ISSN: **3107-6513**

Demystifying Legal Documents Using Artificial Intelligence

Plain-Language Summarization and Document-Grounded Question Answering

Shravan R[#1], Srihari S[#2], Chandan U Achar[#3], Shreyas V H[#4], prof. Komala Devi[#5]
[#1,#2,#3,#4]*Students, Department of Information Science and Engineering*
CMR Institute of Technology, Bengaluru, Karnataka, India
[#5]*Assistant Professor , Department of Information Science and Engineering*
CMR Institute of Technology, Bengaluru, Karnataka, India
[1]Shr23ise@cmrit.ac.in, [2]srs23ise@cmrit.ac.in, [3]chac23ise@cmrit.ac.in , [4]shvh23ise@cmrit.ac.in, [5]komala.k@cmrit.ac.in

Abstract

Legal documents such as contracts, policies, notices, service agreements and court-related records routinely contain domain-specific terminology, deeply nested clauses and unusually long sentences that are difficult for ordinary users to interpret. This paper presents an artificial intelligence based system designed to transform complex legal content into concise, plain-language explanations. The proposed system accepts legal text or portable document format files, extracts machine-readable text, preprocesses and segments long documents, and applies transformer-based natural language processing for legal document summarization and simplification. A retrieval-supported question-answering component allows a user to obtain document-grounded explanations of specific clauses rather than generic legal commentary. The system design uses the Longformer Encoder-Decoder transformer for long documents, together with the legal-domain language model InLegalBERT for downstream legal analysis tasks. The paper describes the complete architecture, the preprocessing and segmentation strategy, the summarization and retrieval pipelines, and a structured evaluation plan covering readability, factual consistency and user usefulness. The objective is to improve accessibility and understanding of legal information while clearly positioning the system as an assistance tool and not a replacement for qualified legal advice.

Keywords—legal document simplification; artificial intelligence; natural language processing; Longformer Encoder-Decoder; InLegalBERT; retrieval-augmented generation; legal text summarization; explainable artificial intelligence

I. INTRODUCTION

A. Motivation

The legal system plays an important role in maintaining justice and enforcing laws. However, legal documents such as policies, contracts and court decisions often contain complex terminology and information that is difficult for the general public to understand. This creates a gap between legal professionals and ordinary users who need to understand their rights, duties and legal procedures.

This gap has practical consequences. Users routinely accept rental agreements, insurance policies, loan terms and digital service contracts without understanding the obligations they are agreeing to. Professional legal interpretation is available, but cost, availability and turnaround time make it impractical for the everyday documents that most people encounter. An automated assistance layer that explains a document in plain language, while remaining grounded in the actual text of that document, addresses a real and widespread need.

Recent developments in artificial intelligence (AI) and natural language processing (NLP) provide an opportunity to make legal information more accessible. AI-based legal systems can analyze complex legal text and transform important information into simpler and more understandable forms. The proposed project focuses primarily on helping users understand legal documents through document extraction, preprocessing, summarization, simplification and document-grounded explanations.

B. Problem Statement

Given an arbitrary legal document supplied as raw text or as a portable document format (PDF) file, the system must produce a faithful, readable explanation of its contents without requiring the user to have legal training. Three properties must hold simultaneously. The explanation must be shorter than the source, it must remain factually consistent with the source, and it must be traceable, so that any statement in the output can be linked back to the region of the document that produced it. Optimizing for brevity alone risks discarding obligations; optimizing for fidelity alone reproduces the original complexity.

C. Contributions

The main contributions of this work are as follows.

1) End-to-end pipeline: a complete workflow spanning PDF extraction, normalization, segmentation, abstractive summarization and document-grounded question answering, rather than a single isolated model.

2) Long-document handling: a segmentation and aggregation strategy that allows documents exceeding conventional transformer input limits to be processed without silently truncating clauses.

3) Traceable explanations: a retrieval-supported answering component that ties every generated response to retrieved source segments, improving transparency over free-form generation.

4) Structured evaluation plan: a multi-dimensional evaluation covering readability, factual consistency,

relevance and human usefulness, together with the reasoning for why automatic metrics alone are insufficient in the legal domain.

The remainder of this paper is organized as follows. Section II surveys related work and identifies the research gap. Section III presents the proposed system architecture. Section IV describes the implementation and methodology in detail. Section V defines the evaluation plan and expected outcomes. Section VI discusses limitations, and Sections VII and VIII conclude the paper and outline future scope.

II. LITERATURE REVIEW

A. Legal Text Summarization

Legal text summarization focuses on identifying the most relevant information from lengthy legal documents. Earlier approaches commonly used statistical and rule-based methods, while recent research increasingly uses deep learning and transformer-based models. Important NLP operations for legal document analysis include tokenization, named entity identification and dependency parsing.

Long legal documents create an additional challenge because they may exceed the input limits of conventional models. Recursive and segmented approaches process a long document in smaller sections and then combine the partial summaries. Extractive and abstractive strategies have also been used to identify salient legal information and generate concise explanations. Jain, Borah and Biswas [2] demonstrate an extract-then-assign strategy for building abstractive datasets over lengthy legal documents, while Sharma and Singh [8] show that a recursive summarization algorithm can preserve structure in documents that exceed model capacity in a single pass.

B. Long-Document Transformer Models

The Longformer Encoder-Decoder (LED) model is designed for sequence-to-sequence tasks involving long input text. Its architecture uses a combination of local sliding-window attention and task-specific global attention to process lengthy documents more efficiently than conventional full self-attention approaches. Because the cost of full self-attention grows quadratically with sequence length, this design is what makes multi-thousand-token legal documents tractable at all. This makes it suitable for legal document summarization, where contextual information must be preserved across many pages and where a definition introduced on the first page may govern a clause on the twentieth.

C. Legal-Domain Language Models

Legal-domain models such as InLegalBERT are trained to capture legal vocabulary and contextual relationships. Such models can support legal classification and analysis tasks. InLegalBERT can process tokenized legal text and produce representations useful for downstream tasks such as case-related classification. For lengthy documents, chunking can be used to divide the text into manageable segments. Domain-adapted encoders are relevant here because general-purpose models frequently mishandle terms whose legal meaning diverges sharply from ordinary usage, such as consideration, party or without prejudice.

D. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) improves factual and domain relevance by retrieving useful source content before generating an answer. Instead of relying only on parameters learned during training, the generator is conditioned on passages retrieved from the document at query time. Mukund and Easwarakumar [7] apply dynamic retrieval-augmented generation with domain-specific adaptation to legal summarization and report improved factual grounding. For the present system this property is essential, because an explanation that is fluent but not anchored in the uploaded document is worse than no explanation at all.

E. Comparative Analysis of Existing Work

Table I summarizes representative approaches and the limitation that motivates the proposed design.

TABLE I. COMPARISON OF EXISTING APPROACHES

Work	Approach	Limitation
[1]	LegalBERT embeddings with AdaBoost for case outcome prediction	Prediction-oriented; produces no user-facing explanation
[2]	Extract-then-assign abstractive dataset building	Dataset construction focus; no interactive querying
[5]	General NLP analysis of legal text	Limited handling of documents beyond model input limits
[7]	Dynamic RAG with domain adaptation	Targets summarization quality, not end-user readability
[8]	Recursive summarization algorithm	No clause-level traceability to the source document

F. Research Gap

Existing work concentrates either on predictive legal tasks or on summarization quality measured by automatic overlap metrics. Comparatively little attention is given to the non-expert reader, to whether the produced text is actually easier to understand, and to whether each claim can be traced back to a clause in the uploaded document. The proposed system targets this gap by combining long-document summarization, plain-language simplification and retrieval-grounded question answering within a single application workflow.

III. PROPOSED SYSTEM ARCHITECTURE

A. System Overview

The proposed system follows a modular architecture consisting of a user interface, a document-processing layer, an AI and NLP layer and an output layer, as shown in Fig. 1. Users can upload a PDF or provide legal text. The document-processing module extracts text and converts it into a machine-readable format. The extracted content is cleaned, normalized and divided into manageable segments.

The AI and NLP layer performs document summarization and simplification. A long-document transformer such as LED processes the segmented legal text and generates an abstractive summary. A retrieval-supported question-answering component identifies relevant sections of the uploaded document and provides explanations based on that context. The output layer presents the simplified content in a user-friendly form alongside the original text, so the two can be compared directly.

