Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

AI-Powered Pipeline for Document Simplification and Automated Drafting

Shubh Gupta
Department of Computer Science and Engineering (Data Science)
Moradabad Institute of Technology, Moradabad

Anchit Gupta
Department of Computer Science and Engineering (Data Science)
Moradabad Institute of Technology, Moradabad

Aviral Gupta
Department of Computer Science and Engineering (Data Science)
Moradabad Institute of Technology, Moradabad

Gauri Agarwal
Department of Computer Science and Engineering (Data Science)
Moradabad Institute of Technology, Moradabad

Dr. Saurabh Srivastava
Department of Computer Science and Engineering (Data Science)
Moradabad Institute of Technology, Moradabad

Abstract - Legal documents, such as contracts, agreements, notices, and policies, often contain complex legal jargon and dense clause structures. This makes them hard for non-experts to understand. As a result, individuals, businesses, and startups must depend on legal professionals for even simple tasks. This reliance leads to higher costs, delays, and possible misunderstandings. LegalEase aims to fill this gap by providing an AI-powered mobile app that turns legal documents into easy-to-understand, plain-language summaries. It also offers a user-friendly way to create legally structured documents on demand. LegalEase accepts various input formats, including PDF, DOCX, plain text, and even photos or scanned images. It uses OCR and NLP technologies to extract text. Once the text is extracted, the system identifies clauses, highlights complex legal terms, and translates them into simple language. Users receive clearly organized key points and summaries. In addition to simplification, LegalEase lets users create new legal documents, such as rental agreements, contracts, and NDAs. They can do this by submitting prompts, selecting templates, and giving custom details. The app also has a conversational AI assistant that allows users to ask questions about the contents of legal documents. They can inquire whether certain clauses exist, if specific terms are fair, or request explanations of unclear sections. The main innovation of LegalEase is its combined chat-style mobile interface. This platform includes document upload, legal-text simplification, clause summarization, document generation, and interactive support for legal questions — all in one place. By using OCR, clause detection, summarization, document drafting, and clause checking, LegalEase seeks to make legal documents understandable and accessible to everyday users without formal legal training. This approach could greatly reduce the need for legal professionals for standard tasks, improve contract transparency, and broaden access to legal services.

Keywords — Natural language processing (NLP), Optical character recognition (OCR), Clause extraction, Document generation, Conversational AI, Legal-tech, Generative AI.

1. Introduction

Legal documents such as contracts, agreements, policies and compliance papers often contain technical language that many people find hard to understand. This challenge can lead to misunderstandings, missed obligations and a greater chance of accepting terms without grasping their consequences. As legal paperwork becomes more embedded in tasks—from renting properties to consenting to online services—the need for clear and understandable legal language is more crucial, than ever. Though digital tools are accessible a significant issue remains: people without legal knowledge struggle to understand clauses check essential terms or identify unfair or vague conditions. Current options offer help. Some offer document summaries others provide templates and a few use OCR to examine documents. However these tools do not combine, lack depth. Do not offer interactive support for customized questions, about document specifics. The issue outlined here reflects the challenge this research aims to solve: Users need an all-in-one platform that can simplify complex legal language, highlight and condense key information ensure clause accuracy and help with personal legal questions all while managing various document types.

To address this the project suggests creating an AI-powered application that combines text simplification, clause-focused summarization, automated document creation, support, for multiple document formats (PDF,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

DOCX, images) and an interactive chat assistant. Integrating OCR with Natural Language Processing (NLP) models the system aims to convert legal texts into clear simple language allowing users to quickly spot essential missing or possibly unfair clauses. Additionally the mobile app includes a chat-style interface that makes it easy for users to engage with the system. This aids navigation through tasks such, as uploading files reviewing clauses and drafting documents based on user inputs. This unified approach not only improves accessibility but also shows how AI can connect legal expertise with everyday understanding. Overall, this research adds to the expanding field of legal technology by offering a complete framework designed to enhance transparency, understanding, and user empowerment when dealing with legal documents

2. THEORETICAL BACKGROUND

A. Natural Language Processing (NLP)

Natural Language Processing is a branch of intelligence that allows machines to comprehend, analyze and produce human language. Within this scope NLP is utilized to examine documents recognize clauses pinpoint important terms and streamline sentence constructions.

