2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN) | 979-8-3503-6717-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIPCN63822.2024.00048

Bettercall: AI based legal assistant

Manikrao Dhore
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
manikrao.dhore@vit.edu

Aryan Vimal
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
aryan.vimal2001@gmail.com

Avish Agrawal
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
avishagrawal999@gmail.com

Raghav Bajaj
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
bajajraghav487@gmail.com

Rohan Barde
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
rohancrrm@gmail.com

***Abstract*—This research introduces an AI-based chat-bot designed to enhance access to legal and judicial information in India. The system leverages advanced natural language processing techniques to transform legal texts into vector embeddings, facilitating semantic search capabilities that improve the user experience in retrieving relevant legal information. The chat-bot aims to provide primary legal aid, support informed decision-making, offer instant access to legal rights and information, and promote legal awareness among users. The methodology includes rigorous data collection, cleaning, and pre-processing processes to ensure high accuracy and reliability. The research also navigates to a potential implementation of the solution and addresses the complexities, results and challenges faced by the authors. The research highlights the novelty of the approach and addresses existing gaps in scalability, multilingual support, and domain coverage, presenting a comprehensive system ready for real-world deployment.**

Index Terms*—Vector Embeddings, Semantic Searching, Web Scrapping , Legal Informatics

I. INTRODUCTION

For a considerable segment of the Indian populace, especially those from marginalised communities or those who lack legal literacy, access to legal information and understanding continues to be an enormous difficulty. A user- friendly and simply available platform that meets the heterogeneous linguistic terrain of India is becoming increasingly necessary as a result of this pressing requirement. The suggestion made in this research to remedy the issue is to create a multilingual, user-friendly digital assistant that can offer legal knowledge and help. With the ability to communicate in several languages, the imagined digital assistant aims to improve accessibility and overcome linguistic obstacles. Delivering clear and simple legal information to users of all devices—including desktop computers, tablets, and smartphones—is its main goal.This platform aims to address common legal queries, offering guidance on citizenship rights, complaint filing procedures, access to legal aid services, obtaining legal documents, and understanding the legal implications of specific actions.

In addition, the digital assistant wants to be a one- stop shop for legal knowledge, doing more than just answering questions and offering details on a wide range of legal subjects like criminal law, family law, property law, and labour law. Through the integration of cutting-edge technology and a user-centered methodology, this study aims to make a positive impact on the democratisation of legal information, promoting greater legal literacy and empowerment across various demographic groups.

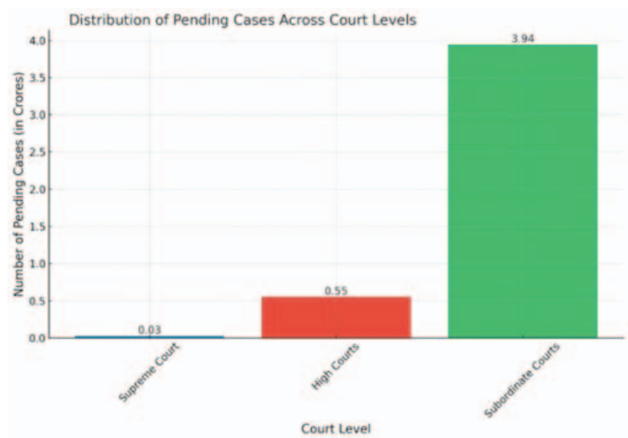

Fig. 1. Distribution of Pending cases across court levels

II. LITERATURE SURVEY

The pervasive problem of sexual assault and its frequently unsolved nature are addressed in the research article ”LAW-U: Legal Guidance Via Artificial Intelligence Chatbot for Sexual Violence Victims and Survivors,” which was presented at IEEE Access in 2021. In order to protect the identity of victims and survivors, the research presents LAW- U, an AI chatbot. The methodology uses Natural Language Processing (NLP) pipelines that were created using a large dataset of 182 instances pertaining to sexual violence that were heard by the Thai Supreme Court. LAW-U’s training is based on mock-up dialogues created by legal experts based on Supreme Court decisions. Testing on a hold-out dataset revealed an astounding

979-8-3503-6717-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICIPCN63822.2024.00048

88.89 percentage accuracy rate in correlating user input with pertinent cases from the Supreme Court, indicating that LAW-U is ready for deployment in real-world legal aid applications.

