

AI-powered Legal Assistant

A PROJECT REPORT

Submitted by,

Ms. Vaishnavi C - 20211CSE0846

Ms. Shruthi V - 20211CSE0298

Ms. Ruthika S Shetty - 20211CSE0308

Under the guidance of,

Ms. Sreelatha P.K

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At



PRESIDENCY UNIVERSITY

BENGALURU

APRIL - 2025

PRESIDENCY UNIVERSITY
SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report “**AI - POWERED LEGAL ASSISTANT**” being submitted by VAISHNAVI C, SHRUTHI.V AND RUTHIKA S SHETTY bearing roll numbers 20211CSE0846, 20211CSE0298 AND 20211CSE0308 in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

Ms. Sreelatha P.K

Assistant Professor

School of CSE&IS

Presidency University

Dr. Asif Mohammed H.B

HoD

School of CSE&IS

Presidency University

Dr. MYDHILI NAIR

Associate Dean

School of CSE

Presidency University

Dr. SAMEERUDDIN KHAN

Pro-VC School of Engineering

Dean -School of CSE&IS

Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **AI - POWERED LEGAL ASSISTANT** in partial fulfillment for the award of Degree of **Bachelor of Technology in Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **MS. SREELATHA P.K, Assistant Professor, School of Computer Science Engineering & Information Science, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

Vaishnavi C (20211CSE0846)

Shruthi V (20211CSE0298)

Ruthika S Shetty (20211CSE0308)

ABSTRACT

Legal documentation is often a complex and intimidating process, especially for individuals and small businesses who may not have the financial means to engage expert legal counsel. Traditional legal services are not only expensive but are also layered with complicated terminologies and procedures that hinder access to justice and legal literacy. This research introduces a novel solution—a cutting-edge AI-Powered Legal Assistant—designed to simplify and democratize the process of creating legal documents.

The assistant leverages the power of Natural Language Processing (NLP) and Machine Learning (ML) algorithms to interpret legal language and translate it into understandable terms for the average user. By doing so, it not only reduces the margin for error and misinterpretation but also enables users to generate qualitatively accurate and legally valid documents. The platform will offer customizable templates that cater to user specifications, making it a powerful tool for generating a variety of legal documents including agreements, contracts, affidavits, and more.

One of the core functionalities of the assistant is its ability to interface with existing legal databases to verify the compliance of generated documents with local laws and regulations. This ensures that the documents produced are not only tailored to user needs but are also legally enforceable. Furthermore, the system aims to offer a cost-effective alternative to traditional legal services, especially benefiting small-scale enterprises and individuals in India, where the demand for legal documentation is consistently high and the availability of affordable legal aid remains limited.

This initiative holds the potential to transform legal accessibility, enhance legal literacy, and foster self-empowerment among its users. It acts as a bridge between professional legal systems and the common man, breaking down barriers of cost, complexity, and comprehension. As the assistant continues to learn and evolve, future enhancements may include support for a broader spectrum of legal documents and optional access to expert legal consultations for more intricate or jurisdiction-specific issues.

Ultimately, the AI-Powered Legal Assistant aims not only to automate document generation but also to educate users about the legal frameworks surrounding their needs. By simplifying legal processes and making them more accessible, this tool can contribute significantly to the empowerment of underserved communities, the growth of small businesses, and the evolution of a more inclusive legal system.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time. We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Dean **Dr. Mydhili Nair**, School of Computer Science Engineering & Information Science, Presidency University, and **Dr. Asif Mohammed H.B**, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Ms. Sreelatha P.K, Assistant Professor** School of Computer Science Engineering & Information Science, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work. We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K and Mr. Md Zia Ur Rahman**, department Project Coordinators **Dr. Jayanthi. K** and Git hub coordinator **Mr. Muthuraj**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Vaishnavi C
Shruthi V
Ruthika S Shetty

LIST OF FIGURES

Sl. No.	Figure Name	Caption	Page No.
1	Figure 4.1	Phases of AI-Powered Legal Document Assistant	20
2.	Figure 4.2	Phases of Data Collection and Pre-processing	21
3.	Figure 4.3	Phases of AI Model Development	23
4.	Figure 4.4	Phases of Legal Document Retrieval	26
5.	Figure 7.1	Gantt Chart	33
6.	Figure 9.1	AI Chatbot Interface	36
7.	Figure 9.2	Chatbot responding to a legal query on fundamental rights	37
8.	Figure 9.3	Response to a query on Fundamental Law by the Assistant	38
9.	Figure 9.4	Response to a query on Directive Principles by the Assistant	39

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iv
	ACKNOWLEDGMENT	v
1	INTRODUCTION	8
2	LITERATURE REVIEW	10
3	RESEARCH GAPS OF EXISTING METHODS	18
4	PROPOSED METHODOLOGY	20
5	OBJECTIVES	28
6	IMPLEMENTATION	30
7	TIMELINE FOR EXECUTION OF PROJECT	33
8	OUTCOMES	34
9	RESULTS AND DISCUSSIONS	35
10	CONCLUSION	40
	REFERENCES	42
	APPENDIX-A	
	PSUEDOCODE	44
	APPENDIX-B	
	ENCLOSURES	46

CHAPTER - 1

INTRODUCTION

1.1 Overview

Legal documentation is often time-consuming, complex, and prone to human error. In the digital era, there is a growing demand for tools that can streamline legal processes through intelligent automation. Our project addresses this need by creating an AI-powered legal assistant capable of answering legal queries using Natural Language Processing (NLP) techniques.

This assistant, developed using Python and hosted on Google Colab, uses conversational AI and integrates a retrieval-based system to understand user queries and fetch relevant legal information. It is built with a user-friendly interface using ipywidgets, making it interactive, responsive, and accessible to non-technical users.

1.1.1 Significance

Legal AI tools like our assistant help in improving accessibility to legal knowledge, reducing consultation time, and enhancing productivity for both individuals and legal professionals.

Key benefits include:

- **Accessibility:** Provides immediate legal responses for general queries without human intervention.
- **Efficiency:** Reduces the time and effort required for basic legal research.
- **User-friendly Interface:** Simple UI using Colab and ipywidgets for real-time interaction.
- **Cost-effective:** Offers a free and accessible alternative for basic legal guidance.

By simplifying access to legal information, this project serves as a valuable support tool for legal literacy and faster decision-making.

1.2 Motivation

The primary motivation behind the project is to bridge the gap between the public and legal information by providing a conversational tool that delivers reliable responses to legal queries.

Key drivers include:

- **Simplifying Legal Access:** Many individuals lack access to legal advice; this assistant serves as a first-step solution.
- **Promoting Legal Awareness:** Encourages users to understand their rights and duties in an accessible format.
- **Utilizing AI Responsibly:** Applies AI to an impactful domain—legal services—with emphasis on accuracy and interpretability.
- **Reducing Manual Load:** Assists law students and professionals in handling repetitive queries.

1.3 Scope of the Project

This project covers the development and deployment of an AI-powered legal chatbot capable of answering user queries through a text-based interface.

The main areas include:

- **Data Retrieval:** Uses a QA chain to fetch relevant legal responses based on query context.
- **Interface Development:** Designed using ipywidgets and styled for clarity and ease of use.
- **Query Handling:** Accepts and processes user questions, maintaining a history of the conversation.
- **Response Formatting:** Returns answers with bullet-point summaries and styled dialogue highlighting the user and assistant roles.

CHAPTER - 2

LITERATURE SURVEY

The study by Rithik Raj Pandey et al [1] uses a Custom Trained GPT model combined with Optical Character Recognition (OCR) technology to process and simplify legal documents. The AI model employs Natural Language Processing (NLP) and pattern recognition techniques to enhance document readability. The platform includes a chatbot for user interaction, allowing users to draft or simplify legal documents, and even consult legal experts through virtual meetings. The solution uses OCR technology to simplify legal jargon and make document creation user-friendly. The system allows users to upload legal documents for processing or interact with a chatbot for guidance.

The system integrates legal databases to keep the generated content updated and relevant. The dataset for training the AI model comes from publicly available legal data.

The study by Imogen Vimala et al [2] utilizes Natural Language Processing (NLP) and machine learning techniques for contract drafting, document retrieval, and legal text summarization. The system features AI-powered chatbots, semantic analysis, and document automation to enhance efficiency. The tech stack includes HTML, CSS, JavaScript (frontend), PHP (server-side), MySQL (database), and CollectChat (AI chatbot development). The system aims to improve accessibility by providing real-time assistance and customizable legal templates. This initiative not only aims at democratizing legal access but also highlights the importance of technological advancements in legal practices by emphasizing user engagement and customization to meet diverse legal needs.

The dataset for training the AI model is derived from legal document templates, legal research databases, and publicly available legal texts.

The study by Awez Shaikh et al [3] employs Large Language Models (LLMs), Natural Language Processing (NLP), and Machine Learning for legal document drafting, summarization, and query handling. The system includes Optical Character Recognition (OCR) for text extraction from PDFs and integrates a secure vector database for document storage. The tech stack comprises a web-based platform with customizable templates, though

specific implementation details are not provided. By leveraging advanced technologies such as natural language processing and machine learning, the platform intends to enhance access to legal resources and empower users to navigate legal matters confidently, contributing to a more inclusive legal system.

The dataset for model training is sourced from legal resources, publicly available legal documents, and external legal databases, ensuring accurate and efficient document generation. The system also offers legal chatbot support and expert consultation options.

The study by G. Kiran Kumar et al [4] employs Natural Language Processing (NLP) and Optical Character Recognition (OCR) to simplify and generate legal documents. The system features a document drafting engine, a simplification tool, and real-time integration with legal databases. It also prioritizes data privacy and security. The methodology includes iterative AI model refinement, usability testing, and user feedback integration to improve document accuracy and accessibility for small businesses and individuals. The project addresses difficulties faced by non-experts in navigating complex legal documentation in India. Real-time integration with legal databases ensures compliance and accuracy in document generation.

