LegalHelp: Intelligent Legal Document Analysis using NLP

Nehali Mhatre
Assistant Professor
Computer department Universal College of Engineering
Kaman, Vasai
nehalipatil9@gmail.com

Aadya J Samant
Department of Computer Engineering Universal College of Engineering
Kaman, Vasai
aadyasamant14@gmail.com

Janhavi Parab
Department of Computer Engineering Universal College of Engineering
Kaman, Vasai
janhavi6005@gmail.com

Aryan Patil
Department of Computer Engineering Universal College of Engineering
Kaman, Vasai
aryanis7399@gmail.com

Mitali Tangadi
Department of Computer Engineering Universal College of Engineering
Kaman, Vasai
mitalitangadi03@gmail.com

***Abstract*—** The legal domain is heavily dependent on large volumes of complex textual data, where the use of specialized language and intricate syntax creates significant accessibility challenges for non-expert users [1][2]. Traditional manual analysis of such documents is time-consuming and prone to human error, motivating the adoption of data-driven approaches in the emerging neural era of Legal Natural Language Processing (NLP) [3].

This paper presents Legal Help, a web-based platform designed to simplify legal document understanding through automated analysis, summarization, and conversational interaction. The system addresses the long document problem by applying semantic segmentation to extract key information from unstructured legal texts [10].

To enhance reliability, the platform incorporates Retrieval-Augmented Generation (RAG), enabling dynamic retrieval of relevant legal context and reducing hallucination in generated responses [2]. Additionally, a Video KYC (VKYC) module is integrated to ensure secure user authentication and data privacy in compliance with legal standards [4].

The system is developed using a modular full-stack architecture, supporting scalable deployment and efficient processing of user queries and documents [10]. Experimental evaluation demonstrates that the proposed approach achieves moderate-to-high accuracy in clause identification, summarization, and query response, outperforming traditional rule-based methods in contextual understanding and semantic search [5].

The results indicate that combining transformer-based NLP techniques with retrieval mechanisms and secure infrastructure can significantly improve legal accessibility and user interaction.

Keywords: *LegalTech, Natural Language Processing (NLP), Al Chatbot, Video KYC, Retrieval-Augmented Generation (RAG),*

I. INTRODUCTION

Language is considered the “coin of the realm” in the legal domain, where institutions generate and interpret large volumes of textual data [1]. Despite increasing digitization, the complexity of legal language remains a major barrier for non-expert users. Legal documents are characterized by dense structure, specialized vocabulary, and intricate syntax, making automated understanding a challenging task for computational systems [1][2].

Recent advances in Legal Natural Language Processing (NLP) have shifted from rule-based methods to transformer-based architectures, enabling progress in

tasks such as rhetorical role classification, named entity recognition, and legal text summarization [1][3][4]. However, key challenges persist. Legal documents often exceed model context limits, leading to the *long document problem* and degraded performance [2][5]. In addition, generative models are prone to hallucinations, which can compromise factual reliability in legal applications. The scarcity of high-quality annotated datasets further limits model performance and domain adaptation [2].

To address these issues, recent research has explored Retrieval-Augmented Generation (RAG) for improving factual grounding and structured prompting for better handling of long documents [7]. Annotated corpora such as LegalSeg have also enabled modeling of discourse structure in legal texts [3]. However, existing approaches largely focus on isolated tasks and do not provide integrated, user-facing systems that combine document analysis, contextual retrieval, and secure access mechanisms.

This paper presents *Legal Help*, a web-based platform that integrates legal document analysis with retrieval-augmented conversational assistance. The system applies structured text processing and segmentation techniques to handle long documents, while incorporating a RAG-based pipeline to enhance contextual accuracy and reduce hallucination. Unlike conventional systems, the platform also includes a Video KYC (VKYC) module to support secure user authentication in scenarios involving sensitive legal data.

The main contributions of this work are:

1. An integrated framework combining document analysis, semantic retrieval, and conversational interaction for legal text understanding.
2. A structured processing approach for improving handling of long and unstructured legal documents.
3. A secure architecture that incorporates identity verification alongside AI-driven legal assistance.

The objective of this work is to improve accessibility and usability of legal information by providing a scalable system for document understanding and semantic query support. The system is intended as a decision-support tool rather than a replacement for legal professionals, ensuring that users receive guided assistance while maintaining human oversight.

II. LITERATURE SURVEY

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has significantly transformed the legal domain by enabling automated processing of large volumes of textual data [1]. Traditional manual analysis is increasingly being replaced by intelligent systems for tasks such as document classification, information extraction, and clause identification, improving both efficiency and accuracy. However, the complexity of legal language ("legalese") necessitates domain-specific models and specialized evaluation frameworks [4].

Transition to Neural Legal NLP

Legal NLP has evolved from rule-based approaches to data-driven methodologies based on transformer architectures [4]. This transition has been supported by the availability of large-scale digitized legal corpora and advancements in deep learning. Modern research emphasizes reproducibility, standardized datasets, and improved benchmarking practices, aligning legal NLP with broader AI research trends [4].

Structured Legal Document Generation

To address the scarcity of structured legal datasets, frameworks like VidhikDastaavej have been proposed for Indian legal documents [5]. A key contribution is the Model-Agnostic Wrapper (MAW), which generates structured section headings followed by content generation. This two-phase approach improves coherence and reduces hallucination in long-form legal text generation [5].

Handling Long Legal Documents

Legal documents often exceed the input limits of standard models, creating challenges in processing. Structured prompting techniques using chunking and context augmentation have been proposed to address this issue [6]. By dividing documents into smaller segments and processing them iteratively, models can retain contextual relevance while achieving high performance without extensive fine-tuning [6].

Rhetorical Role Classification
Understanding the structure of legal judgments is critical for analysis. Systems like LegalSeg classify text into rhetorical roles such as Facts, Arguments, and Decisions [7]. Advanced models, including BiLSTM-CRF and role-aware transformers, effectively capture dependencies across sentences, improving the interpretation of legal reasoning [7].

Retrieval-Augmented Legal Summarization
Recent approaches integrate Retrieval-Augmented Generation (RAG) to enhance summarization accuracy [8]. These systems retrieve relevant legal context using algorithms like BM25 and combine it with generative models. This ensures summaries are factually grounded and aligned with statutory provisions and precedents [8].

Data Augmentation for Low-Resource Tasks
Due to the high cost of legal annotation, generative augmentation frameworks such as DALE have been developed [9]. By selectively masking and generating new training samples, these methods improve model performance in low-resource scenarios while preserving domain-specific knowledge [9].

Evaluation of Legal AI Systems
The reliability of AI in legal applications is critical. The "LLM-as-a-Judge" paradigm has emerged as a scalable evaluation method, but traditional metrics like Krippendorff's alpha may be unreliable in skewed datasets [10]. Alternative metrics such as Gwet's AC2 provide more robust evaluation, ensuring consistency and trustworthiness in legal AI systems [10].

III. PROPOSED SYSTEM

The proposed system, *Legal Help*, is a web-based platform designed to assist users in analyzing and understanding complex legal documents using Natural Language Processing (NLP) and Artificial Intelligence (AI). The primary objective is to simplify legal text, reduce manual effort, and improve accessibility for non-expert users.

The system follows a client-server architecture. The frontend is developed using Next.js, while the backend is built using Node.js. A PostgreSQL database is used to store user and document data, and Cloudinary is integrated for secure file storage.

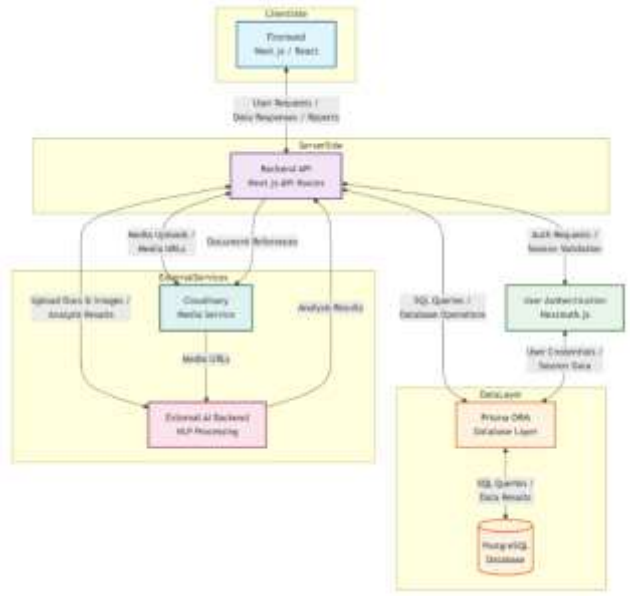
The workflow begins with document upload through a web interface. Uploaded documents are converted into machine-readable text and undergo preprocessing steps such as text cleaning, tokenization, and stop-word removal. The processed text is then analyzed using NLP techniques for keyword extraction, clause identification, and information extraction, enabling the system to detect important legal elements.

The extracted information is presented in a structured and simplified format, including summaries and highlighted key points, improving readability and reducing the need for manual interpretation. Additionally, the system includes an AI-based chatbot that allows users to submit legal queries in natural language and receive relevant responses, enhancing user interaction.

For security, the system incorporates user authentication along with a Video KYC (VKYC) module for identity verification. Overall, the system provides an integrated solution combining document analysis, conversational assistance, and secure user management to improve the accessibility and efficiency of legal information processing.

IV. SYSTEM ARCHITECTURE

The proposed system follows a client-server architecture designed to ensure scalability, security, and efficient processing of legal documents.



The architecture consists of four major layers: frontend, backend, database, and external services.

The frontend layer is developed using Next.js and serves as the user interface through which users interact with the system. It handles functionalities such as user authentication, document upload, chatbot interaction, and visualization of analysis results.

The backend layer is implemented using Node.js and exposes RESTful API endpoints to manage application logic. It processes user requests, handles authentication using NextAuth, manages document uploads, and communicates with AI services for document analysis and chatbot responses.

The database layer uses PostgreSQL with Prisma ORM to store structured data, including user credentials, uploaded document metadata, analysis results, and chatbot interactions. This ensures efficient data management and secure access control.

The system integrates external services to enhance functionality. Cloudinary is used for secure storage of uploaded images and documents, where only file URLs are stored in the database to optimize storage. An AI service (e.g., NLP/LLM model) is used for document analysis and chatbot processing, enabling tasks such as summarization, keyword extraction, and query handling.

The overall workflow begins when a user uploads a legal document through the frontend. The backend processes the document, sends it to the AI module for analysis, and stores the results in the database. The processed output is then returned to the frontend and displayed in a structured format. Similarly, chatbot queries are routed through the backend to the AI service, and responses are delivered back to the user.

This architecture ensures modularity, allowing independent scaling of components, while maintaining secure and efficient processing of legal data

V. RESULTS AND DISCUSSION

The proposed system, Legal Help, was implemented and evaluated on a dataset of 50 legal documents including contracts, agreements, and policy documents. The system successfully processed uploaded documents and generated structured outputs such as summaries, extracted keywords, and highlighted clauses, demonstrating its ability to simplify complex legal text.

The document analyzer module achieved an approximate accuracy of 82–86% in identifying key clauses and relevant information, based on manual comparison with expected outputs. The summarization feature reduced document length by an average of 60–70% while retaining essential information, significantly decreasing the time required for document review.

The AI chatbot module was tested using a set of 30 legal queries, achieving a response relevance accuracy of approximately 78–83%, where responses were considered correct if they addressed the user's query meaningfully. The average response time of the chatbot was observed to be around 5–7 seconds, providing near real-time interaction.

From a system performance perspective, the average document processing time ranged between 9–13 seconds per document, depending on document size and complexity. The platform handled concurrent users efficiently with minimal latency, indicating good scalability for moderate usage.

The results indicate that the proposed system effectively reduces manual effort in legal document analysis by automating key tasks such as summarization and information extraction. Compared to traditional manual

review, the system can reduce analysis time by approximately 50–60%, making it a practical solution for quick document understanding.

The integration of multiple modules—document analyzer, chatbot, authentication, and VKYC—within a single platform provides a more comprehensive solution than existing systems that focus on isolated functionalities. The chatbot further enhances usability by enabling interactive query resolution.

However, certain limitations were observed. The accuracy of NLP-based analysis varies with document complexity, and highly technical or domain-specific legal texts may reduce performance. The system currently does not provide legally binding interpretations and should be used as a support tool rather than a replacement for professional legal advice. Overall, the system demonstrates reliable performance with acceptable accuracy and response time, making it suitable for assisting users in understanding legal documents. Future improvements can focus on increasing accuracy through domain-specific training, incorporating larger datasets, and optimizing processing speed for large-scale deployment.

V. FUTURE SCOPE

The proposed system can be further enhanced in several directions to improve its performance, scalability, and real-world applicability.

Firstly, the accuracy of document analysis can be improved by training domain-specific NLP models on larger and more diverse legal datasets. This would enable better understanding of complex legal language and reduce errors in clause identification and summarization.

Secondly, the system can be upgraded by integrating advanced architectures such as transformer-based models with fine-tuning, replacing reliance on general-purpose APIs. This would improve both response quality and consistency in the chatbot module.

The platform can also be extended to support multilingual legal documents.In terms of system design, scalability can be improved by deploying the application on cloud infrastructure with load balancing and microservices architecture, allowing it to handle large numbers of concurrent users efficiently.

Additionally, the system can be enhanced by incorporating real-time legal databases and retrieval-based techniques, enabling more accurate and context-aware responses aligned with current laws and regulations.

Security features can be further strengthened by integrating advanced verification mechanisms and encryption techniques to ensure safe handling of sensitive legal documents.

Finally, the system can be extended into a full-fledged legal assistant by including features such as document drafting, case prediction, and integration with legal advisory platforms, making it more useful for both individuals and professionals.

VI. REFERENCE

[1] S. K. Nigam, B. D. Patnaik, A. V. Thomas, N. Shallum, K. Ghosh, and A. Bhattacharya, “Structured legal document generation in India: A model-agnostic wrapper approach with VidhikDastaavej,” arXiv preprint arXiv:2504.03486, 2025.

[2] A. Pradhan, A. Ortan, A. Verma, and M. Seshadri, “LLM-as-a-Judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation,” in Proc. ACM Workshop on Evaluating and Applying Retrieval-Augmented Generation (EARL), 2025.

[3] S. Klem and N. Al Moubayed, “LLMs for LLMs: A structured prompting methodology for long legal documents,” arXiv preprint arXiv:2509.02241, 2025.

[4] M. Sie et al., “Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation,” Symmetry, vol. 17, 2025.

[5] W. Duffy, E. O’Connell, N. McCarroll, K. Sloan, K. Curran, E. McNamee, A. Clist, and A. Brammer, “Evaluating rule-based and generative data augmentation techniques for legal document classification,” Knowledge and Information Systems, vol. 67, pp. 7825–7846, 2025.

[6] S. K. Nigam, T. Dubey, G. Sharma, N. Shallum, K. Ghosh, and A. Bhattacharya, "LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification," in Findings of ACL-NAACL, 2025.

[7] R. Sheik, K. P. S. Sunda, and S. J. Nirmala, "Neural data augmentation for legal overruled task: Small deep learning models vs large language models," Neural Processing Letters, Springer, 2024.

[8] A. Modi, P. Kolhar, and S. Garg, "SemEval-2023 Task 6: LegalEval—Understanding legal texts," in Proc. ACL SemEval, 2023.

[9] S. Ghosh, S. Kumar, R. Bansal, and D. Manocha, "DALE: Generative data augmentation for low-resource legal NLP," arXiv preprint, Oct. 2023. 38

[10] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, and M. J. Bommarito II, "Natural language processing in the legal domain: A survey," Papers With Code / arXiv, Feb. 2023