely the legal documents that are retrieved match the query input.

**I. Ethical Considerations and Future Enhancements**

1. **Data Privacy and Security**

Because legal information is sensitive, security precautions include:

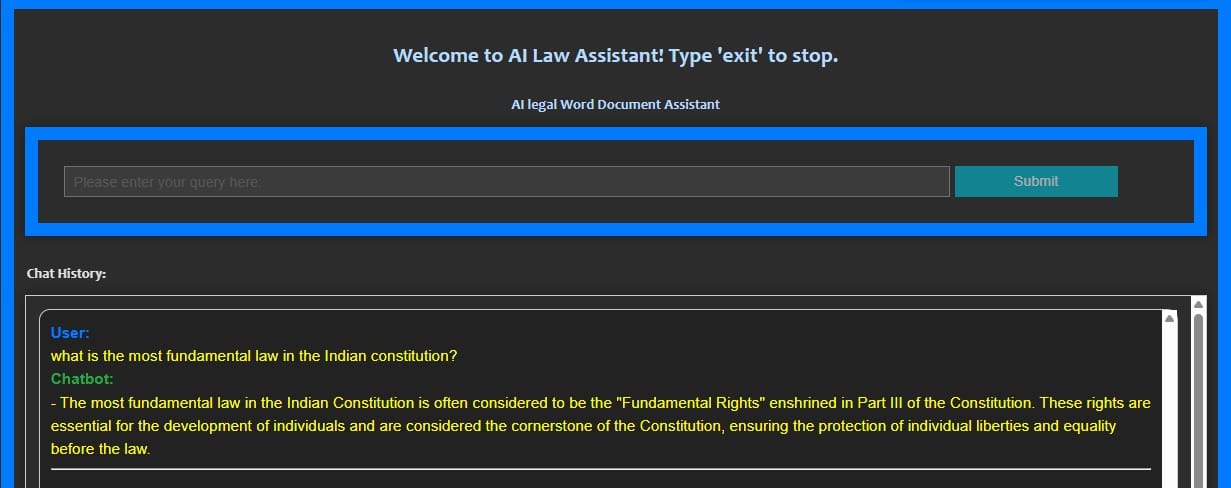
* Sensitive Data Anonymisation: This involves redacting personal information to preserve user privacy.
* Encryption Mechanisms: Protecting transmitted and stored legal data.
* Privacy Regulation Compliance: Following the General Data Protection Regulation (GDPR) and other data protection laws.

1. **Future Enhancements**

To expand functionality, the system aims to:

* **Support Multiple Languages** – Enhancing accessibility for non-English legal documents.
* **Integrate Real-Time Legal Consultation** – Providing AI-powered assistance for legal queries.

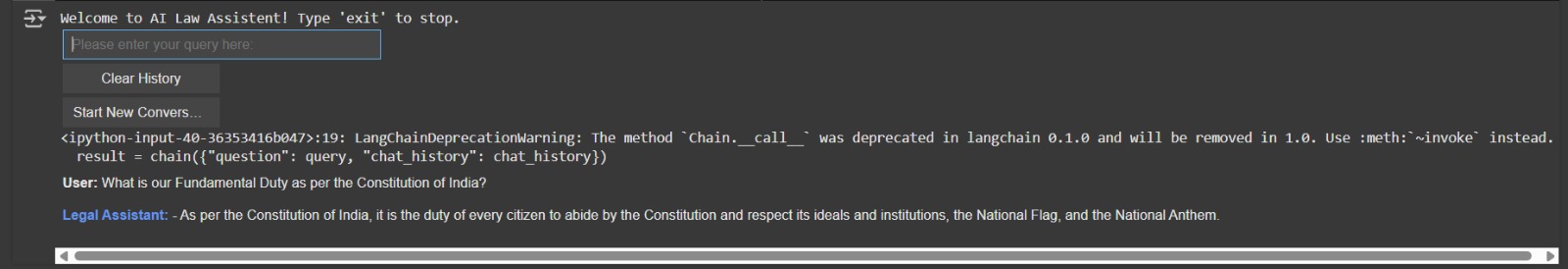
The AI-powered Legal Documentation Assistant will effectively retrieve and analyse legal texts with high accuracy thanks to this methodical approach. The system benefits legal professionals and organisations by improving legal documentation processes through the integration of text embeddings, semantic search, and conversational retrieval chains.