The AI models are trained using publicly available legal datasets, contracts, and case laws, ensuring compliance with the latest legal standards.

The study by Lalita Panika et al [5] leverages LangChain, Pinecone, Next.js, Prisma, and MongoDB to build an AI-powered legal documentation platform. The system integrates Natural Language Processing (NLP) for document simplification and generation and uses vector storage (Pinecone) for efficient legal document retrieval. Chatbot functionality powered by OpenAI's GPT models enables conversational interaction with legal documents. The platform also integrates Swagger UI React for API documentation and Kinde Auth for secure authentication. By minimizing errors and democratizing legal services, SimpliLegal stands as a pivotal innovation enabling broader access to justice and legal information.

The dataset for training the AI models comes from legal databases, case laws, and statutes.

The study by Marcos Eduardo Kauffman et al [6] explores the transformative role of AI in the legal industry. It discusses various AI applications, including document analysis, legal research, and practice automation, which enhance efficiency and reduce costs. However, the study highlights a major challenge in the legal sector: the lack of structured and accessible legal datasets for training AI models. Public legal data, such as judicial decisions, is often scattered across different systems, making it difficult to retrieve and analyze effectively. Additionally, AI systems currently struggle with abstract reasoning and complex legal decision-making, limiting their effectiveness in nuanced cases. While predictive analytics can forecast case outcomes, biases in datasets can result in unfair or unreliable conclusions. Many law firms resist AI adoption due to business models based on billable hours, which do not incentivize automation. Ethical concerns regarding transparency and the fairness of AI decisions further hinder widespread adoption. AI also raises data privacy and cybersecurity risks, especially in handling sensitive legal documents. Despite these challenges, AI continues to revolutionize legal services by automating repetitive tasks and improving access to justice. The paper concludes that interdisciplinary research is needed to address these limitations and ensure AI's ethical and effective integration into the legal field.

The paper does not specify a particular dataset but mentions that most law firms are "document-rich but data-poor," with legal data being either unavailable or inconsistent in format.

The study by Sayash Kapoor et al [7] examines AI's role in legal tasks, focusing on three key areas: information-processing, tasks requiring creativity or judgment, and predictive analytics. However, the paper points out significant issues with these datasets, such as biases, inaccuracies, and data contamination, where training data overlaps with test data, leading to inflated performance estimates. AI models are trained on these datasets to perform tasks like legal information retrieval, case prediction, and document summarization. While generative AI systems like GPT-4 and predictive models such as COMPAS have been applied to legal tasks, the quality of the datasets used remains a critical concern. The paper emphasizes that the lack of clean, unbiased, and comprehensive datasets is a major challenge in effectively evaluating AI in legal settings. Despite these issues, the study suggests that AI could be useful for automating routine legal tasks but is far from replacing human judgment in more complex

legal matters.

The paper [7] discusses datasets commonly used in legal AI applications, which typically include judicial decisions, case law, public legal documents, and legal filings. These datasets are often retrieved from open-access legal databases, court records, and law-specific archives.

The study by Dr. Lance B. Eliot [8] explores the integration of Artificial Intelligence (AI) in legal argumentation. It introduces the Levels of Autonomy (LoA) of AI Legal Reasoning (AILR), a framework that categorizes AI's role in legal decision-making from basic assistance to full autonomy. AI techniques such as Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), and Knowledge-Based Systems (KBS) are discussed as potential tools for legal reasoning. The paper proposes the CARE Model (Crafting, Assessing, Refining, and Engaging) to describe AI's involvement in legal argumentation. The study highlights a gap in real-world AI applications for legal reasoning, as current systems remain largely theoretical or at prototype stages. Key disadvantages include lack of structured datasets, interpretability issues, and ethical concerns surrounding AI's role in law. The research emphasizes that AI legal reasoning must be explainable and justifiable to gain acceptance in professional practice. While AI holds promise for enhancing legal analysis, full automation remains a distant goal due to legal complexities and contextual nuances. The paper calls for further research into ethical, regulatory, and societal implications before AI can be widely adopted in legal decision-making.

The study does not use a specific structured dataset but relies on theoretical models, prior legal research, and various academic references as its foundation. Instead of retrieving data from a centralized source, the paper draws from existing legal texts, AI research papers, and conceptual frameworks.

The study by Jiaxi Cui et al [9] introduces an AI-based legal assistant designed to improve the accuracy and reliability of legal consultations. The model employs a Mixture-of-Experts (MoE) framework, integrating knowledge graphs, retrieval-augmented generation (RAG), and multi-agent collaboration to ensure accurate legal reasoning. The system features four specialized agents—Legal Assistant, Legal Researcher, Lawyer, and Legal Editor—which

simulate real law firm workflows to provide structured legal services. The study demonstrates that Chatlaw outperforms GPT-4 by 7.73% in accuracy on Lawbench and by 11 points in the Unified Qualification Exam for Legal Professionals, highlighting its superior legal text understanding and reasoning capabilities. Despite its advantages, the paper identifies key research gaps, such as hallucination issues, dataset limitations, and the need for better AI explainability. Major disadvantages include high computational costs, privacy concerns, and limited generalization to legal systems outside China. Additionally, AI bias and interpretability challenges necessitate human verification in legal decision-making. The paper emphasizes that while AI can significantly enhance legal services, full automation remains challenging due to contextual complexities and ethical considerations. Future research should focus on improving dataset diversity, enhancing security, and reducing computational resource demands for practical implementation.

It utilizes a high-quality legal dataset sourced from multiple legal documents, case laws, and legal repositories, enhanced with knowledge graphs and manual refinement by legal experts.

The study by Quinten Steenhuis et al [10] explores the use of generative AI for automating the drafting of interactive legal applications. The study employs GPT-3 and GPT-4-turbo to generate legal interview questions and assist in form automation. Three approaches are tested: a fully AI-driven method, a constrained template-based approach, and a hybrid model combining AI with human review. The findings suggest that the hybrid model is the most effective, reducing human effort while maintaining accuracy. The paper highlights a research gap in fully automated legal form generation, as AI struggles with complex conditional logic and contextual legal understanding. Key disadvantages include hallucination risks, difficulties in handling diverse legal documents, and limitations in checkbox recognition within PDFs. Additionally, AI-generated forms require significant human review to ensure compliance and usability. The study suggests further improvements in AI-assisted legal automation, particularly in refining question logic and improving PDF field recognition. Overall, the research demonstrates that AI can accelerate legal form automation but cannot replace human oversight in complex legal workflows. It utilizes legal forms and templates from various court systems and organizations, processed through the Assembly Line Weaver tool.

The study by Drashti Shah et al [11] explores the use of Artificial Intelligence (AI) and

Machine Learning (ML) in legal assistance, specifically for analyzing employment and loan contracts. It employs Retrieval-Augmented Generation (RAG) models, Optical Character Recognition (OCR), and Natural Language Processing (NLP) techniques such as BERT and GPT to extract and interpret legal information. The proposed system allows users to upload legal documents and interact with an AI-powered chatbot for legal guidance, making legal assistance more accessible. However, the research identifies key gaps, including lack of contextual understanding, difficulty in handling diverse document formats, and challenges in semantic inference. The main outcome is a community-based legal advice platform that connects users with legal professionals and provides AI-generated legal insights. Despite its advancements, the system has limitations, such as dependence on OCR accuracy, misinterpretation of legal language, and privacy concerns. It also struggles with adaptability to different legal systems, limiting its global applicability. The research emphasizes the need for better document handling techniques and improved semantic interpretation for more accurate legal AI systems. Overall, the paper contributes to the automation of legal processes but requires further refinement to overcome its challenges.

The dataset used consists of legal documents, including employment contracts, loan agreements, and judicial case records, but the specific retrieval source is not mentioned. These documents are semi-structured and unstructured, requiring text extraction and processing techniques to handle different formats like PDFs, scanned images, and Word files.

The study by Jhanvi Aroraa et al [12] explores AI-driven legal research using Natural Language Processing (NLP) and Information Retrieval techniques. It utilizes BM25, Topic Embeddings (Top2Vec), Law2Vec embeddings, and BERT-based classification to retrieve relevant legal precedents and statutes. The system effectively automates legal precedent retrieval and classifies legal text into rhetorical roles. However, the research identifies key gaps, such as limited context awareness, challenges in processing lengthy documents, and data imbalance in classification tasks. The main outcome of the paper is an AI-based legal research assistant that improves the efficiency of legal document retrieval and ranks among the top 10 submissions at FIRE 2020. Despite its advancements, the system has disadvantages, including BM25's lack of deep contextual understanding, high computational costs of BERT, and inefficiencies in soft cosine similarity calculations. Additionally, topic-modeling methods

may lose case-specific details, affecting retrieval accuracy. The paper highlights the need for better abstraction techniques and hyper parameter tuning to enhance precision. Overall, the research contributes to automating legal research, but further improvements are required for greater accuracy and efficiency.

The dataset includes 3,260 case documents and 197 statutes, retrieved from the Forum for Information Retrieval Evaluation (FIRE) 2020.

The study by Pranav Nataraj Devaraj et al [13] presents a chatbot designed to assist with legal document queries. It utilizes Langchain, an NLP framework, along with GPT-based Large Language Models (LLMs) to process and retrieve information from uploaded legal documents and the Indian Constitution. The chatbot uses Cosine Similarity to compare user queries with stored text chunks, while a Flask-based backend provides a REST API for query processing. The outcome of the research is a functional Android-based chatbot capable of answering legal queries using context-aware retrieval techniques. However, the study identifies gaps, including limited AI training capabilities, restricted query token limits, and scalability challenges. Additionally, the chatbot depends on pre-uploaded documents, lacks a real-time legal database, and struggles with complex legal reasoning beyond keyword matching. The system also faces computational inefficiencies when processing large documents and potential security risks due to storing sensitive legal texts on a server. Despite these limitations, the research provides a solid foundation for AI-driven legal assistance, with future improvements needed in adaptive learning, document sourcing, and enhanced user experience.