A thorough analysis of large-scale language model (LLM) assessment techniques may be found in the research paper "A Survey on Evaluation of Large Language Model," which was released on arXiv in 2023. Three crucial dimensions—what to evaluate, where to assess, and how to evaluate—are the centre of attention. The principal goal is to provide scholars studying LLM evaluation with crucial information that will aid in the creation of more advanced models. This survey summarises and analyses LLMs' evaluation techniques in an attempt to go beyond simply choosing the "best" benchmark or evaluation process. By doing this, it hopes to pinpoint both good and bad examples, spot novel trends in assessment procedures, and—most importantly—mark out fresh difficulties and prospects for further study. The results demonstrate how well LLMs generate text using accurate and fluid language expressions, performing exceptionally well in language tasks such as text classification and sentiment analysis. The study does, however, highlight the tendency for biassed outcomes as a result of innate biases and errors in the production process. Furthermore, LLMs struggle with difficult logic and reasoning skills, which frequently leads to misunderstandings and mistakes in challenging situations.[2]

The research paper titled "Artificial Intelligence in Law," published in IEEE in 2020, delves into the exploration of artificial intelligence (AI) technologies within the realm of law. Focused on current applications, capabilities, and limitations of AI algorithms in legal processes, the paper aims to inform readers about the diverse ways in which AI is influencing legal practices and systems. Through a thorough review and analysis, the paper examines various applications of AI in the legal domain, drawing on case studies, tools, and platforms that leverage AI algorithms for automating legal tasks. Noteworthy applications discussed include predicting case outcomes, estimating case durations, drafting legal documents, and simplifying complex cases, particularly in areas such as copyright and patents where the subject matter is vast. The overarching objective is to comprehensively analyze the impact of AI in the legal domain, emphasizing its multifaceted roles, such as analyzing extensive data sets to identify patterns, predicting legal outcomes and behaviors, and assisting in the precise drafting of legal documents. The paper concludes by underscoring AI's transformative role in streamlining legal procedures, allowing legal professionals to focus on intricate tasks, and ultimately enhancing the overall efficiency of legal practices.[3]

The research paper titled "Attention Is All You Need," presented at EuroVis 2017, introduces the Transformer neural network architecture as a groundbreaking solution for sequence-to-sequence tasks. The primary objective is to overcome the limitations of traditional recurrent and convolutional models by proposing a more efficient and parallelizable architecture. The Transformer architecture comprises an encoder and decoder model. The encoder transforms words into vectors, ensuring semantically related words have similar vectors, and introduces positional encoding. It then employs self-attention to generate attention vectors for each word based on its context. These attention vectors are subsequently fed into the decoder, enabling accurate predictions of the next word. The Transformer architecture surpasses existing models in machine translation tasks, achieving state-of-the-art results on benchmark datasets. The model's success underscores the effectiveness of attention mechanisms and parallelization in sequence processing. This innovative approach has since become a foundational architecture in various natural language processing tasks.[4]

In the research paper "An Approach to Get Legal Assistance Using Artificial Intelligence," the goal is to improve word meaning disambiguation through the use of semantic linkages and Word2vec-based word embeddings. Main objective is to achieve better context-sense similarity computations and outperform current unsupervised systems, especially in the SENSEVAL-3 English lexical sample challenge. The method uses Word2vec to build sense definition vectors and context sentence vectors. To determine the similarity scores between these vectors, cosine similarity is used. The approach uses sense relations from WordNet to improve sense definitions. Potential answer sensations are identified when the similarity score is less than a threshold, in which case it is paired with sense distribution probabilities extracted from a sense-tagged corpus, SEMCOR. The suggested approach shows remarkable effectiveness with a 50.9 percent (or 48.7 percent in the absence of sense distribution probability) result. In the SENSEVAL-3 English lexical sample task, its performance outperforms multiple unsupervised systems and exceeds a number of baselines, including the original, simplified, modified, and LSA Lesk. The results highlight how well the suggested strategy works to advance word sense disambiguation in the context of legal help.[5]

The research paper titled "An Approach to Get Legal Assistance Using Artificial Intelligence," presented at IEEE in 2020, introduces a Virtual Legal Assistant (VLA) as a novel solution to address the challenge of pending legal cases in India through the utilization of Artificial Intelligence (AI) technologies. The methodology outlined in the paper involves the development of a Virtual Legal Assistant (VLA) using AI, consisting of various components. These include Text Analytics for case preprocessing and similarity calculation, a Knowledge Base for storing relevant data, a Question Generation Engine to transform context into queries, and a Bot for user interaction. The primary goal is to assist users with legal cases by leveraging AI-driven technologies. Furthermore, the paper presents a conceptual framework for the Virtual Legal Assistant (VLA) that harnesses AI to potentially streamline legal cases, offer guidance, predict outcomes, and enhance public engagement with the legal system. This innovative approach holds promise for addressing the complexities associated with pending legal cases in India, showcasing the potential impact of AI in the legal assistance domain.[6]