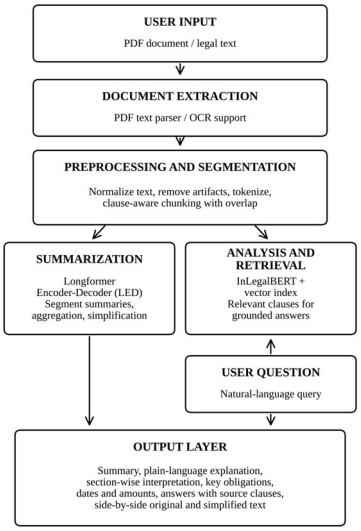

Fig. 1. Proposed system architecture.

Modularity is a deliberate design decision. Each layer communicates through a defined interface, so the summarization model can be replaced, or an additional legal-domain analyzer inserted, without altering extraction or presentation logic.

B. Legal Document Processing Flow

The workflow begins with text input or PDF upload. The PDF text is extracted and preprocessed by removing unnecessary whitespace and formatting artifacts such as headers, footers and page numbers. The text is then tokenized and segmented according to the capacity of the selected model. The summarization model generates a concise representation of the original content, which is postprocessed into a readable plain-language explanation. Segment boundaries and offsets are retained throughout, which is what later allows an answer to be mapped back to its source clause.

C. Key Output Features

The proposed system can provide a short summary, a detailed explanation, section-wise interpretation, key obligations, important dates or amounts and answers to user questions. The original and simplified text can be presented together to improve transparency and allow the user to compare the explanation with the source content. Each generated section carries a reference to the segment it was derived from, so a user who doubts an explanation can inspect the underlying clause immediately.

D. Technology Stack

Table II lists the components selected for each layer of the system and the role each plays.

TABLE II. SYSTEM TECHNOLOGY STACK

Layer	Component	Role
Interface	Web application	Upload, query and result display
Extraction	PDF text parser	Convert PDF to machine-readable text
Preprocessing	NLP tokenizer	Normalization, tokenization, segmentation
Summarization	LED transformer	Abstractive long-document summarization
Analysis	InLegalBERT	Legal-domain representation and classification
Retrieval	Vector index	Segment retrieval for grounded answering

E. Data Flow Design

Data moves through the system in a single forward direction with one feedback path. Raw input becomes normalized text, normalized text becomes indexed segments, and indexed segments feed both the summarizer and the retriever. The feedback path arises only at query time: a user question re-enters the retriever, selects a subset of the existing segments, and conditions the answer generator. No user question modifies the stored document representation, which keeps repeated queries deterministic with respect to the uploaded file.

IV. IMPLEMENTATION AND METHODOLOGY

A. Input Handling and Preprocessing

The application accepts text and PDF documents. For PDF input, the document text is extracted and converted into machine-readable form. Preprocessing includes normalization of whitespace, removal of unnecessary artifacts and segmentation of lengthy content. Because legal documents can be very long, segmentation is important for processing the content within transformer model limits. Ligature and encoding artifacts common in scanned or exported legal PDFs are corrected at this stage, since they otherwise propagate into the summary as malformed tokens.

B. Segmentation Strategy

Naive fixed-length splitting is avoided because it frequently divides a clause in the middle, leaving each half semantically incomplete. Instead, segmentation prefers structural boundaries in the following order of priority: numbered clause markers, section headings, paragraph breaks, and only then sentence boundaries. A bounded overlap is maintained between consecutive segments so that a definition appearing at the end of one segment remains visible at the start of the next. The overlap is expressed as a fraction of the segment length and is the main tunable parameter of this stage.

The number of segments produced for a document of length L with segment size S and overlap ratio r is approximated as

$$N = L / (S \times (1 - r)), \text{ rounded up} \quad (1)$$

Increasing the overlap ratio improves context continuity but raises the number of forward passes and therefore the processing time. In practice a modest overlap gives most of the benefit at acceptable cost.

C. Document Summarization Using LED

The summarization module is based on a transformer sequence-to-sequence approach. The LED architecture is suitable for long text because its encoder can efficiently handle longer sequences. The document is tokenized, passed through the encoder and decoded into an abstractive summary. The output can be controlled according to the desired summary length, such as short, medium or detailed.

During decoding, the model generates output token by token based on the input document and previously generated tokens. Postprocessing removes unwanted special tokens and improves readability. The resulting summary is intended to

preserve the important meaning of the legal document while reducing complexity. When a document has been split into multiple segments, per-segment summaries are produced first and then aggregated into a single coherent explanation, with duplicate statements merged.

D. Summarization Algorithm

The overall procedure is summarized below.

```
Input: document D, summary length mode m
Output: plain-language explanation E
T ← extract_text(D)
T ← normalize(T)
S ← segment(T, size, overlap)
for each s in S do
    u ← LED_summarize(s, m)
    U ← U + u
end for
E ← aggregate(U)
E ← simplify(E)
return E with segment references
```

E. Legal Information Analysis

In addition to summarization, the system can identify important legal information such as parties, dates, monetary values, obligations and referenced clauses. Legal-domain models such as InLegalBERT may be integrated for future legal analysis tasks. For documents exceeding model limits, chunking and aggregation strategies can be applied to process each segment while retaining relevant context. Extracted entities are surfaced as a structured summary panel, which is often the fastest way for a user to answer practical questions such as what is owed, by when, and by whom.

F. Document-Grounded Question Answering

For a user question, the system retrieves the most relevant text segments from the uploaded document. These segments are supplied as context to the answer-generation component. This design improves traceability because the response is connected to the user's document rather than being generated without document context. If retrieval returns no segment above a relevance threshold, the system reports that the document does not appear to address the question instead of producing an unsupported answer.

G. User Interface

The interface is organized around a two-pane comparison view. The original document occupies one pane and the generated explanation the other, with corresponding regions highlighted on selection. A query box below the panes accepts natural-language questions, and answers appear with the retrieved clause quoted alongside. Summary length is exposed as a simple control rather than as a token count, since non-expert users reason more naturally about short, medium and detailed than about generation parameters.

V. EXPECTED RESULTS AND EVALUATION

A. Evaluation Objectives

The proposed system is expected to improve the accessibility and readability of legal documents for nonexpert users. Evaluation considers readability, factual consistency, relevance and user usefulness. Readability measures whether the explanation is easier to understand, while factual consistency measures whether important obligations and conditions from the source document are preserved.

B. Evaluation Metrics

Table III lists the metrics planned for each evaluation dimension.

TABLE III. PLANNED EVALUATION METRICS

Dimension	Metric	Purpose
Compression	Compression ratio	Extent of length reduction
Readability	Readability index	Ease of understanding for non-experts
Fidelity	Overlap and entailment scores	Retention of source meaning
Retrieval	Relevance of retrieved segments	Quality of grounding for answers
Human	Expert and user ratings	Practical usefulness and correctness

The compression achieved for a document of length Ld and summary length Ls is defined as

$$CR = 1 - (Ls / Ld) \quad (2)$$

A higher compression ratio is desirable only while factual consistency is maintained; the two are reported together rather than independently, because a very short summary can score well on compression while omitting the obligations that matter most.

C. Baseline Comparison

Two baselines are planned. The first is an extractive baseline that selects the highest-scoring sentences from the source without rewriting them. The second is a general-purpose abstractive model applied with truncation rather than segmentation. Comparing against the first isolates the benefit of abstraction, and comparing against the second isolates the benefit of the segmentation strategy for long documents.

D. User Study Design

For model-based evaluation, standard NLP and classification metrics are used where applicable. For summarization, human evaluation is important because automatic metrics alone may not fully determine whether legal meaning has been preserved. A study involving both legal professionals and non-expert participants is planned, in which participants answer comprehension questions about a document after reading either the original or the generated explanation. User feedback can also measure whether the generated explanation helps users understand the original document.

E. Ethical and Legal Considerations

The system should clearly communicate uncertainty and limitations. Generated explanations are intended for information and accessibility support and should not be treated as a legally binding interpretation or replacement for professional legal consultation. Uploaded documents frequently contain personal and financial information, so document handling, retention and deletion policies must be explicit, and users must be able to remove an uploaded document and its derived representations completely.

VI. LIMITATIONS

Several limitations follow directly from the design. Scanned documents without a text layer require optical character recognition, and recognition errors propagate into every downstream stage. Jurisdictional variation means that a

clause carrying one meaning in one legal system may carry another elsewhere, and the system does not currently reason about jurisdiction. Abstractive generation can still produce fluent statements that are not supported by the source, which is why retrieval grounding and side-by-side comparison are treated as required components rather than optional additions.

Processing time also grows with document length, since longer documents require more segments and therefore more forward passes. For very large documents the system is better characterized as an offline analysis tool than as an interactive one.

VII. CONCLUSION

This paper presents an AI-based approach for improving access to complex legal information. The proposed system combines PDF and text input handling, preprocessing, segmentation, transformer-based legal document summarization and document-grounded question answering. Long-document models such as LED provide a suitable approach for processing lengthy legal content, while legal-domain models such as InLegalBERT offer potential support for future legal analysis tasks.

The main contribution of the proposed system is to present complex legal information in a more understandable form for ordinary users. By combining AI and NLP with a user-friendly application design, the system can help reduce the gap between complex legal language and public understanding. Equally important is what the system does not claim: professional legal judgment remains necessary for important legal decisions, and the design deliberately keeps the source document visible so that the user is never asked to trust the explanation alone.

VIII. FUTURE SCOPE

Future work can extend the system with multilingual support for regional languages, improved optical character recognition for scanned legal documents and stronger clause-level source citations. Advanced retrieval methods and explainable AI can further improve transparency and trust. The system can also be expanded with personalized explanations based on user literacy level and additional legal-domain analysis features.

Further development may include secure document storage, role-based access, real-time notifications and optional lawyer consultation features. Cross-jurisdictional support and larger legal datasets can improve scalability and broader applicability. A clause-level risk indicator, highlighting terms that are unusual relative to comparable agreements, is a natural extension once a sufficiently large corpus is available. Future evaluation should involve both legal professionals and non-expert users to measure accuracy, faithfulness and practical usefulness.

ACKNOWLEDGMENT

The authors would like to thank Dr. Jagadishwari V, Professor and Head of Department, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, for providing an excellent academic environment and continuous support. The authors also express their sincere gratitude to Prof. K. Komala Devi, Assistant Professor, Department of Information Science and Engineering, CMR Institute of Technology, for her invaluable guidance and mentorship throughout the development of this work.

REFERENCES

[1] A. George et al., “Enhancing legal decision making: WRIT case outcome prediction with LegalBERT embeddings and AdaBoost classifier,” in Proc. 2024 IEEE Int. Conf. Contemporary Computing and Communications (InC4), 2024.

[2] D. Jain, M. D. Borah, and A. Biswas, “Summarization of lengthy legal documents via abstractive dataset building: An extract-then-assign approach,” Expert Systems With Applications, vol. 237, 2024.

[3] J. Wang et al., “Graph reasoning with supervised contrastive learning for legal judgment prediction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 2, 2025.

[4] Y.-H. Liu and Y.-F. Lin, “LASG: Streamlining legal adjudication with AI-enabled summary generation,” in Proc. IEEE COMPSAC, 2024.

[5] D. Singh, “Legal documents text analysis using natural language processing (NLP),” in Proc. 2024 2nd Int. Conf. Self Sustainable Artificial Intelligence Systems, 2024.

[6] L. Yue et al., “A circumstance-aware neural framework for explainable legal judgment prediction,” IEEE Trans. Knowl. Data Eng., vol. 36, no. 11, 2024.

[7] S. A. Mukund and K. S. Easwarakumar, “Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation,” Symmetry, vol. 17, no. 5, 2025.

[8] S. Sharma and P. P. Singh, “Advancing legal document summarization: Introducing an approach using a recursive summarization algorithm,” SN Comput. Sci., vol. 5, 2024.

[9] S. Pandi et al., “Enlightening justice: Empowering society through AI-driven legal assistance,” in Proc. 2024 2nd Int. Conf. Advances in Information Technology, 2024.

[10] T. Y. Basha et al., “Generative artificial intelligence in legal drafting,” in Proc. 2024 Int. Conf. Computational Intelligence for Green and Sustainable Technologies, 2024.

[11] T. Ansari, H. S. Dhillon, and M. Singh, “Machine learning model to predict results of law cases,” in Proc. 2024 4th Int. Conf. Advances in Electrical, Computing, Communication and Sustainable Technologies, 2024.

[12] P. Szemerényi, “LED large book summary model,” Hugging Face.

[13] Legal AI Lab, “InLegalBERT model,” Hugging Face.