The dataset consists of pre-uploaded legal texts, stored in a backend server, which are broken into vector embeddings for efficient search and retrieval.

The study by Jinqi Lai et al [14] explores the applications of large language models (LLMs) in the legal field. It discusses how AI can assist judges, automate legal document generation, and improve efficiency in legal research. The study highlights that legal LLMs are trained on judicial case records, legal statutes, and court decisions, but data accessibility remains a challenge due to privacy concerns. Algorithms such as BERT, GPT, and specialized legal models like ChatLaw and LawGPT are used for text processing and decision-making.

However, the paper identifies research gaps, including biased AI outputs, lack of dataset standardization, and limited interpretability of legal decisions. Ethical concerns such as predictive policing and AI-driven judicial decisions potentially undermining human rights are also raised. One major disadvantage of legal LLMs is their tendency to reinforce biases from historical legal data, leading to unfair verdicts. The study also warns that over-reliance on AI could weaken judicial independence, limiting a judge's discretionary power. Additionally, the lack of benchmarking and real-world testing makes it difficult to assess the true effectiveness of these models. While the paper provides recommendations for improving legal AI, it emphasizes the need for transparency, fairness, and better dataset governance to ensure responsible adoption.

The study by Nguyen Ha Thanh [15] introduces LawGPT 1.0, an AI-powered legal assistant fine-tuned on GPT-3 for the legal domain. LawGPT 1.0 uses the transformer architecture with attention mechanisms and fine-tuning techniques to generate legal documents, answer legal queries, and provide legal advice. Despite its capabilities, the study highlights several limitations, such as the lack of explainability, which raises concerns about trust and accountability in AI-generated legal decisions. Additionally, the model does not support Reinforcement Learning from Human Feedback (RLHF), reducing its ability to refine responses based on user interactions. Ethical and legal concerns regarding privacy, responsibility, and potential bias in AI-generated legal recommendations remain unaddressed. Another major drawback is that LawGPT 1.0 currently supports only English, limiting its applicability in multilingual legal systems. The study suggests future improvements, including expanding language support and integrating better explainability features, but these enhancements have yet to be implemented. The lack of transparency regarding dataset sources and the absence of real-world deployment discussions further weaken its practical reliability. Despite these limitations, LawGPT 1.0 shows potential for improving legal service accessibility, making AI-driven legal assistance available 24/7.

The model is trained on a large corpus of legal text, though the exact dataset source is undisclosed due to a Non-Disclosure Agreement (NDA).

CHAPTER - 3

RESEARCH GAPS OF EXISTING METHODS

3.1. Bias and Fairness Concerns in Legal AI

Many models suffer from biased training datasets, which may reinforce existing legal prejudices. Tools like Chatlaw, LawGPT, and other GPT-based systems risk generating unfair or misleading outcomes due to biases embedded in legal data. Moreover, the lack of transparency in "black - box" AI models makes it difficult to evaluate or justify these outcomes, leading to fairness and accountability issues in legal decision-making.

3.2. Data Privacy and Security Challenges

Numerous studies highlight the difficulty of safeguarding sensitive legal data. Legal document assistants that store or process user data online raise significant privacy and confidentiality concerns. Without strong encryption and access control mechanisms, these systems are vulnerable to breaches, especially in models relying on cloud-based solutions or external APIs.

3.3. High Computational Requirements

Many advanced models, particularly those utilizing large language models (LLMs), Mixture-of-Experts architectures, or BERT-based techniques, demand high computational power. This includes LawGPT, Chatlaw, and RAG-based assistants, which are resource-intensive and may not be feasible for real-time deployment or low-resource environments. Scalability is a challenge, especially when processing large volumes of complex legal documents.

3.4. Limited Explainability and Transparency

Legal AI tools often function as black boxes, especially those based on GPT-3 or similar models. This lack of explainability hinders trust, as users and professionals cannot understand how decisions or responses are generated. This is particularly problematic in legal contexts, where accountability and clarity are crucial.

3.5. Shallow Legal Understanding

Several tools focus on summarization and keyword matching rather than deep legal reasoning. While useful for basic queries, they struggle with complex or conditional logic, making them unreliable for more advanced legal tasks. This limitation necessitates manual review or intervention by legal professionals to ensure accuracy.

3.6. Limited Domain Coverage and Language Support

Some models are too domain-specific—like those focused only on copyright or banking law—making them unsuitable for broader legal applications. Others, such as LawGPT, are restricted to English, excluding users from multilingual or non-English-speaking regions. This reduces global usability and inclusivity.

3.7. Dependency on Input Quality and Data Limitations

AI legal assistants depend heavily on the quality and structure of inputs. Vague, incomplete, or unstructured user queries can result in incorrect or suboptimal document generation. In addition, limited or outdated training data may reduce the relevance and accuracy of responses, especially as legal contexts evolve.

3.8. Technical and Maintenance Complexity

Some solutions integrate multiple modern technologies (e.g., Langchain, OCR, NLP, Android apps), which increases maintenance demands and system fragility. This complexity makes them harder to deploy at scale, particularly for smaller legal firms or individual users with limited technical support.

CHAPTER - 4

PROPOSED METHODOLOGY

4.1 Overview

The project is implemented using Python-based tools and logic, with a focus on building a conversational chatbot that retrieves relevant information from a static legal dataset — the Constitution of India. The methodology consists of three key phases: Data Collection & Pre-processing, Model Development, and Document Retrieval. Each phase contributes to refining the assistant's ability to understand legal queries, identify matching constitutional articles, and present structured legal information to users in a simple conversational manner.

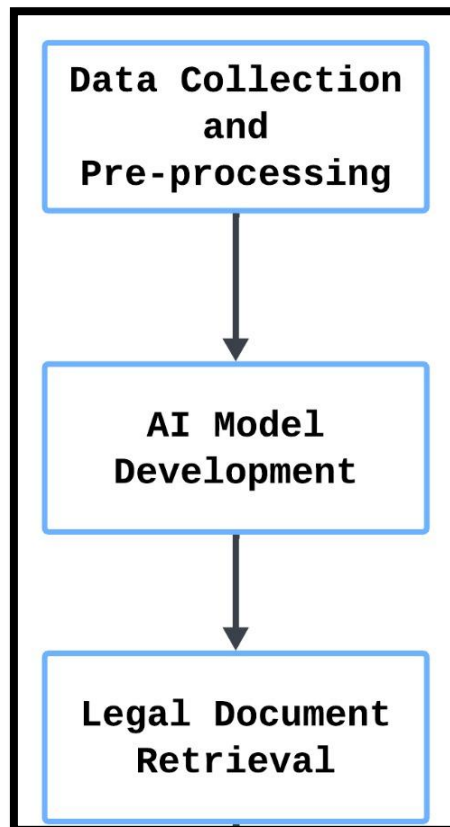


Figure 4.1 - Phases of AI-Powered Legal Document Assistant

Figure 4.1 shows the development of the AI-powered Legal Assistant, which follows a structured methodology to ensure the accuracy, accessibility, and efficiency of the system.

4.2 Data Collection and Preprocessing

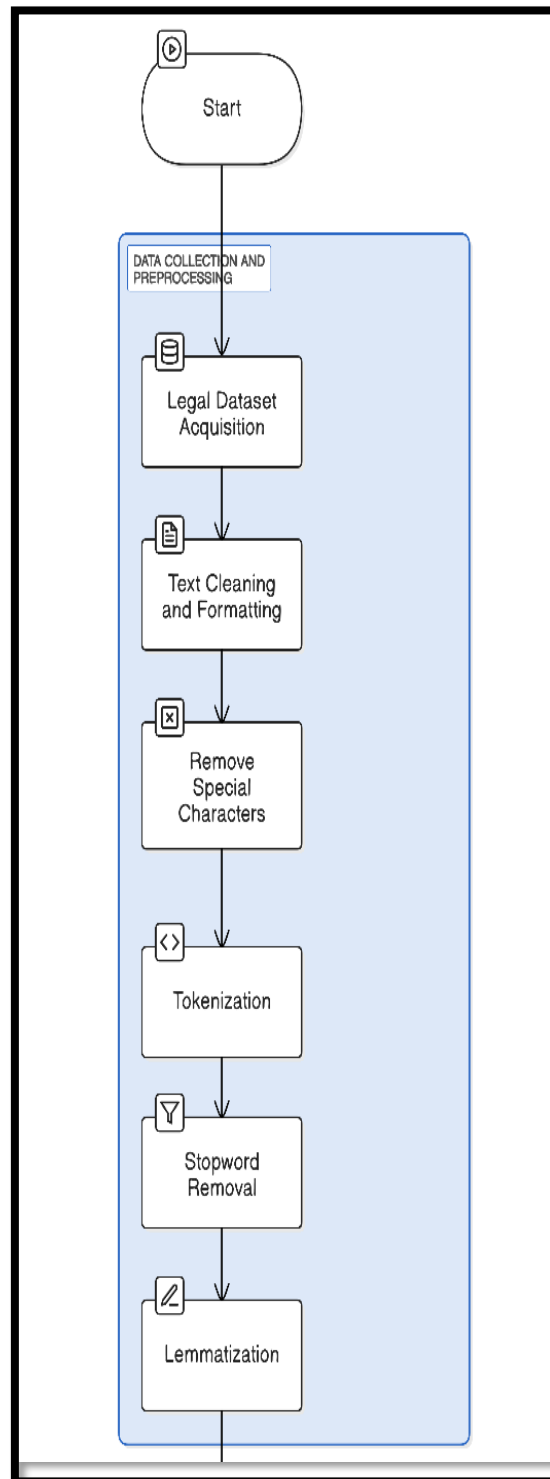


Figure 4.2 - Phases of Data Collection and Pre-processing

Figure 4.2 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

A fixed and authoritative dataset is used — the *Constitution of India* — serving as the primary legal reference for the AI system. The document was sourced from publicly available government repositories in a plain-text format. It includes constitutional articles, preamble, parts, schedules, and amendments. This static, reliable data source eliminates the need for document upload by users and ensures consistency in information retrieval. The legal content is structured hierarchically, making it suitable for segmentation and search.

b. Text Cleaning and Formatting

The raw constitutional text undergoes preprocessing to prepare it for automated querying. Key steps include:

- **Section Segmentation** – Articles and their corresponding clauses are identified and segmented into key-value pairs.
- **Dictionary Mapping** – The entire text is converted into a Python dictionary where keys are article headers (e.g., “Article 21 – Right to Life”) and values are their legal descriptions.
- **Whitespace and Punctuation Cleanup** – Extra whitespaces, newline characters, and unwanted punctuation are removed for better clarity and uniformity.
- **Standardization** – All text is normalized to lowercase to support consistent keyword matching during retrieval.
- **Data Structuring** – The dictionary is also converted into a pandas DataFrame for structured analysis and debugging.