The research paper titled "Improving Access to Justice with

Legal Chatbot," published by MDPI in 2020, addresses the objective of enhancing access to justice through the development of AI-powered chatbots. These chatbots are designed to provide legal information to both the general public and employees of the National Bank of Canada. The primary focus of the chatbots is to offer answers to frequently asked questions (FAQs), with specific emphasis on topics related to immigration and banking. The authors employed Information Retrieval (IR)-based chatbot technology to create two chatbots dedicated to providing legal information on immigration issues and addressing legal concerns relevant to the job tasks of bank employees. The paper delves into the utilization of Bag-of-Words (BoW) and Term Frequency- Inverse Document Frequency (TF-IDF) techniques in this context. Real interactions from approximately 30 users were collected to create a test dataset, ensuring its representativeness of actual user questions. The authors discuss the challenges associated with fine-tuning models on large datasets, highlighting the extreme costs involved, particularly when dealing with legal datasets for an entire country. The findings contribute to the discourse on leveraging AI-driven chatbots to facilitate access to legal information, particularly in the realms of immigration and banking, thereby advancing the goal of improving access to justice.[7]

The research paper titled "Interactive Analysis of Word Vector Embedding," presented at EuroVis 2018, aims to elucidate the concepts of vectorization and word embeddings, exploring their intricacies through the lens of visualization. The methodology employed involves a comprehensive literature survey, encompassing 111 papers from diverse communities, with a particular focus on domains such as Human-Computer Interaction. The authors further enriched their understanding through collaborative sessions with various experts. The paper serves as a valuable resource in unraveling the underlying concepts of vector embeddings by leveraging visualizations. It not only contributes to a deeper comprehension of these concepts but also delves into the usability and implementation aspects of the methods discussed. This research provides insights into the interactive analysis of word vector embedding, shedding light on the role of visualization in enhancing our understanding of this critical aspect of natural language processing.[8]

The research paper titled "Evaluating the Stability of Embedding-based Word Similarities" published in the Transactions of the Association for Computational Linguistics in 2018, investigates the variability of word embedding models. The authors, affiliated with Cornell University, explore how minor changes in training data can significantly impact the nearest-neighbor distances across various algorithms. The study highlights the sensitivity of embeddings to training corpus variations, particularly in smaller datasets, and recommends averaging over multiple bootstrap samples to ensure reliable distance calculations. The methodology involves a thorough analysis of embedding algorithms' sensitivity to changes in document order and presence, employing both fixed and bootstrap sampling techniques. The findings underscore the need for caution when using single embedding models for linguistic analysis, advocating for the use of multiple models to mitigate variability. The conclusion emphasizes the importance of acknowledging and addressing the inherent instability in embedding-based word similarity measures to enhance the robustness and reliability of linguistic research.[9]

The research paper titled "Efficient Training and Inference of Large Language Models: A Comprehensive Review" published in 2024, meticulously explores the methodologies and technological advancements in training and deploying Large Language Models (LLMs). It delves into the evolution of LLMs, emphasizing cost-efficient strategies in data preprocessing, training architectures, model compression, and parallel computation. The findings highlight the transformative impact of the Transformer architecture on LLM effectiveness across various applications, including machine translation and sentiment analysis. The paper concludes by identifying future research directions, underscoring the importance of refining LLM training and inference for broader and more efficient application in diverse fields.[10]

The research "Experiences of Intimate Partner Violence during Lockdown and the COVID-19 Pandemic," by Minna Lyons and Gayle Brewer, was published in the Journal of Family Violence in 2021. The study examines the effects of the COVID-19 lockdown on individuals who were already victims of intimate relationship violence (IPV). Through a qualitative thematic analysis of 50 Reddit forum postings, the study discovers four main themes: the use of COVID-19 by abusers, service disruption, leaving prepared, and elements that worsen abuse or pain. The findings show how the epidemic has considerably exacerbated IPV, emphasising the need for targeted interventions and victim crisis support.[11] The study "LMSYS-CHAT-1M: A Large-Scale Real-World LLM Conversation Dataset" was published in September 2023. It introduces LMSYS-Chat-1M, a dataset of one million real-world conversations with large language models (LLMs), aiming to provide insights into human-LLM interactions. This dataset, unique for its scale and diversity, was gathered from public engagements on the Vicuna demo and Chatbot Arena website. It supports research in various areas, including content moderation models, safety benchmarks, instruction-following models, and the generation of challenging benchmark questions. The paper underscores the potential of LMSYS-Chat-1M to advance LLM capabilities through detailed analysis and practical applications, highlighting the importance of understanding human interactions with LLMs in real-world scenarios.[12]