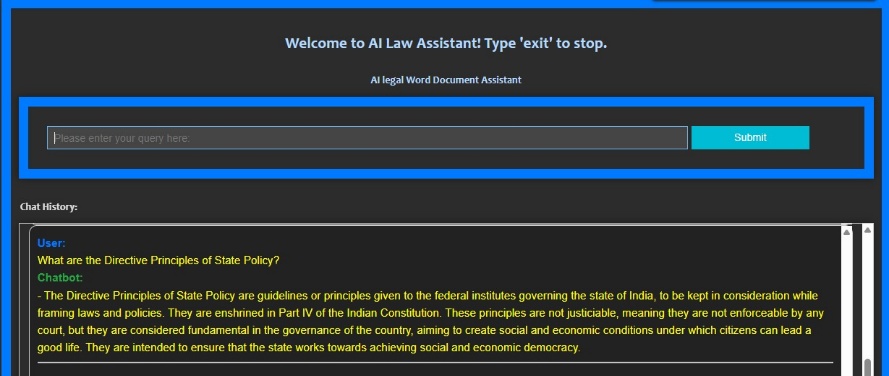
**IV. RESULTS**

**Fig. 1**

Fig. 1 shows the output interface of the AI Legal Word Document Assistant, a chatbot driven by AI and intended to help users with legal enquiries. The interface greets 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. Below this, the chat history section displays the conversation between the user and the chatbot. 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.

A document-based retrieval system is used to generate the response. The system first uses PyMuPDF (fitz) to extract and preprocess legal text from PDF files when a user inputs a query. Because the text is organised into headings and related content, the system can find pertinent information quickly. Upon receiving a query, it is analysed to find important terms and compared to the legal text that has been extracted. To find pertinent sections, the system mostly uses text-matching strategies and keyword-based search. After being retrieved, the data is formatted into a structured response and added to the chat history.

**Fig. 2**

In Fig. 2, 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. 3**

The user enquired, "What are the Directive Principles of State Policy?" in the image above. The chatbot gave a thorough explanation in response, claiming that Part IV of the Indian Constitution contains the Directive Principles of State Policy (DPSP), which serve as rules for governance in India. It emphasises that these ideas are essential guidelines to guarantee social and economic democracy even though they are not justiciable, or enforceable by courts.

**V. CONCLUSION**

Modern Natural Language Processing (NLP) methods, such as text embeddings, FAISS-based similarity search, and conversational retrieval, are used by the AI-powered Legal Documentation Assistant to expedite the processing of legal documents. Without starting from scratch, the system efficiently extracts pertinent legal clauses from a carefully curated knowledge base, guaranteeing high accuracy and contextual relevance. The assistant improves legal document retrieval by using vector-based similarity matching and sentence embeddings, which facilitates effective clause extraction and question answering. Real-time answers to legal queries are made possible by the FAISS indexing technique, which greatly maximises search efficiency. Precision, recall, and relevance scores are used to assess the system's performance, guaranteeing solid and trustworthy legal aid.

Additionally, the system's comprehension of intricate legal jargon is enhanced by the incorporation of legal text preprocessing methods such as named entity recognition (NER), lemmatisation, and stopword removal. Data encryption and anonymisation are examples of security and privacy measures that guarantee adherence to moral and legal requirements.   
By adding multi-language support, real-time legal expert validation, and wider legal domain coverage, future improvements seek to increase the system's functionality. For legal professionals, companies, and individuals looking for precise and organised legal document retrieval, the AI-powered assistant offers a scalable and effective solution with its current framework.

**VI. 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.