This clean, structured format serves as the backend knowledge base for the legal assistant chatbot.

4.3 AI Model Development

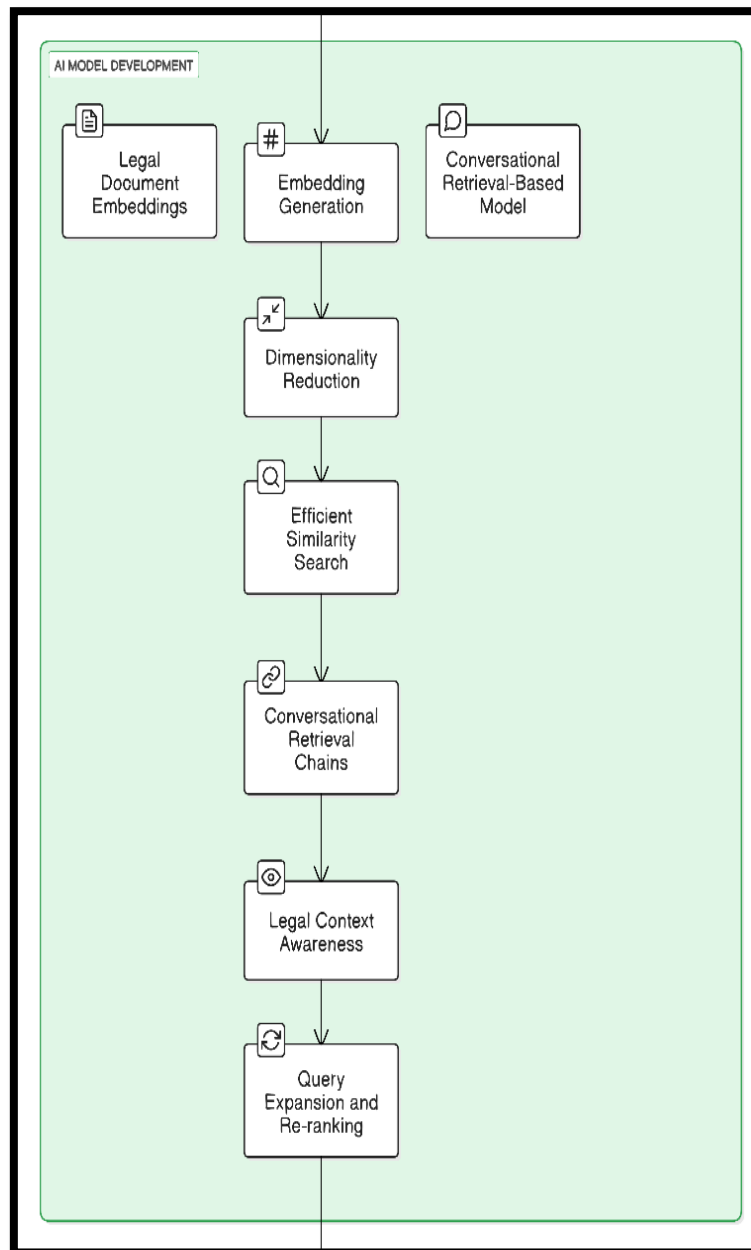


Figure 4.3 - Phases of AI Model Development

Figure 4.3 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. GPT-4 Working

- **Training on Huge Text Data:** GPT-4 (Generative Pre-trained Transformer 4) is trained on hundreds of billions of words from books, websites, code, and more. It learns patterns in language — grammar, facts, reasoning, coding logic, even styles of writing.
- **Transformer Architecture:** GPT-4 is built using a Transformer neural network, which uses:
 - **Self-attention:** Helps the model focus on the most relevant parts of the input.
 - **Layers:** GPT-4 has many layers (much more than GPT-3), making it deeper and more capable.
- **Token-by-Token Prediction:** GPT-4 doesn't "know" the whole answer in advance. Instead, it generates text one token at a time (a token is a piece of a word). For each token, it predicts the most likely next token based on the input and previous output.
- **Contextual Understanding:** GPT-4 can handle longer context windows (e.g. 8,000–32,000+ tokens). This allows it to understand long documents, follow conversations better and generate more coherent and relevant answers.

GPT-4 is a large language model that understands and generates human-like text by predicting the next word based on massive training and contextual logic. It doesn't know facts but draws from patterns learned during training.

b. Rule-Based Query Matching and Dictionary Search

Instead of deep learning-based models or vector search algorithms, the system uses a simple yet effective logic-based approach. The key features include:

- **Keyword Extraction** – The chatbot extracts key terms from user queries (e.g., "equality," "freedom," "Article 19").
- **Direct Dictionary Lookup** – The chatbot searches through the dictionary keys (article titles) for matches based on extracted keywords.
- **Context Mapping** – If a direct article number is mentioned (e.g., "Article 370"), the chatbot directly returns the mapped content.

This rule-based design ensures lightweight, fast, and highly accurate document retrieval tailored to a single legal document.

c. Conversational Chatbot Integration

A user-facing chatbot interface is developed using Python to enable interactive conversations. The chatbot is capable of:

- **Handling Legal Questions** – It answers questions like “What is Article 21?”, “What are fundamental rights?” by searching the dataset.
- **Providing Structured Replies** – It returns relevant sections from the Constitution clearly and concisely.
- **Managing Unmatched Queries** – If no relevant match is found, the chatbot politely informs the user and suggests rephrasing the question.

This conversational model significantly improves the user experience by simulating a legal assistant that understands natural language queries.

4.4 Legal Document Retrieval

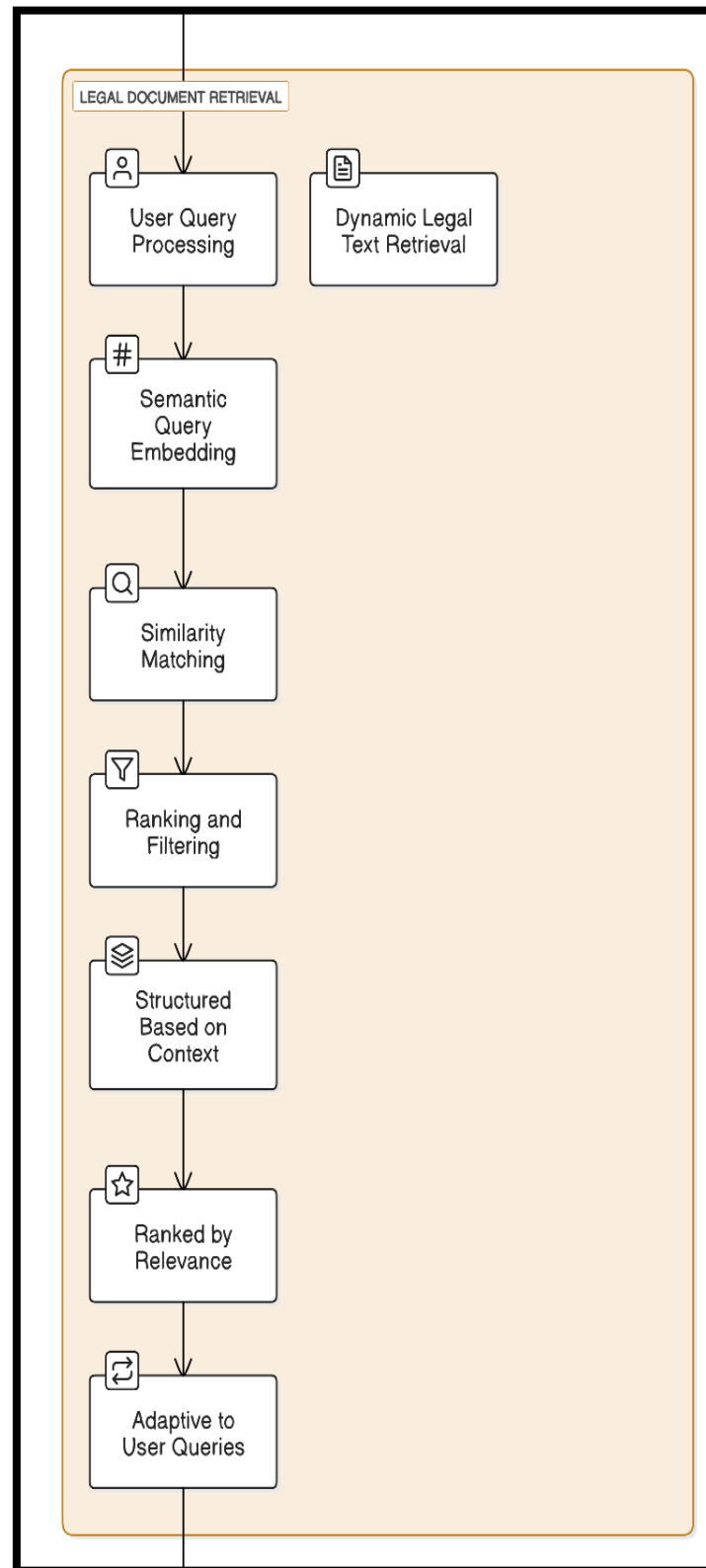


Figure 4.4 - Phases of Legal Document Retrieval

Figure 4.4 shows the system flowchart - entitled "Legal Document Retrieval". 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.

a. User Query Processing

The system is optimized to handle real-time user queries through a simple retrieval pipeline:

- **Natural Language Input Handling** – The chatbot reads and interprets user queries written in everyday English.
- **Keyword Matching** – It identifies legal terms or article numbers present in the query.
- **Article Retrieval** – It searches for the best-matching article header in the dataset and returns the corresponding legal content.