The paper "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective" published in PeerJ Computer Science in 2016, presents a pioneering study on forecasting the outcomes of cases at the European Court of Human Rights (ECtHR) based solely on the textual content of case documents. Employing a binary classification approach, the research leverages N-grams and topic models to analyze the text, achieving an average prediction accuracy of 79 percent. The study underscores the substantial role of the factual aspects of cases in predicting decisions, aligning with legal realism theories that suggest judicial decisions are significantly influenced by the case facts. This work illuminates the potential of Natural Language Processing in enhancing the understanding and prediction of judicial decisions, offering valuable tools for legal professionals and researchers.[13]

The paper ”MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction” published by the Harbin Institute of Technology and the Institute of Electronic and Information Engineering of UESTC in Guangdong, introduces a novel neural network model for predicting legal judgments. With funding from the NNS Foundation of China and the NKRD programme of China, this model attempts to increase trial efficiency by offering thorough judgement forecasts that include relevant charges, legal articles, and prison sentences. With an emphasis on the facts, the defendant’s persona, and pertinent legal articles, the method applies attention-based neural networks to analyse textual case descriptions, attaining state-of-the-art performance on assessment criteria across real-world datasets.[14]

The article ”Artificial Intelligence in Law” by K. Nikolskaia and V. Naumov, presented at the 2020 FarEastCon, discusses the application of AI in legal practices. Highlighting the potential of AI to automate routine legal work, it explores various AI applications in law, including prediction algorithms for court case outcomes and legal analytics platforms. The paper emphasizes that while AI cannot fully replicate legal thinking, it can significantly enhance the efficiency and accuracy of legal analyses and decision-making processes.[15]

The literature survey identified a significant gap regarding the lack of accessible legal assistance solutions for the general public. Existing studies emphasized the challenges in data processing and the need for more accurate and efficient legal query systems. This gap underscored the necessity for an advanced AI-driven solution to enhance the accuracy and inclusivity of legal assistance services, which led to our proposed solution to this problem.

III. PROPOSED SOLUTION

We propose the use of semantic searching. Semantic searching is an advanced search technique that aims to improve the accuracy of search results by understanding the contextual meaning of the search terms as opposed to just matching keywords. Unlike traditional keyword-based searches that rely heavily on exact word matches, semantic searching seeks to comprehend the intent behind the query and the contextual relationships between terms. This approach leverages natural language processing (NLP), machine learning, and knowledge graphs to deliver more relevant and precise results.

To understand semantic searching, one must take note of the following terms:

1) Tokenization: Tokenization refers to the process of breaking the textual data into smaller units called tokens. These tokens can be words, phrases, or even smaller units like characters, depending on the specific needs of the application.

Research Area	Key Focus	Methodologies/Technologies Used
Legal Guidance Chatbots	Providing legal advice to sexual violence victims	NLP pipelines, Supreme Court case data
Evaluation of Large Language Models	Reviewing evaluation methods for LLMs	Benchmarking, analysis of LLMs' proficiency
AI in Legal Practices	Exploring AI applications in law	Predictive algorithms, legal analytics platforms
Legal Judgment Prediction	Predicting legal judgments using neural networks	Multichannel attentive neural networks
AI for Sexual Violence Victims	Legal assistance through chatbots	NLP, Supreme Court case data
Improving Access to Justice with Legal Chatbot	Enhancing access to legal information	IR-based chatbot technology, BoW, TF-IDF
Interactive Analysis of Word Vector Embedding	Understanding vectorization and word embeddings	Visualization techniques, literature survey

Fig. 2. The identified scope of work

2) Vectorization: Vectorization is a process that converts textual data into a numerical representation as ML models cannot directly work on raw text. The vectors generated capture the context of words with similar meaning. This results in words with similar meaning having similar vector representations, also known as vector embeddings.
3) Cosine Similarity: Cosine Similarity[4][5] is a metric used to measure how similar two vectors are. This technique is used as traditional methods for vector checking would be highly inefficient in searching for similarity in vectors of large sizes. It measures the angle of two vectors in a multi-dimensional space. The formula for Cosine Similarity is as follows:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

4) Chunking: To enhance the working of LLM with vector embeddings formed using pre-trained transformers, it is essential to have details that have both semantic and contextual similarity to the text which is being converted into vector embedding. To streamline this, large texts are segmented into several smaller chunks which allows for more accurate results. The process of chunking is extremely important as it is a big factor that can contribute to the accuracy and reliability of the model.