B. Optical Character Recognition (OCR)

OCR is a technology that transforms text found in images scanned files or photographs into text that machines can read. This allows the system to handle documents even if they are not, in formats that can be edited (such as photos of contracts or scanned PDFs).

C. Clause Extraction

Clause extraction involves pinpointing and isolating legal provisions within a document. This enables users to grasp the intent of each provision examine duties and confirm the inclusion of legal terms.

D. Document Generation

Document generation involves the creation of legally formatted documents—like contracts, agreements and notices—using user inputs or predefined templates. This minimizes manual. Guarantees uniformity and accuracy.

E. Conversational AI

Conversational AI lets users communicate with the system using natural language queries like having a conversation with an assistant. It enables users to inquire about clauses seek clarification, on meanings or ask for help regarding information.

F. Legal-Tech

Legal-tech encompasses technology-based advancements designed to enhance procedures boost efficiency and broaden access, to legal services. LegalEase operates in this field by utilizing AI to streamline and produce documents.

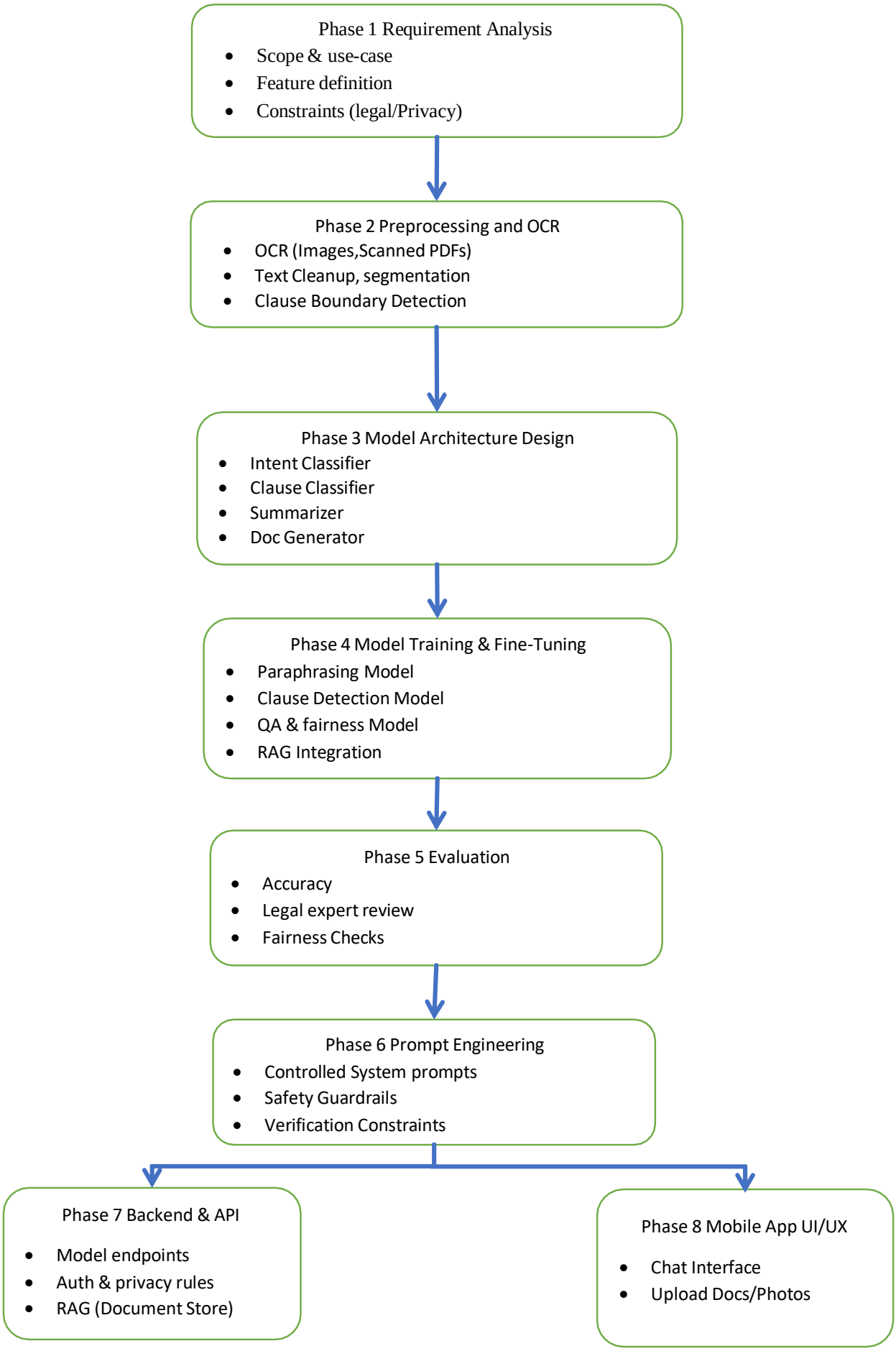
G. Generative AI

Generative AI represents a branch of intelligence that can produce original content—like text, documents or summaries—derived from user prompts. Within this project it serves to compose documents and make current ones easier to understand.

3. MODEL DEVELOPMENT

The creation of the model, for this project adhered to an methodical approach aimed at guaranteeing precision, dependability and practical relevance. Every stage was meticulously designed to tackle an aspect of the system—spanning from requirement analysis and data preprocessing to model training, assessment and implementation. By integrating OCR, natural language processing, retrieval mechanisms, and a conversational interface, the development process aimed to create an end-to-end solution capable of simplifying legal documents, identifying clauses, generating structured drafts, and responding intelligently to user queries. This section outlines the complete development workflow and the rationale behind each phase.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Phase 0: Requirement Analysis

The project commenced with an in-depth requirement analysis to thoroughly grasp the objective, boundaries and user demands for the system. This stage entailed pinpointing the issues users encounter when dealing with legal documents, including challenges in comprehending complex legal terminology difficulties, in validating key clauses and absence of user-friendly tools for document creation. Functional specifications were established to cover OCR processing, simplification of text, identification of clauses question answering and document generation driven by prompts. Non-functional criteria including precision, confidentiality, dependability and mobile friendliness were also defined. Moreover legal and ethical factors were assessed to guarantee that the system safeguards user information refrains from offering counsel and adheres to applicable privacy laws.