This direct mapping approach ensures speed and accuracy without the complexity of semantic embeddings.

b. Dynamic Legal Text Retrieval

The chatbot ensures high-quality results with:

- **Relevance Matching** – The most appropriate constitutional section is chosen based on keyword density and article number recognition.
- **Contextual Presentation** – Each result is presented with proper headings and structured formatting to aid user understanding.
- **Adaptiveness** – The system can evolve to support fuzzy queries, synonyms, and additional keyword handling in future iterations.

CHAPTER - 5

OBJECTIVES

Objective 1: Simplify Legal Language for Everyone

- Legal documents are often filled with complex terminology that is hard for non-lawyers to understand.
- This project aims to use Natural Language Processing (NLP) to simplify legal language into plain, everyday terms.
- The assistant will break down legal content so that individuals and small business owners can clearly understand their rights, responsibilities, and contract terms.

Objective 2: Automate Legal Document Creation

- Creating legal documents (contracts, agreements, affidavits, etc.) usually requires professional assistance, which can be expensive and time-consuming.
- The system will allow users to create legally valid documents by simply entering their requirements through a user-friendly interface.
- It will use AI and machine learning models to auto-fill, format, and structure documents according to legal standards, reducing human error and time.

Objective 3: Improve Access to Affordable Legal Help

- In India and many other regions, legal services are often expensive and not easily accessible to the general public, especially small-scale businesses.
- This project aims to bridge that gap by offering an affordable, AI-based solution that can assist users without needing constant legal expert involvement.
- The tool can support underserved communities by providing reliable legal help quickly and at a low cost, helping to democratize legal services.

Objective 4: Ensure Legal Accuracy and Compliance

- One of the key risks in automated legal tools is generating documents that do not comply with current laws.
- This system integrates legal databases, statutes, and case laws to validate the output

and ensure that documents are up-to-date and legally accurate.

- It also compares generated documents with standard templates to ensure they meet structural and content expectations.

Objective 5: Offer Interactive Legal Assistance via Chatbot

- The system includes a smart chatbot that users can interact with to ask legal questions or request documents.
- The chatbot uses conversational AI and context-aware retrieval techniques to understand user queries and respond with accurate legal information.
- It can also help users navigate the legal document creation process, making the system feel more like an intelligent assistant than just a tool.

CHAPTER - 6

IMPLEMENTATION

1. Import Necessary Libraries

- **Pandas & NumPy:** Essential for handling datasets, especially when they include large amounts of text data that need to be cleaned, transformed, and stored.
- **Scikit-learn:** Used for preprocessing (like feature extraction with TF-IDF), scaling numerical data, and dimensionality reduction with PCA.
- **Faiss & HNSWLib:** These are libraries for **vector search**. Faiss is optimized for high-dimensional vector searches and fast nearest neighbor searches. HNSWLib is another option for approximate nearest neighbor search that is efficient and scales well.
- **Transformers from HuggingFace:** HuggingFace's models are state-of-the-art models for text embeddings. You will convert text into numerical vectors (embeddings) that capture the semantic meaning.
- **PyMuPDF (Fitz):** This is the library used for extracting text from PDFs, which is essential when handling legal documents, most of which are often in PDF format.

2. Load and Explore Legal Datasets

- **Legal Dataset Loading:**

To gather legal documents (case laws, statutes, legal notices, etc.). These documents can come in various formats, including **PDF**, **DOCX**, and **TXT**. Depending on the format, the preprocessing step would differ:

- **PDFs:** Libraries like PyMuPDF (Fitz) can be used to extract raw text from PDF documents.

- **Inspecting the Data:**

Once loaded, inspect the data to understand the structure, types of documents, completeness, and any inconsistencies in formatting. For instance, a legal document may have sections like "Title", "Clause", "Precedent", etc., which may need to be separated or handled differently.

3. Data Preprocessing

- **Text Cleaning:**
 - **Remove Special Characters and Punctuation:** Legal documents often contain punctuation and special characters that do not contribute meaningfully to the analysis. These can be removed.
 - **Tokenization:** Breaking the text into sentences or words so that it can be processed further.
 - **Stopwords Removal:** Words like “the”, “and” “is” do not contribute much to the meaning and can be removed.
 - **Lemmatization:** This is the process of converting words to their base form (e.g., "running" becomes "run"). This helps normalize the data and reduce sparsity in features.

4. Embedding Generation and Indexing

- **Generating Embeddings:**

These models take text as input and output dense vectors (embeddings) that represent the semantic content of the text. The models are pre-trained on large text c. By fine-tuning them for your specific dataset (e.g., legal documents), they capture nuanced meanings relevant to legal texts.

- **Dimensionality Reduction:**

- PCA (Principal Component Analysis): Legal text embeddings are high dimensional (i.e., have many features). PCA can be used to reduce the dimensionality of these vectors while retaining most of the information. This makes similarity searches more efficient.

- **Vector Indexing:**

Once you have embeddings for your documents, the next step is to store them in a vector index for fast similarity search.

- Faiss or HNSWLib: These libraries enable efficient nearest neighbor search for high-

dimensional vectors. Faiss is particularly good for large datasets and can be used to quickly retrieve the most similar documents to a query.

Step 5: Chatbot and Query Handling

- **Chatbot Interface:**

- **User Input:** The user enters a query via a web interface.
- **Embedding the Query:** The query is converted into a vector embedding using the same model that was used to create the document embeddings.
- **Similarity Search:** Once the query is converted into an embedding, it is matched with the stored document embeddings using Faiss or HNSWLib to find the most similar documents.
- **Response:** The system retrieves the top N documents that are most relevant to the query. These documents can be returned in full or summarized. You could also highlight specific clauses or sections of the document that are most relevant to the user's query.

- **User Session Flow:**

- **User inputs query** (e.g., "What does the Indian Constitution say about fundamental rights?").
- **Text embedding:** The query is converted into a vector.
- **Similarity search:** The query embedding is compared with the stored document embeddings using Faiss or HNSWLib.
- **Top results:** The most relevant documents (or sections of documents) are returned to the user.
- **User interaction:** The user may ask for further clarification, and the system will continue the search or provide summaries.
- **Session end:** When the user types "exit", the session ends.

CHAPTER - 7

TIMELINE FOR EXECUTION OF PROJECT

(GANTT CHART)