With knowledge of the above terms, the working of semantic searching can be divided into 4 steps:

- Input processing - It involves extracting data and performing the tokenization and chunking processes.
- Embedding generation - Embeddings are generated of the chunked data.
- Similarity calculation - Cosine similarity is performed
- Result ranking - The most relevant chunks are retrieved and ranked based on the output of the previous step.

The above steps are explained in greater detail in the Methodology section below.

IV. Methodology

The methodology elaborates on the comprehensive working of the proposed solution, divided into the Training and Application phases, detailing the processes involved in each stage. The Training phase contains creating and updating the database while the Application phase contains the actual flow of the user's query.

The methodology also pulls points from the system that was developed by us to check out the validity of the proposed solution and to figure out metrics like accuracy and results.

The training phase is critical for preparing the legal assistant to process and understand legal queries effectively. It involves several key steps the intricate workings are as follows:

A. Training phase:

Given in Figure 3. is the training phase, training involves 5 steps:

1) The first step involves collecting the required legal data, essential for training the AI model. The knowledge base used for this is the website, a highly reliable and up-to-date online repository of Indian legislation. The data is collected using various web-scraping techniques in a section-wise format for every act, ensuring a comprehensive collection of legal texts. .
2) After accumulation, the data has to undergo cleaning and formatting. It is essential to not only extract the text of the sections but also to collect additional metadata such as the name of the act, act number, section number, and section title. This metadata is pivotal for effectively responding to user queries and facilitating research by enabling users to navigate through legal documents efficiently. Additionally, data cleaning also includes processes like removing unnecessary whitespaces, unknown characters & any additional metadata that is not needed.
3) The clean data then undergoes the processes of chunking and tokenization as mentioned in section III above.
4) The chunks that get generated as a result in the above step are then vectorized, i.e. vector embeddings are generated from those chunks. (see Vectorization in section III).
5) These data points are then stored in a database that supports vector embeddings alongwith their textual and metadata in order to interact with the user queries.

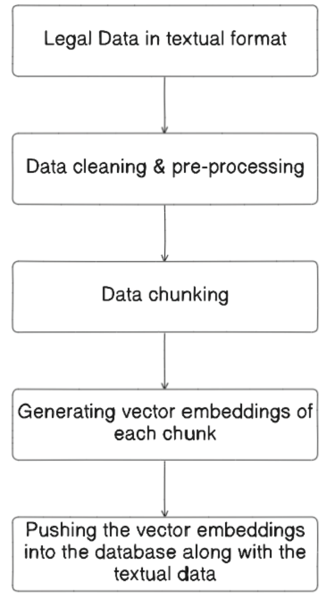

Fig. 3. Training Phase

B. Application phase:

Given in Fig. 4 is the end-to-end user query flow of the application. The application phase outlines the operational flow of the application that was developed to test the solution in responding to user queries, involving the following steps:

1) The user interface, features an input field that allows users to submit their legal queries The frontend would then pass this query to the backend.