Phase 1: Data Collection

This stage concentrated on collecting the datasets required for training and assessing the AI models. A diverse selection of documents—contracts, agreements, NDAs, rental papers and policies—was gathered to reflect authentic real-life situations. Clause-level labels were created to mark boundaries and types of clauses including payment terms, confidentiality and termination. Datasets, with versions featuring plain-language paraphrases of legal clauses were assembled to train the summarization system. Samples of OCR data consisting of scanned documents and images were gathered well to enhance the effectiveness of text extraction. The quality and variety of this collection are essential, for guaranteeing model training.

Phase 2: Preprocessing & OCR

After gathering the data, preprocessing and OCR processes were carried out to transform documents into clean, uniform machine-readable text. OCR software such, as Tesseract or EasyOCR was employed to extract text from images and scanned PDFs, followed by noise elimination, alignment adjustment and line segmentation. Text normalization methods were utilized to eliminate characters, correct inconsistent spacing and standardize formatting. Subsequently the documents were divided into paragraphs and clauses with clause boundaries automatically identified using rules and patterns. Effective preprocessing guaranteed that the models that followed were provided with input data.

Phase 3: Model Architecture Design

During this stage the complete AI system structure was created by specifying the functions of each model element. The framework featured an intent classifier tasked with detecting user intentions, like summarization or clause verification. A clause classifier was developed to assign clauses to specific categories and a summarization/paraphrasing model transformed legal text into straightforward language. Additionally a document generator that could create drafts, from user instructions was included. Retrieval-Augmented Generation (RAG) was incorporated to improve precision by fetching legal templates or previously cataloged clauses. This design guaranteed scalability and effective task distribution.

Phase 4: Model Training & Fine-Tuning

The gathered and processed datasets served to train and refine the AI models. The paraphrasing model underwent training on to-plain-language pairs to produce precise and legally accurate simplifications. The clause detection model was developed to recognize the existence, category and limits of clauses inside a document. The question-answering and fairness assessment model was fine-tuned to understand user inquiries, about clause requiredness or fairness. The RAG pipeline was further enhanced by creating embeddings cataloging cases and improving retrieval effectiveness. These adjustments guaranteed that the system delivered appropriate and precise results.

Phase 5: Evaluation

Following training the system was rigorously assessed to evaluate its effectiveness, dependability and compliance with standards. Automated assessments utilized accuracy rates, precision-recall statistics, F1-scores for categorization and readability measures, for summarization. Additionally human evaluations were carried

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

out with legal specialists examining clauses key-point summaries and created documents to confirm accuracy and semantic integrity. Safety verifications guaranteed that the system did not fabricate information exclude clauses or generate deceptive interpretations. This stage confirmed if the system fulfilled the anticipated quality criteria.

Phase 6: Prompt Engineering

Prompt engineering was carried out to enhance the models behavior during interactions. Crafted system prompts directed the model's tone organization and consistency of outputs. Task-tailored prompts guaranteed peak effectiveness in summarization, document creation, clause interpretation and answering questions. Safety measures were integrated via prompts to avoid legal advice. Output verification rules were incorporated into prompts to maintain formatting, thoroughness and accuracy. This step improved system. Minimized erratic model responses.

Phase 7: Backend & API Development

The backend architecture was created to unify all AI elements into one system. Models were hosted on protected servers. Made available via API endpoints that the mobile application could access. Security measures were put in place to safeguard user information and guarantee entry. A vector database was established to facilitate RAG functions allowing retrieval of legal clauses. The backend managed document storage, request handling and inter-module communication. This maintained performance, minimal delay and secure functionality.

Phase 8: Mobile App UI/UX

The mobile app interface was crafted to offer users an practical experience. A chat-oriented UI was developed to facilitate communication with AI models supporting functions, like uploading files summarizing text posing clause-specific queries and creating legal drafts. The design focused on clarity, ease of use and seamless navigation. Dedicated screens were designed to show extracted clauses, simplified summaries, main points and produced documents. This stage guaranteed that users were able to engage with AI systems via a straightforward and easy-, to-use interface.

Phase 9: End-to-End Testing

Extensive testing was carried out to confirm that all parts operated smoothly when combined. Functional assessments confirmed that document uploading, OCR, summarizing, clause identification and user query processing performed correctly. Stress testing evaluated performance under loads or extensive documents. User acceptance testing (UAT) involved sample users to collect feedback on ease of use and comprehensibility. Detected problems, like misclassification, interface bugs or slow responses were resolved through repeated fixes. This stage guaranteed system reliability and preparedness for launch.

Phase 10: Deployment & Monitoring

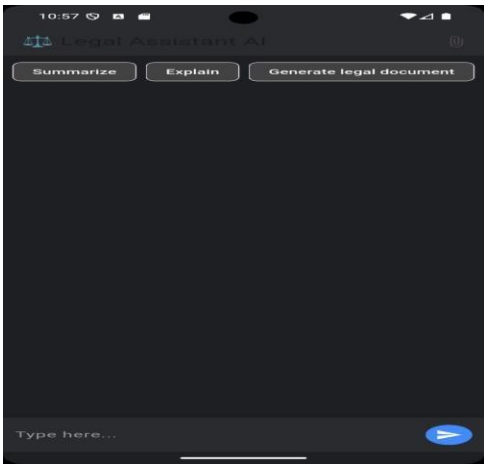
The system was launched in a production setting allowing actual users to engage with it. Ongoing monitoring solutions observed model effectiveness, system stability, response times and user activities. Techniques for detecting model drift were applied to spot declines, in performance as time progressed. Logs and user feedback were reviewed to plan retraining intervals and enhance models. Consistent updates guaranteed the system stayed precise, scalable and responsive.

Phase 11: Legal & Compliance

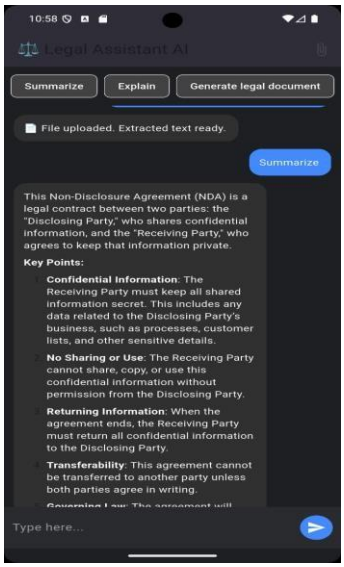
Finally, legal and compliance measures were incorporated to ensure safe operation. Data privacy policies were implemented to define how user data is collected, stored, processed, and deleted. Compliance with relevant data protection laws was maintained. The application included clear legal disclaimers stating that the system provides informational assistance rather than professional legal advice. Encryption and secure access protocols protected sensitive documents throughout the workflow. This phase ensured ethical, transparent, and legally responsible system use.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

4. RESULT & DISCUSSION



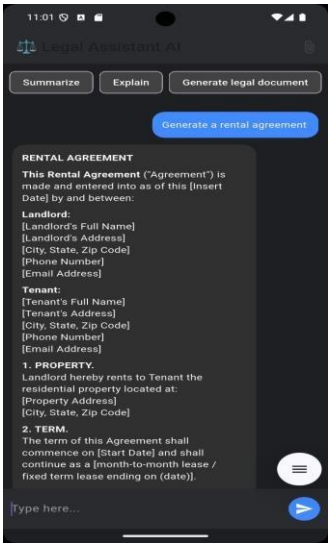
The picture shows the part of the LegalEase mobile application where you can have a conversation. This application is meant to give people help in a way that is easy to understand. The LegalEase mobile application has a background and a simple design which makes it easy, on the eyes and looks professional for talking about legal things. At the top of the interface you will see three buttons that say Summarize, Explain and Generate Legal Document. These buttons help users find what they need to do fast without having to look around a lot. The main part of the interface is simple and easy to look at so users can just focus on what they're doing. There is a text box, at the bottom where users can talk to the system using language like they were chatting with someone. This makes it feel like talking to a person like those AI chat things that are popular now. The system is designed to be easy to use so users can just type what they want in the text box. Get started with Summarize Explain and Generate Legal Document tasks. This design allows the system's backend to dynamically interpret user intent and execute the appropriate legal processing task. Overall, the interface effectively abstracts sophisticated AI-driven legal analysis into a user-friendly mobile experience, aligning with the project's objective of making legal document understanding and generation more accessible to non-technical users.



The picture shows how the LegalEase mobile application works when you upload a document and get it ready to use. Here someone is uploading a Non-Disclosure Agreement (NDA) through the part of the app where you can talk to it.The LegalEase mobile application then shows a picture of the file right in the conversation window. This makes it clear that the document was uploaded correctly.After that the LegalEase mobile application sends

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

a message saying "File uploaded. Extracted text ready."The Optical Character Recognition module and the text preprocessing pipeline have done their job. This means the Optical Character Recognition module and the text preprocessing pipeline can take picture documents and turn them into text that computers can understand. The Optical Character Recognition module and the text preprocessing pipeline make sure the parts of the documents like clauses and sections are still, in the order.The application is easy to use because it lets people upload documents look at them and get feedback from the system in one place. This makes it simple for people to work with the Optical Character Recognition module and the text preprocessing pipeline. The Optical Character Recognition module and the text preprocessing pipeline are important for this to work. This functionality demonstrates the system's capability to handle non-editable legal document formats and serves as a crucial foundation for subsequent operations such as summarization, explanation, clause analysis, and legal query resolution, thereby validating the practical applicability of the proposed LegalTech solution in real-world scenarios.



The picture shows what you see when you first open the LegalEase application. This is what the LegalEase mobile application looks like at the beginning. The LegalEase mobile application has an dark look that makes it easy to read and looks professional, which is what you want for something related to the law. At the top of the screen you can see the name of the application, which's Legal Assistant AI and a picture of a legal symbol. The Legal Assistant AI title and the legal symbol at the top tell you what the LegalEase mobile application is for and what it is, about. The website has three buttons that you can see right away. These buttons are called Summarize, Explain and Generate Legal Document. They help you do what you want to do with the law stuff without having to look through a lot of menus. This makes it easy for you to use the website because it shows you what to do and how to do it. The middle part of the website is simple, on purpose. This is so you can focus on what you're doing. At the bottom there is a box where you can type what you want to say. You can type what you mean in your words. The buttons are there to help you with things like Summarize, Explain and Generate Legal Document. This chat-based interaction model enables flexible communication, where users can either rely on predefined actions or describe their requirements freely. Overall, the interface effectively abstracts complex AI- driven legal processing into a simple and intuitive mobile experience, aligning with the research objective of making legal document analysis and generation accessible to non-technical users.

5. CONCLUSION

The people behind LegalEase made an app that uses artificial intelligence to make legal documents easier to understand. This app can create drafts of legal documents and help people get answers to their legal questions in a conversation. The app uses tools like Optical Character Recognition and Natural Language Processing to find important parts of legal documents and simplify the language. It can even summarize documents and generate new text. The goal of LegalEase is to help people understand language that is often too complicated for everyday people to comprehend. LegalEase is trying to bridge the gap, between legal language and what people can easily understand. The way we develop things has steps. We start by looking at what the system needs to do. Then we get everything ready. After that we train the model. See how well it works. We also make sure the system

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

understands what people are asking for. Finally we put the system on devices. This approach makes sure the system is good from a standpoint and also easy for people to use. The development process is very important for the system. The system has to be good, at doing its job. People have to be able to use it easily.The way we set up the experiment shows that we can take documents and turn them into simple summaries that people can understand. We do not change what the law says. We look at each part of the document. Use special tools to help answer questions that people ask. For example we can check if a certain part is, in the document find parts that might not be fair and see if the document has everything it needs. We also have a way for people to talk to the system on their phones. They can just type what they want to do like they were talking to someone and the system will help them with their tasks. This way of doing things makes legal documents easier to understand and use for everyone. It also reduces the amount of thinking and money that people usually have to spend when they try to figure out what legal documents mean. Legal documents can be really hard to understand so this approach helps with that. It helps make legal document interpretation easier and cheaper which is a deal, for legal documents.

REFERENCES

[1] “Enhancing Legal Document Summarization for Professionals: An Extractive Approach,” in IEEE Conference Publication, 2025.

[2] “Deep Learning Techniques for Legal Text Summarization,” in IEEE Conference Publication, 2025.

[3] M. Akter, E. Çano, E. Weber, D. Dobler, and I. Habernal, “A Comprehensive Survey on Legal Summarization: Challenges and Future Directions,” arXiv preprint, Jan. 2025.

[4] Proceedings of the Natural Legal Language Processing Workshop 2024, Association for Computational Linguistics, Nov. 2024.

[5] A. Hosabettu and H. Shah, “Transformer-Based Extraction of Statutory Definitions from the U.S. Code,” arXiv preprint, Apr. 2025.

[6] M. S. Anbarasi, A. Mohammed A., D. R., M. V. and M. S., “Abstractive Summarization of Indian Legal Documents Using T5 & QLoRA,” Int. Educ. Res. J., 2024.

[7] V. Suryavanshi and D. Naikwadi, “Legal Case Document Summarization Using AI,” Int. J. Eng. Techniques, vol. 10, no. 2, Mar. 2024.

[8] “Effectiveness in Retrieving Legal Precedents: Exploring Text Summarization and Language Models,” Artificial Intelligence and Law, Feb. 2025.

[9] S. Chakraborty, A. Gupta, and P. Sharma, “Legal Document Classification and Retrieval Using BERT Variants,” in Proc. IEEE Int. Conf. on Data Science & Engineering (ICDSE), 2024.

[10] N. Dahal and W. K. Ng, “Transformer Models for Legal Text Analysis: An Evaluation on Indian Legal Corpora,” IEEE Access, vol. 12, pp. 54023–54037, 2024.

[11] A. V. M. Rao and R. R. Chand, “Automated Clause Extraction from Legal Contracts Using Deep Learning Models,” IEEE Trans. on Neural Networks and Learning Systems, vol. 35, no. 4, pp. 2890–2902, 2025.

[12] M. D. Nguyen, S. Wang, and J. Lee, “Legal Question Answering with Contextual Language Models,” in Proc. IEEE Int. Conf. on Big Data (BigData), 2023.

[13] R. S. Patel and K. R. Shah, “Summarization Techniques for Legal Case Documents: A Comparative Study,” IEEE International Conference on Knowledge Engineering, 2025.

[14] R. Joshi and P. Singh, “AI-Based Legal Contract Assistant: Clause Identification and Risk Flagging,” IEEE Consumer Electronics Magazine, vol. 14, no. 3, pp. 58–66, 2025.

[15] B. Kumar and T. Reddy, “Neural Networks for OCR-Assisted Legal Document Text Extraction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 1, pp. 77-90, 2025.