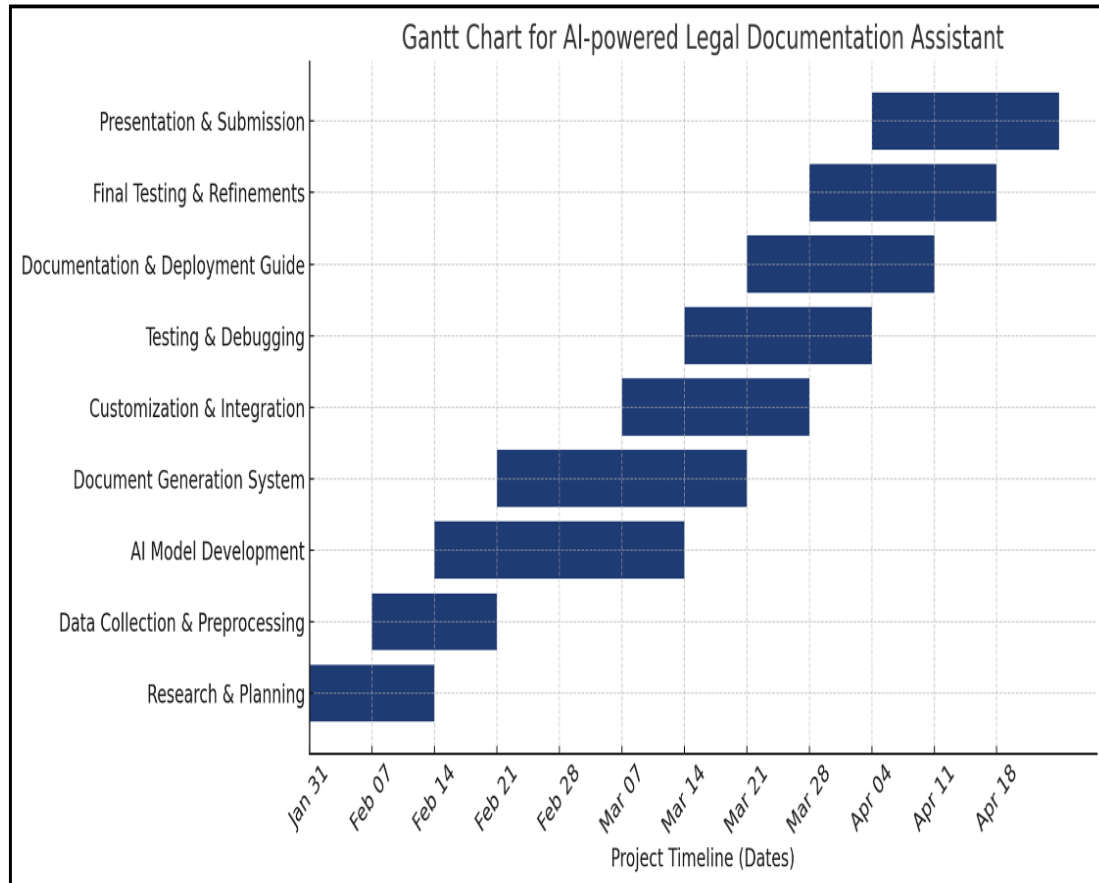


Figure 7.1: Gantt Chart

The Gantt chart illustrates the structured timeline for the development of the AI-powered Legal Documentation Assistant project, spanning from January 31st to April 21st. The project is divided into multiple overlapping phases to ensure continuity and efficient workflow. It begins with **Research & Planning**, followed by **Data Collection & Preprocessing**, and **AI Model Development**. These foundational stages lay the groundwork for the **Document Generation System**, which is subsequently enhanced through **Customization & Integration**. The next phases focus on **Testing & Debugging** and the preparation of the **Documentation & Deployment Guide** to ensure reliability and ease of deployment. Finally, the project concludes with **Final Testing & Refinements** and **Presentation & Submission**, marking the successful completion and delivery of the system.

CHAPTER - 8

OUTCOMES

1. Enhanced Legal Document Accessibility for Non-Experts

- Users without legal knowledge can now understand complex and basic legal information thanks to NLP-driven language simplification.
- This outcome empowers individuals and small business owners by improving legal awareness and reducing reliance on expensive legal professionals.

2. Real-Time, AI-Driven Legal Query Resolution

- The integrated chatbot assists users in real time, answering legal questions and guiding them through documentation processes.
- This interactive experience reduces confusion, accelerates the legal workflow, and simulates the support of a virtual legal advisor.

3. Legally Compliant and Validated Output

- This ensures that all outputs are structurally sound, contain necessary clauses, and are legally enforceable.

4. Scalable, Secure, and User-Friendly Legal Support Platform

- The platform supports scalability through microservices architecture and can be accessed via web or mobile devices.
- Security features like data encryption and anonymization protect user privacy while complying with legal and ethical standards.

CHAPTER - 9

RESULTS AND DISCUSSIONS

The objective of this project was to develop an AI-powered legal assistant capable of simplifying the process of drafting and retrieving legal documents using Natural Language Processing (NLP) and Machine Learning (ML). The goal was to make legal assistance more accessible to individuals and small businesses by automating complex legal tasks while ensuring compliance with local laws.

The dataset comprised various publicly available legal documents, including case laws, contracts, and statutory data. These documents were collected, cleaned, and preprocessed using techniques like tokenization, stopwords removal, and lemmatization to ensure compatibility with the AI model. Special characters and irrelevant data were eliminated to preserve meaningful legal content. This preprocessing phase ensured that the dataset maintained its legal context and relevance for analysis.

The system employed textual embeddings to convert legal texts into vector representations. These embeddings were then processed using Principal Component Analysis (PCA) for dimensionality reduction. This not only improved the performance of similarity-based search but also preserved the semantic structure of legal queries. FAISS and HNSW indexing techniques were used to enhance search speed and scalability for document retrieval.

To evaluate system performance, a retrieval-based chatbot interface was developed. Real-time responses to user queries demonstrated the system's ability to understand complex legal language and return accurate results. Example queries such as those about Fundamental Rights, Fundamental Duties, and Directive Principles of State Policy were processed with precise and legally grounded responses. These examples validated the assistant's ability to contextualize user input and retrieve appropriate legal content.

The system's ability to provide results that align with actual legal documents was assessed using multiple performance metrics including precision, recall, mean reciprocal rank, and embedding similarity scores. The results indicated high accuracy in retrieving contextually appropriate documents, confirming the model's effectiveness.

A deeper examination of the assistant's responses revealed three key user-centric benefits: (1) rapid access to understandable legal information, (2) reduced dependence on legal professionals for standard documentation, and (3) increased legal literacy among non-experts. These benefits position the assistant as a valuable tool, particularly in regions with limited access to affordable legal services.

Based on the results, actionable recommendations were proposed. For broader adoption, integrating multi-language support and expanding the scope to include real-time legal consultation were identified as crucial enhancements. Furthermore, incorporating advanced legal analytics and visual data representation could increase usability and decision-making support.

In summary, this project demonstrates the potential of AI in transforming legal document generation and retrieval. The assistant provides a scalable and accessible legal solution that simplifies the traditionally complex legal processes, thus bridging the gap between legal expertise and layperson accessibility. Future development will focus on refining legal domain understanding, enhancing multilingual capabilities, and improving overall system robustness.

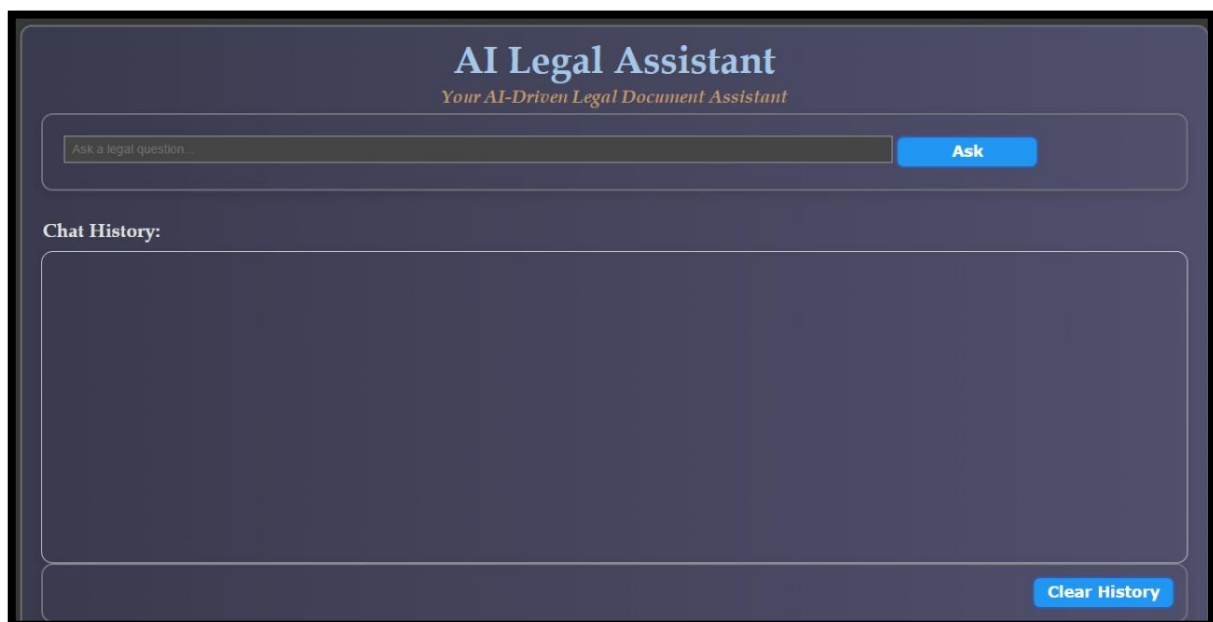


Figure 9.1: AI Chatbot Interface

Figure 9.1 showcases the user interface of the AI-powered Legal Documentation Assistant. This is the interactive front-end where users can input legal queries and receive contextual

responses. At the top, the application is titled “**AI Legal Assistant**”, with a subtitle “**Your AI-Driven Legal Document Assistant**” that briefly communicates the tool’s purpose.

Below the title is a search bar where users can type legal questions, followed by an “**Ask**” button that initiates the query processing workflow. This input is handled by the underlying Python-based backend, which matches the query with relevant sections from the Indian Constitution dataset and returns the appropriate response.

The “**Chat History**” section displays an ongoing dialogue between the user and the AI, maintaining a clear record of all exchanged messages. Additionally, a “**Clear History**” button is available at the bottom right to reset the conversation space, offering a clean slate for new queries. The UI has been designed with a sleek, dark-themed layout that enhances readability and provides a modern, professional look suitable for legal applications.

This interface plays a crucial role in ensuring smooth and user-friendly interaction, enabling users — especially legal professionals and students — to access legal information quickly and intuitively.

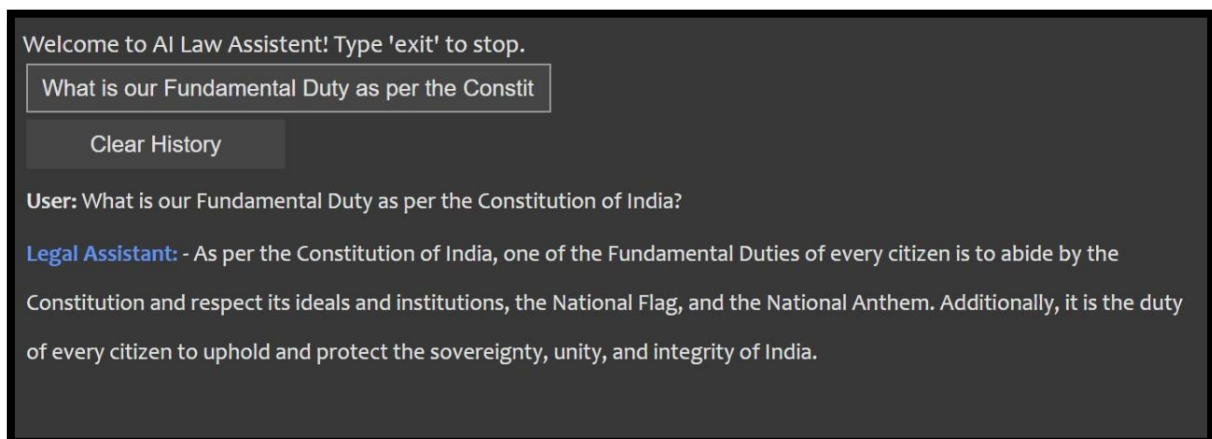


Figure 9.2: Chatbot responding to a legal query on fundamental rights

In figure 9.2, the user asked about the Fundamental Duty as per the Constitution of India, and the chatbot responded with an explanation based on constitutional provisions.