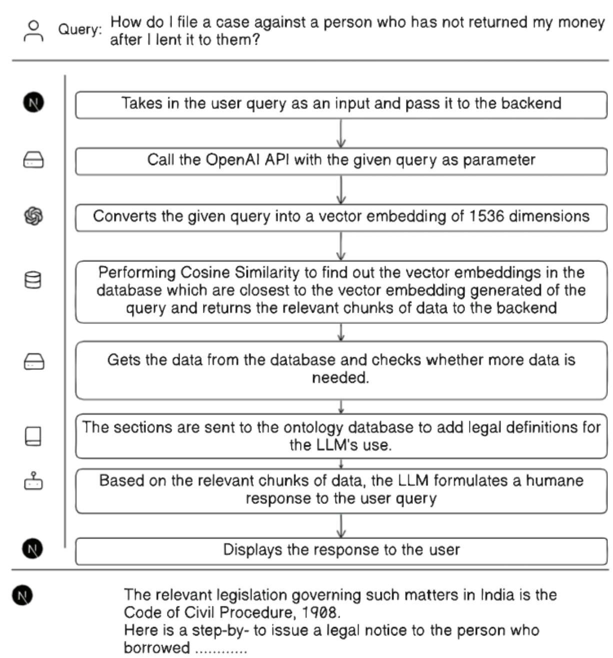

Fig. 4. Application Phase

2) Upon receiving a user's query, the LLM(Large Language Model) generates a vector embedding of the query.
3) Then the system's backend interfaces with the database to execute a Cosine Similarity operation, utilizing the vector embedding produced from the user's query. This process facilitates rapid comparisons, given that the project's vector embeddings operate within a 1536-dimensional space. After this, legal sections that most closely match the user's inquiry, based on semantic similarity, are identified and retrieved for further processing by the backend.
4) Upon receival of the data from the vector database, the system employs an ontology database (more on ontology in section V). Using the metadata received in step 3, the system gets the ontology for the particular act number of the sections. After receiving the ontology, extra optimizations were done to minimize space, cost and time by filtering out only those words that were used in the sections needed.
5) Combining the retrieved sections, user query, and ontology, the AI model generates a comprehensive, human-like response.
6) Finally, the generated response is delivered to the user through the frontend interface, completing the query-response cycle.

‘

V. CHALLENGES

Due to the absence of access to a pre-existing database containing acts and their sections, it was necessary to construct a database utilizing data publicly accessible from the indiacode.nic.in website. In the context provided below, ”data” denotes the content within a section, ”metadata” refers to the section's number, title, the act's name, and year it belongs to, and ”site” denotes the indiacode.nic.in website.

Initially, the decision was made to download the acts' PDFs and systematically scrape them, dividing the content into paragraphs of approximately 300-500 tokens each. This strategy aimed to ensure a consistent token count across all chunks within the database.

Challenges emerged during this process:

While implementing this approach, it became evident that chunks often encompassed data from multiple sections, potentially leading to lines from one section being divided into separate chunks. This presented difficulties in accurately storing metadata for each chunk, particularly in isolating the section number it pertained to. Consequently, defining metadata fields within the database became problematic, posing challenges for addressing queries such as ”What does a specific section of an act state?”

Given the inefficiency of the PDF-scraping approach, a pivot was made to a web-scraping strategy. The structure of indiacode.nic.in facilitated the extraction of both data and metadata directly from the site, promising a more organized database structure and enhanced responsiveness of the model. However, continual adjustments to the web-scraping approach were necessary to address emerging challenges.

Versions of the ideation process unfolded as follows:

Initially, the plan involved extracting data from all sections, embedding it, and subsequently integrating it into a relational vector database (Supabase being the chosen platform for this project) along with corresponding metadata.

In practice, upon receiving a user query, the model generated a vector embedding from the query, employing Cosine Similarity to identify sections whose embeddings closely matched the user query's embedding in the vector space. These identified sections were then provided to the Language Model (LM), specifically OpenAI's GPT-3.5 in this case, as context to generate a response. While this approach generally proved effective, ontology emerged as a concern, particularly regarding keywords and their contextual meanings within the documents. Inaccurate interpretation of keywords could yield disparate outcomes. However, sections containing these crucial definitions tended to be extensive, making it impractical to pass the entire document ontology due to space limitations.

To address this, initially, exploration was conducted into appending keywords and their definitions from the ontology to each section. This approach aimed to equip each section with its ontology, facilitating more precise responses from the LM. However, this method encountered challenges:

- Data duplication: Appending the keywords to every section would mean that if a keyword is in multiple sections (which is very very common), all those sections will have the same data in them, which would lead to data duplication and overall lower storage efficiency as every section will be bloated from the keywords.
- Reduced accuracy: Since all sections would now contain some data that is similar to other sections, the embeddings generated of these sections would be closer together in the vector space, which would prove to generate less accurate results for the Cosine Similarity operation. This in turn, would reduce the accuracy of the sections that would be received for the user query and in turn, a less accurate handling of the user query.
- Context duplication: As mentioned above, multiple sections were bloated with the same keywords. In cases where the Cosine Similarity operation results were such that the resultant sections contained the same keywords, the LLM would be passed the keywords multiple times as context which would again lead to less space for the user query.
- Higher cost: Since, the keywords would be a part of the section itself, each section would require a higher number of tokens while generating a vector embedding which would result in a higher cost to build and maintain the data. This would also play a part in a higher cost in the input for the LLM.

To mitigate these issues, a decision was made to relocate the ontology segment of the data from the relational vector database to a separate, key-value based non-relational database, specifically MongoDB. Since the reliance was primarily on keywords rather than vector embeddings, this decision proved beneficial. The MongoDB database contained keyword definitions mapped to the respective acts they belonged to.

Upon receiving a list of sections from the Cosine Similarity operation, cross-referencing was done with the corresponding document names to identify keywords within the relevant documents.

VI. RESULTS AND DISCUSSION

In assessing the efficacy of our AI-based legal chatbot tailored for Indian legal and judicial data, two key evaluatory parameters were considered and meticulous examination of each parameter provides a comprehensive analysis of the system's efficacy in addressing the needs of its target users. Both parameters provide unique insights into the performance and effectiveness of the system, contributing to a comprehensive understanding of its capabilities. In conducting the evaluatory analysis for precision and recall values of the AI-based legal assistant chatbot, we undertook a meticulous process to ensure a comprehensive assessment of the system's performance. Here's an overview of the steps that were followed: Selection of Queries: A diverse set of legal queries, representative of potential user inquiries spanning various legal domains and complexities were compiled , to thoroughly evaluate the chatbot's response capabilities. Standard Creation: For each query within this dataset, we manually identified and labeled correct responses or relevant legal documents/information, establishing a "gold standard" for comparison purposes. Query Processing and Response collection: Each query from the test dataset was fed into the chatbot and responses were recorded. This process was critical in capturing the system's ability to reference legal documents, sections of law, or provide direct answers. Relevance Assessment: The chatbot's responses for each query were compared against the pre-defined standard to determine relevance, categorizing the outcomes into true positives, false positives, and false negatives.

- True Positives (TP): Responses correctly identified by the chatbot as matching the gold standard.

- False Positives (FP): Irrelevant responses provided by the chatbot not aligning with the gold standard.

- False Negatives (FN): Missed responses that were relevant according to the gold standard but not identified by the chatbot.

Calculation of Precision and Recall:

Precision: We calculated the precision to assess the accuracy of the chatbot's responses relative to the total responses generated, using the formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: We determined the recall to evaluate the chatbot's efficacy in identifying all relevant information, utilizing the formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Legal Domain	Precision	Recall
Family Law	0.85	0.80
Property Law	0.90	0.88
Labor Law	0.87	0.85
Criminal Law	0.93	0.90

Fig. 5. Precision and Recall scores across legal domains

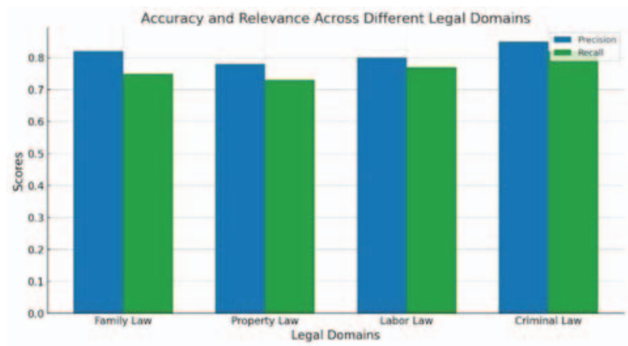

Fig. 6. Accuracy and Relevance across legal domains

The precision and recall analysis presented in Table 5 reveal the system's performance in delivering accurate and relevant legal information across various legal domains. High precision scores indicate the system's ability to minimize false positives, while respectable recall scores signify its capability to retrieve relevant information effectively.

Figure 5 illustrates the precision and recall scores across different legal domains, providing a visual representation of the system's performance in delivering accurate and relevant legal information

Accuracy and reliability in the model depended primarily on the data extracted and the chunking of data. The bot needed to always fetch the right chunk relevant to the query to provide accurate results. The chunking mechanism used in this solution divided each act by its sections and subsections to ensure a comprehensive collection of legal text, along with a legal ontology to provide further details for the terms used in the legal text. The data extracted were from official government websites and were constantly updated to include any amendments, thereby keeping the performance of the model up to date. The model was continuously tested against predefined responses to achieve and maintain highly accurate and reliable results. Additionally, the model was further evaluated by taking user feedback, as illustrated in the process below.

User satisfaction and usability were evaluated through user feedback and usability testing. Positive feedback and high usability scores indicate that the chat-bot interface is intuitive, responsive, and user-friendly, contributing to overall user satisfaction. The survey consisted of four questions that cover key aspects of the user experience :

- How satisfied were you with the accuracy of the legal information provided?
- How would you rate the ease of use of the legal assistant?