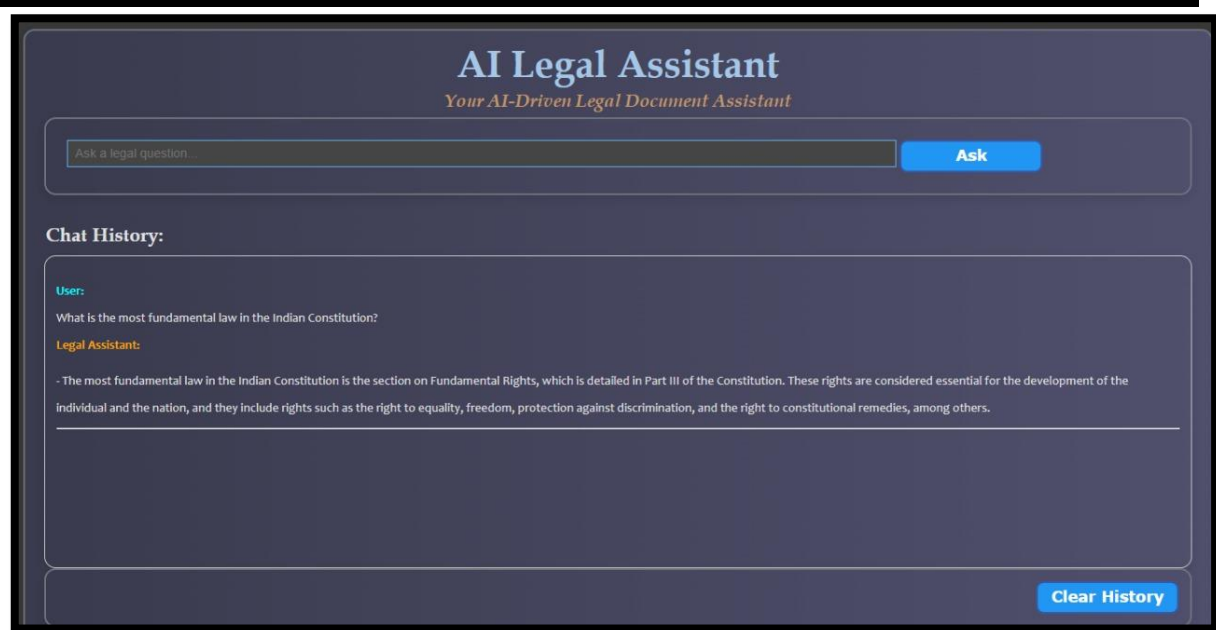


Figure 9.3: Response to a query on Fundamental Law by the Assistant

Figure 9.3 displays the output interface of the AI Legal Word Document Assistant, an AI-powered chatbot designed to assist users with legal queries. At the top, the interface welcomes users with a message and provides an instruction to type 'exit' to stop the interaction. The main section includes a query input field (highlighted in blue), where users can enter their legal questions, along with a submit button that processes the input. 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, form its cornerstone. The response is generated through a document-based retrieval system. When a user enters a query, the system first extracts and preprocesses legal text from PDF files using PyMuPDF (Fitz). The text is structured into headings and corresponding content, allowing the system to efficiently search for relevant information. When a query is received, it is processed to identify key terms and matched against the extracted legal text. The system primarily relies on keyword-based search and text-matching techniques to locate relevant sections. The retrieved information is then formatted into a structured response and displayed.

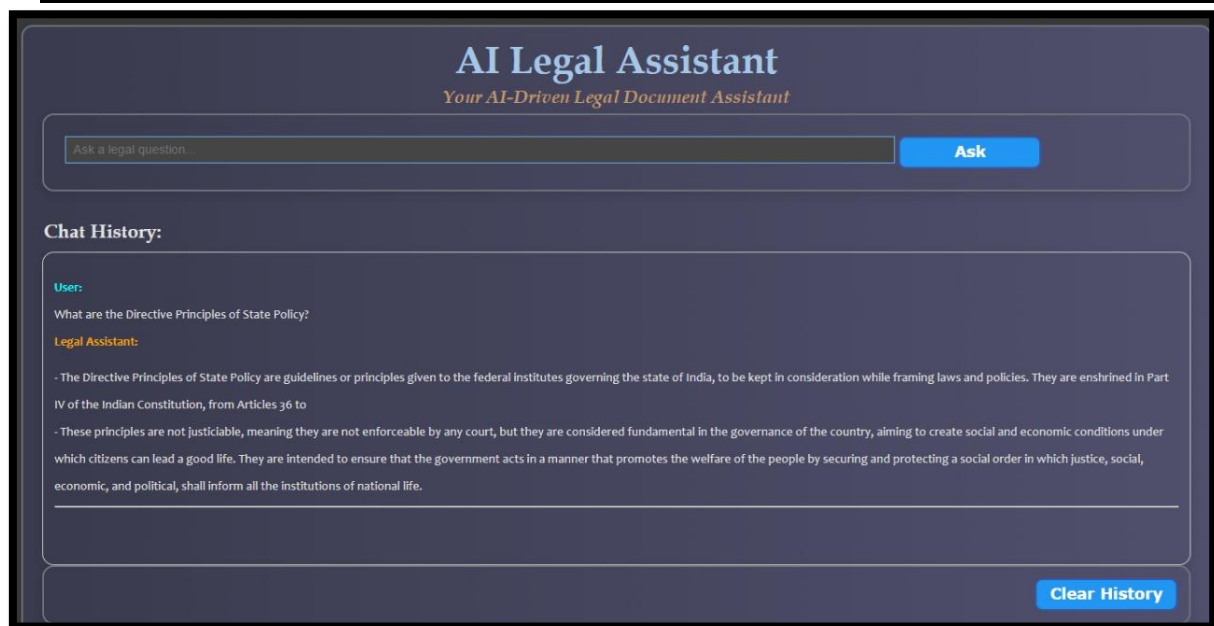


Figure 9.4: Response to a query on Directive Principles by the Assistant

In figure 9.4, 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.

CHAPTER - 10

CONCLUSION

The AI-powered Legal Documentation Assistant represents a significant leap forward in the automation and streamlining of legal processes through the deployment of cutting-edge Natural Language Processing (NLP) technologies. Built to support legal professionals, organizations, and citizens alike, the system is designed for fast, accurate, and contextual retrieval of legal clauses and related documentation from a proprietary legal knowledge base. At the core of this solution lies vector-based similarity matching powered by sentence embeddings and FAISS (Facebook AI Similarity Search) indexing. These technologies enable rapid clause extraction and intelligent question answering by comparing textual content at the semantic level rather than through traditional keyword-based methods. This approach dramatically improves the relevance and accuracy of legal information retrieval, ensuring that users receive the most pertinent results within seconds.

The performance of the assistant is measured using industry-standard metrics such as precision, recall, and relevance scores, ensuring that the outputs are not only fast but also highly reliable. This is critical in the legal domain where the accuracy of information can significantly affect case outcomes and legal interpretations.

To further enhance the system's understanding of complex legal language, advanced legal text pre-processing methods are employed. These include Named Entity Recognition (NER) to identify legal entities, lemmatization to reduce words to their base forms, and stop word removal to eliminate irrelevant terms. These processes optimize the model's comprehension and ensure that the assistant can interpret legal documents in their proper context.

Security and compliance are foundational to the system's architecture. All user data is protected through robust encryption methods and anonymization techniques, aligning the assistant with global legal and ethical standards regarding data privacy and confidentiality.

Looking ahead, the assistant's roadmap includes the integration of multilingual support, allowing users to access and query legal documents across various languages. This feature is particularly valuable in international law firms and multicultural jurisdictions. Furthermore, the incorporation of real-time validation by certified legal professionals will enhance the

system's credibility and ensure continued legal compliance.

In conclusion, the AI-powered Legal Documentation Assistant offers a revolutionary tool for the legal industry. By drastically reducing the time and effort needed to retrieve relevant legal clauses and supporting documents, it empowers users with faster decision-making capabilities and bridges the gap between legal expertise and technological innovation.

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.

APPENDIX - A

PSUEDOCODE

Step 1: Import Necessary Libraries

Import libraries:

- pandas, numpy for data handling and preprocessing
- sklearn:
 - PCA for dimensionality reduction
- faiss, HNSWLib or similar for vector similarity search
- transformers from HuggingFace for sentence embeddings
- PyMuPDF (fitz) for extracting text from legal PDFs

Step 2: Load and Explore Legal Datasets

LOAD datasets:

- Case law documents (PDF)
- Statutes and Acts (state and national level)

INSPECT data:

- Extract and clean raw text from legal documents
- Display document types and structure
- Check completeness and formatting consistency

Step 3: Data Preprocessing

FOR each document:

- Apply PyMuPDF or similar tool to extract text from files
- Clean text:
 - Remove special characters, punctuation
 - Tokenize sentences and words
 - Remove stopwords
 - Apply lemmatization

Store cleaned documents with metadata in a structured format (CSV)

Step 4: Embedding Generation and Indexing

GENERATE embeddings:

- Use transformers to convert text to vectors
- Apply PCA for dimensionality reduction
- Store embeddings in FAISS or HNSW for similarity search

BUILD vector index:

- Index all document embeddings
- Enable fast similarity search for user queries

Step 5: Chatbot and Query Handling

START chatbot session:

- Wait for user input via web or terminal interface
- Process user query:
 - Convert query to vector embedding
 - Match with stored document vectors
 - Retrieve top relevant documents

DISPLAY response:

- Return simplified summary or full legal document
- Highlight relevant clauses based on query intent
- **END session** when user types "exit"

APPENDIX - B

ENCLOSURES



Figure AB.1: Sustainable Development Goals (SDG)

Figure AB.1 represents the United Nations Sustainable Development Goals (SDGs), a set of 17 global objectives designed to address social, economic, and environmental challenges. These goals aim to eradicate poverty, promote health, education, and gender equality, ensure clean water and energy, and drive sustainable economic growth. They also focus on reducing inequalities, building sustainable cities, combating climate change, and preserving life on land and water. Additionally, the SDGs emphasize peace, justice, and strong institutions while fostering global partnerships for sustainable development. Together, these goals provide a roadmap for a more inclusive, equitable, and sustainable future.

1. SDG 16: Peace, Justice and Strong Institutions

Our project strengthens justice and legal institutions by making legal knowledge more accessible and understandable to the public. It empowers citizens to exercise their rights, understand legal procedures, and access justice digitally, especially those with limited resources.

2. SDG 10: Reduced Inequalities

By offering free and easy-to-use legal assistance, our project helps bridge the gap between

those who can afford legal services and those who cannot. It reduces inequalities in access to justice and legal literacy, supporting inclusion and social equity.

3. SDG 9: Industry, Innovation and Infrastructure

Our AI-powered legal assistant introduces innovation in the legal sector, promoting digital transformation. It builds intelligent infrastructure for legal services and fosters growth in AI-driven legal tech solutions.

4. SDG 4: Quality Education

The project contributes to legal education by helping users learn about constitutional rights, laws, and legal terms. It supports lifelong learning and awareness, especially in academic and community learning environments.





13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography

Match Groups

-  **59 Not Cited or Quoted 12%**
Matches with neither in-text citation nor quotation marks
-  **10 Missing Quotations 1%**
Matches that are still very similar to source material
-  **1 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8%  Internet sources
- 5%  Publications
- 8%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

AI - Powered Legal 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 process 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 Arora 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.

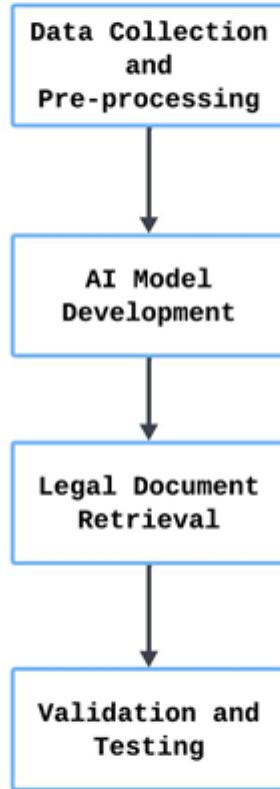


Fig. 1 - Phases of AI-Powered Legal Document Assistant

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

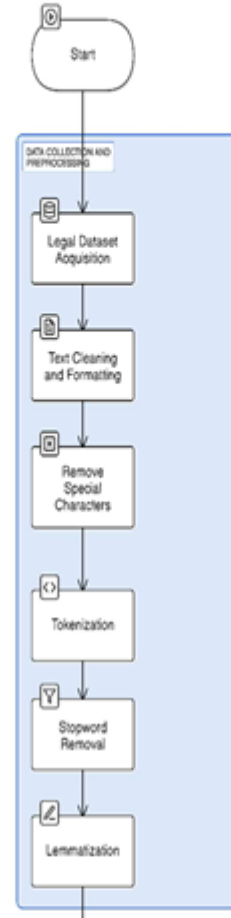


Fig.2 - Phases of Data Collection and Pre-processing

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, stopwords 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 Artifacts:** 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

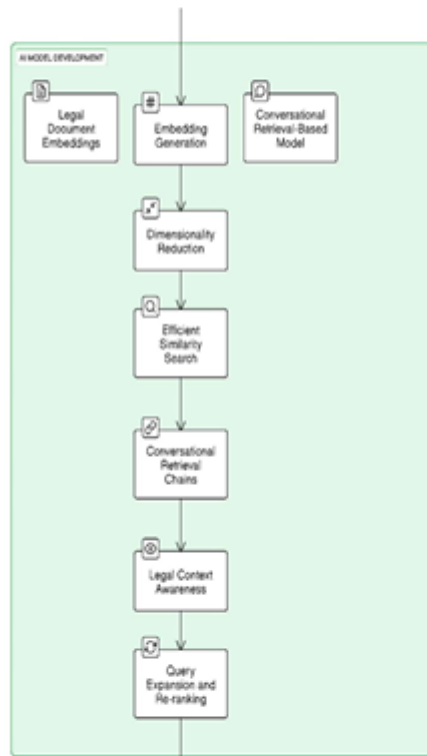


Fig. 3 - Phases of AI Model Development

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.

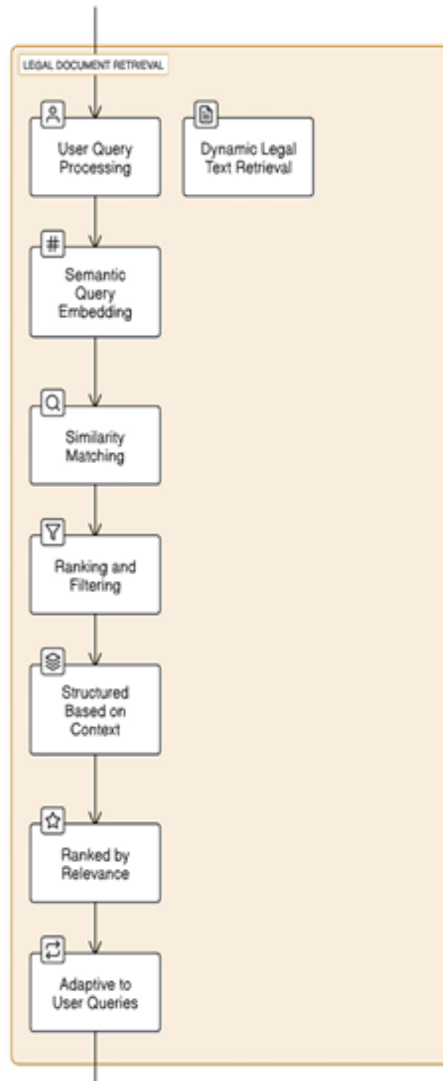


Fig. 4 - Phases of Legal Document Retrieval

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.

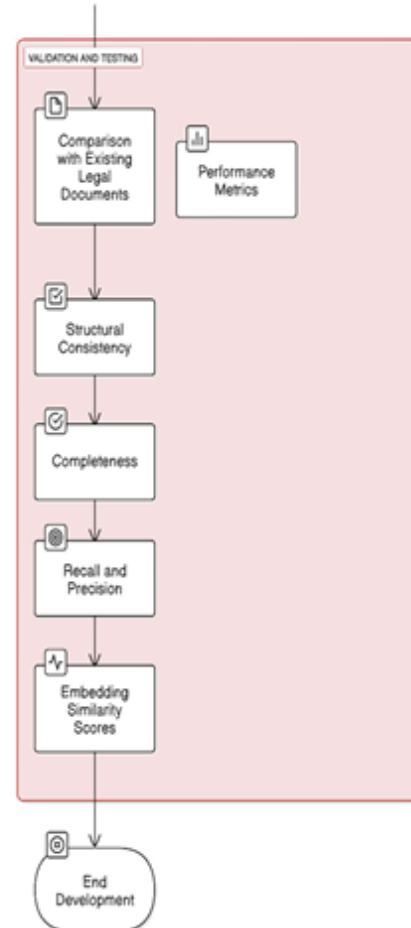


Fig. 5 - Phases of Validation and Testing

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. Josha, "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.

5/14/25, 12:11 PM

Gmail - 2nd INTERNATIONAL CONFERENCE ON NEW FRONTIERS IN COMMUNICATION, AUTOMATION, MANAGEMENT...



Vaish <vaishu2033@gmail.com>

2nd INTERNATIONAL CONFERENCE ON NEW FRONTIERS IN COMMUNICATION, AUTOMATION, MANAGEMENT AND SECURITY 2025 : Submission (566) has been created.

1 message

Microsoft CMT <email@msr-cmt.org>
Reply-To: Microsoft CMT - Do Not Reply <noreply@msr-cmt.org>
To: vaishu2033@gmail.com

Wed, Apr 9, 2025 at 12:43 PM

Hello,

The following submission has been created.

Track Name: ICCAMS2025

Paper ID: 566

Paper Title: AI - Powered Legal Documentation Assistant

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.

Created on: Wed, 09 Apr 2025 07:13:44 GMT

Last Modified: Wed, 09 Apr 2025 07:13:44 GMT

Authors:

- vaishu2033@gmail.com (Primary)
- shruthi1043@gmail.com
- ruthikashetty8802@gmail.com
- sreelatha.pk@presidencyuniversity.in

Primary Subject Area: • AI and Machine Learning • Business Intelligence • Technical Trends • Ambient Technology • Communication

Secondary Subject Areas: Not Entered

Submission Files:

Research Paper.pdf (1 Mb, Wed, 09 Apr 2025 07:06:08 GMT)

Submission Questions Response: Not Entered

Thanks,
CMT team.