Legal Domain	Average Score (Out of 5)
Family Law	4.2
Property Law	4.5
Labor Law	4.3
Criminal Law	4.7

Fig. 7. Distribution of User Satisfaction scores

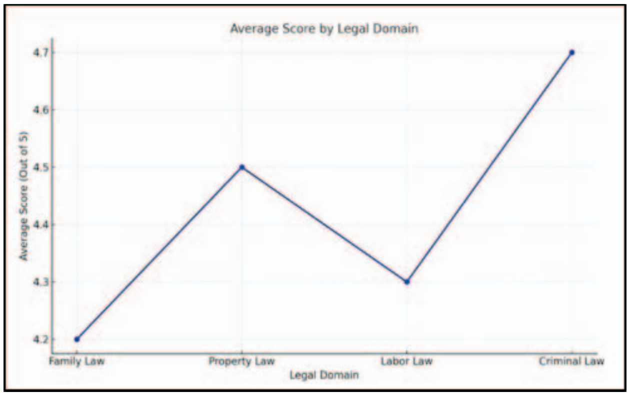

Fig. 8. User Satisfaction Level

- How satisfied were you with the response time of the legal assistant?
- Overall, how satisfied are you with the legal assistant?

For each question, we calculated the average score by summing all the ratings given by the respondents and then dividing by the total number of responses. This gave the mean satisfaction score for each question. The overall average satisfaction score was calculated by taking the mean of the average scores for each question.

Figure 7 showcases the distribution of user satisfaction ratings, highlighting the overall positive user feedback received.

Figure 8 showcases a graph of user satisfaction ratings, highlighting the overall positive user feedback received.

Figure 9 showcases the comparison of proposed solution and Traditional approach

This research contributed to the field of legal informatics by providing a scalable, efficient solution for legal assistance. As AI continues to evolve, future work will likely include the continuous improvement of the chatbot's capabilities, expansion of its legal ontology, and refinement of its multilingual functions

Parameter	Proposed Solution	Traditional Solution
Data Acquisition	Utilizes web scraping for efficient data extraction.	Relies on manual search methods and raw data.
Text Representation	Utilizes vector embeddings to represent legal text to capture semantic relationships between words.	Legal text is mostly represented in its raw format not taking word context in consideration.
Search Technique	Implements semantic searching using context of the query to search and match with database.	Uses a keyword-based search model, where keywords are extracted from text and then searched for in the database.
Efficiency	Can process complex user queries in quick time using vectors.	Takes longer time to search and map keywords from large user queries.
Accessibility	Resolves legal queries in a comprehensible manner, breaking down complex legal terms for common users.	The respective legal text/sections are returned as is, which is difficult for users to comprehend.

Fig. 9. Comparison of proposed solution and Traditional approach

CONCLUSION

The AI-based legal assistant chatbot developed in this study has demonstrated its potential to significantly enhance access to legal information within the Indian judicial system. The chatbot, powered by advanced NLP techniques and leveraging the robust OpenAI API, demonstrated high precision and recall values across various legal domains, indicating a strong ability to deliver accurate and relevant legal advice. User satisfaction scores were notably high, reflecting the system's ease of use and the quality of interaction experienced by the users.

The paper also navigated the challenges encountered during the development phase, such as data acquisition and processing.

REFERENCES

[1] V. Socatiyanurak *et al.*, "LAW-U: Legal Guidance Through Artificial Intelligence Chatbot for Sexual Violence Victims and Survivors," in *IEEE Access*, vol. 9, pp. 131440-131461, 2021, doi: 10.1109/ACCESS.2021.3113172.

[2] Yupeng Chang, Xu Wang, Jindong Wang, Yuan , Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi: A Survey on Evaluation of Large Language Model. arXiv 2023

[3] K. Nikolskaia and V. Naumov, "Artificial Intelligence in Law," *2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*, Vladivostok, Russia, 2020, pp. 1-4, doi: 10.1109/FarEastCon50210.2020.9271095.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need

[5] N. Jain and G. Goel, "An Approach to Get Legal Assistance Using Artificial Intelligence," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2020, pp. 768-771, doi: 10.1109/ICRITO48877.2020.9198029.

[6] Korawit Orkphol, WuYang: Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet.

[7] Marc Queudot, Éric Charton and Marie-Jean Meurs: Improving Access to Justice with Legal Chatbot.

[8] F. Heimerl and M. Gleicher: Interactive Analysis of Word Vector Embedding.

[9] Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding- based Word Similarities. Transactions of the Association for Computational Linguistics

[10] Liu, Y., He, H., et al. (2024). Efficient Training and Inference of Large Language Models: A Comprehensive Review. Journal of Machine Learning Research

[11] Lyons, M., & Brewer, G. (2021). Experiences of Intimate Partner Violence during Lockdown and the COVID-19 Pandemic. Journal of Family Violence

[12] Zheng, L., Chiang, W.-L., Sheng, Y., Li, T., et al. (2023). LMSYS-CHAT-1M: A Large-Scale Real-World LLM Conversation Dataset

[13] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective

[14] Li, S., Zhang, H., Ye, L., Guo, X., & Fang, B. (Year). MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction

[15] MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction