

ARTIFICIAL INTELLIGENCE

Legal AI: An Intelligent Legal Document Simplification Using NLP

K N. Dhanya, R. Akshay Kumar, L S. Bhoomika, P. Divyanethri, S S. Kiran

Department of Computer Science & Engineering, Mysuru Royal Institute of Technology, Mandya, Karnataka, India

Published October 2026

Abstract

Legal documents such as rental agreements, employment contracts, service agreements, and policy documents often contain complex legal terminology, lengthy clauses, and technical language that make them difficult for non-legal users to understand. Identifying important conditions related to notice periods, penalties, and payment obligations requires considerable time and effort. To address these challenges, this paper presents Legal AI, an intelligent legal document simplification system using Natural Language Processing (NLP). The proposed system supports multiple document formats, including PDF, DOCX, TXT, images, and scanned documents, and uses Optical Character Recognition (OCR) for extracting text from scanned documents. The extracted content is processed to identify important legal clauses, generate simplified summaries, highlight potential risks, extract deadlines, analyze financial obligations, and provide document-based question answering through an AI assistant. Retrieval-Augmented Generation (RAG) is incorporated to support responses grounded in the uploaded document. The system is implemented as a web-based application using Python, Flask, MySQL, HTML, CSS, and JavaScript. The proposed platform aims to provide an accessible preliminary document-understanding tool for students, professionals, and general users by presenting complex legal information in a simplified and interactive form. The system is intended for preliminary document understanding and does not replace professional legal advice.

Keywords: Legal Document Simplification; Natural Language Processing; Optical Character Recognition; Retrieval-Augmented Generation; Clause Detection; Risk Analysis; AI Chatbot

1 Introduction

The rapid digitalization of legal documentation has transformed how contracts, agreements, and official documents are created, stored, and shared. Rental agreements, employment contracts, insurance policies, loan documents, and service agreements are increasingly available in digital formats, allowing faster communication and easier record management. However, despite this convenience, understanding these documents remains a major challenge for individuals without legal expertise. Legal documents frequently contain lengthy clauses, technical terminology, cross-references, and complex sentence structures that require careful interpretation.

Manual review of legal documents is both time-consuming and error-prone. Users often struggle to locate important clauses such as confidentiality, termination conditions, payment obligations, arbitration procedures, liability limitations, and governing law. Missing these clauses may result in unexpected penalties, financial loss, or contractual disputes. Existing AI-based legal tools generally focus on enterprise-level contract management and legal professionals, making them less suitable for students, small businesses, or ordinary users seeking quick document understanding.

To overcome these limitations, this research proposes Legal AI, an intelligent legal document simplification system that integrates Natural Language Processing (NLP), Optical Character Recognition (OCR), and Retrieval-Augmented Generation (RAG) into a unified web application. The system allows users to upload legal documents in multiple formats, automatically extract text from digital and scanned documents, identify critical legal clauses, generate simplified summaries, detect potential risks, extract deadlines, create compliance checklists, perform financial clause analysis, and answer document-specific questions through an AI assistant.

The developed prototype is implemented using Python [14] and Flask [12] for backend processing, MySQL [13] for document management, Tesseract OCR [11] for scanned document processing, and HTML [15], CSS, and JavaScript for an interactive user interface. The dashboard provides multiple integrated analysis modules, including Simplify Text, Summarize, Risk Detection, Clause Detection, Timeline Extraction, AI Assistant, Audio Explanation, Action Items, Compliance Checklist, Financial Analyzer, and Report Generation, allowing users to perform comprehensive document analysis within a single platform.

1.1 Problem statement

Understanding legal documents remains difficult for non-legal users due to complex language, lengthy clauses, and scattered contractual obligations. Existing solutions often require legal expertise or focus on limited functionalities, creating the need for an accessible AI-assisted legal document analysis platform.

1.2 Objectives

The proposed system aims to:

- Extract text from digital and scanned legal documents.
- Detect important legal clauses automatically.
- Generate simplified summaries.
- Identify contractual risks.
- Extract important deadlines.
- Generate compliance checklists.
- Analyse financial obligations.
- Provide document-based AI question answering.
- Support report generation for easier review.

1.3 Major contributions

The primary contributions of this work include:

1. A unified legal document simplification framework integrating NLP, OCR, and Retrieval-Augmented Generation.
2. Multi-format legal document processing, including digital and scanned documents.
3. Automated identification and analysis of important legal clauses, risks, deadlines, and financial obligations.
4. A document-grounded AI assistant for interactive question answering based on uploaded legal documents.

2 Literature survey

Artificial Intelligence has significantly transformed legal document processing through Optical Character Recognition (OCR), Natural Language Processing (NLP), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG). Recent studies demonstrate improvements in document summarization, clause classification, legal question answering, and contract analysis. However, most existing solutions concentrate on enterprise legal workflows or individual AI techniques instead of providing a complete user-oriented document analysis platform. The literature reviewed below highlights recent developments and identifies the research gap addressed by Legal AI.

2.1 OCR-based legal document processing

Legal documents frequently exist as scanned images or handwritten copies, making text extraction an essential first step before analysis. Modern OCR systems have considerably improved text recognition accuracy compared to traditional rule-based approaches.

OCR technology enables text extraction from scanned and image-based documents, making such documents suitable for subsequent NLP-based analysis. However, OCR performance can be affected by factors such as document quality, image resolution, font characteristics, and scanning conditions [11].

However, OCR performance still depends heavily on document quality, lighting conditions, and font variations.

In Legal AI, OCR is integrated through Tesseract OCR, enabling users to upload scanned agreements, notices, and contracts while maintaining compatibility with digital documents [11].

2.2 Natural language processing for legal documents

Natural Language Processing has become the foundation for understanding legal text. Unlike ordinary documents, legal documents contain long sentences, specialized terminology, references to previous clauses, and structured obligations that require contextual interpretation.

A comprehensive systematic mapping study published in IEEE Access (2024) reviewed the evolution of Legal NLP and found that transformer-based language models substantially improved legal document understanding, particularly for classification, summarization, and information extraction. The study also highlighted the growing importance of explainable AI for legal applications [1].

Legal AI adopts this direction by combining NLP-based clause identification with simplified explanations rather than only classifying legal content [10].

2.3 Legal document summarization

Automatic summarization has become one of the most useful applications of AI for legal professionals and ordinary users. Instead of reading dozens of pages, users can quickly understand important obligations through concise summaries.

Tekade et al. presented an NLP-based framework for simplifying legal document understanding through automated summarization techniques. Their work demonstrated that NLP significantly reduces the effort required to interpret lengthy legal documents, although the system focused primarily on summarization rather than complete document interaction [2].

Legal AI extends this idea by integrating summarization with risk detection, compliance analysis, and conversational assistance.

2.4 Retrieval-augmented generation for legal AI

Large Language Models can produce convincing answers, but they may generate incorrect legal interpretations when they rely only on pretrained knowledge. Retrieval-Augmented Generation addresses this limitation by retrieving relevant document content before generating responses.

Research on Enhancing Legal Document Understanding Using Retrieval-Augmented Generation (2025) demonstrated that grounding responses in retrieved legal passages improves response relevance while reducing unsupported answers [3].

Legal AI incorporates this principle by using document-grounded responses for questions related to the uploaded legal document.

2.5 Large language models in law

The rapid growth of Large Language Models has expanded legal AI applications beyond simple classification tasks. A 2025 IEEE Access study reviewing LLM applications in law discussed their use in areas such as legal document analysis, question answering, and other legal tasks [4].

Recent research has also explored reusable prompting frameworks for applying LLMs to legal tasks [8].

Legal AI incorporates LLM-based interaction to support document understanding and document-based question answering.

2.6 Multilingual legal AI

Legal systems increasingly require multilingual support because users may access legal documents in different languages. Recent research on legal translation has highlighted the challenges of translating legal documents while preserving terminology and legal meaning across languages [5].

Legal AI currently focuses on English legal documents and provides a translation capability to improve accessibility for users.

2.7 AI-assisted legal document review

Another important research direction involves AI-assisted contract review instead of fully autonomous legal decision-making.

An experimental study on in-context learning for legal document review demonstrated that carefully designed prompting strategies significantly improve review consistency while maintaining document context [6].

Legal AI incorporates this idea through guided conversational assistance rather than replacing legal professionals.

2.8 Agentic retrieval systems

The latest research introduces Agentic Retrieval-Augmented Generation, where intelligent agents perform multi-step reasoning before answering legal questions.

A recent IEEE Access (2026) paper showed that agent-based retrieval improves legal question-answering performance by combining retrieval with structured reasoning [7].

Although the current version of Legal AI does not implement autonomous agents, its modular architecture supports future integration of advanced retrieval strategies.

2.9 Research gap

The reviewed literature clearly indicates that existing research has made significant progress in OCR, NLP, LLMs, and Retrieval-Augmented Generation individually. However, several limitations remain:

- Most systems perform only one primary task, such as summarization or question answering.
- Enterprise legal solutions often require specialized legal knowledge.
- Few platforms combine OCR, clause detection, risk identification, compliance checking, financial analysis, and conversational assistance within a single interface.
- Limited emphasis is placed on improving accessibility for students and ordinary users.

Legal AI addresses these limitations by integrating multiple AI techniques into a unified web application that supports complete preliminary legal document analysis while maintaining an intuitive user experience.

3 Proposed methodology

Legal AI follows a modular document-processing architecture that combines Optical Character Recognition (OCR), Natural Language Processing (NLP), and Retrieval-Augmented Generation (RAG) to assist users in understanding legal documents. Instead of relying on a single AI model, the system processes documents through multiple independent modules, allowing different types of analysis to be performed from a common document repository.

The overall workflow begins when a user uploads a legal document through the web interface. The application accepts PDF, DOCX, TXT, image, and scanned documents. Digital documents are processed directly, while scanned documents undergo OCR using Tesseract to convert images into machine-readable text. The extracted text is cleaned, segmented, and stored before entering the NLP pipeline for clause identification and summarization, while relevant document content is retrieved for conversational querying.

Unlike conventional legal document analyzers that primarily focus on contract review, the proposed system integrates multiple assistance modules including risk detection, compliance checklist generation, financial analysis, deadline extraction, and AI-powered question answering within a single dashboard.

3.1 System architecture

The proposed architecture (Figure 1) consists of three major layers:

1. Presentation Layer
2. Processing Layer
3. Data Management Layer

The Presentation Layer provides an interactive web dashboard developed using HTML, CSS, and JavaScript. Users can upload documents and access different analysis modules through a unified interface.

The Processing Layer contains the Flask backend, OCR engine, NLP pipeline, clause detection module, summary generator, and Retrieval-Augmented Generation module.

The Data Management Layer stores uploaded documents, extracted text, and analysis history using MySQL.

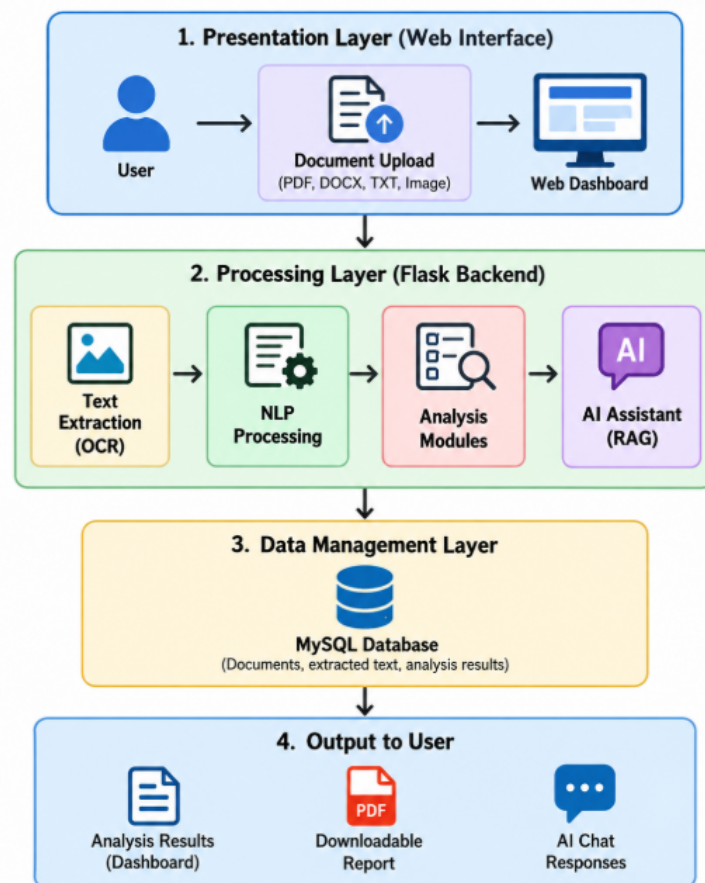


Figure 1. Proposed architecture of Legal AI.

3.1.1 Processing modules

- **Upload Module:** Accepts multiple document formats
- **OCR Engine:** Extracts text from scanned documents
- **NLP Engine:** Processes legal language
- **Clause Detector:** Identifies important clauses
- **Risk Analyzer:** Highlights risky conditions
- **Summary Engine:** Simplifies legal text
- **Timeline Extractor:** Finds important dates
- **Compliance Checker:** Generates checklist
- **Financial Analyzer:** Identifies payment-related information
- **AI Assistant:** Answers document-based questions

- **Report Generator:** Creates downloadable reports

3.2 System workflow

The document-processing workflow is illustrated in Figure 2.

The workflow performs the following sequence:

1. User uploads a document.
2. OCR extracts text when necessary.
3. Text preprocessing removes unwanted formatting.
4. NLP identifies important legal clauses.
5. Risk analysis highlights important obligations.
6. Summary generation simplifies the content.
7. Timeline extraction identifies deadlines.
8. Compliance checklist organizes obligations.
9. Financial analyzer extracts monetary information.
10. AI chatbot answers document-specific questions.
11. Report generator produces a downloadable summary.

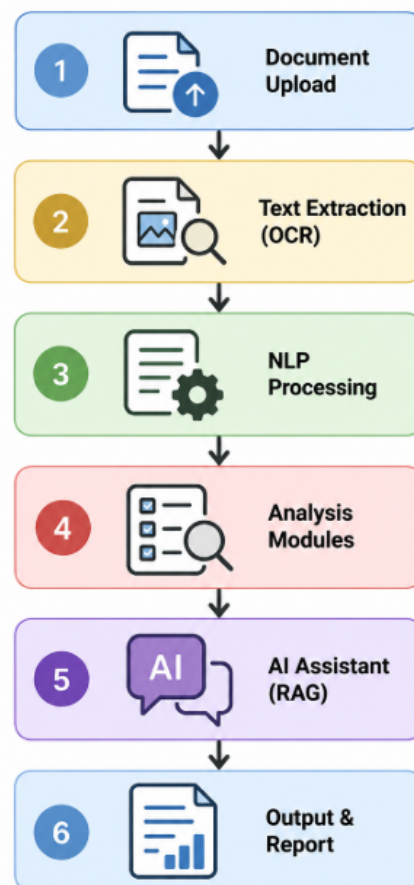


Figure 2. Workflow of the proposed Legal AI system.

3.3 Technology stack

Table 1. Technology stack of the prototype.

Component	Technology
Frontend	HTML, CSS, JavaScript
Backend	Flask(Python)
Database	MySQL
OCR	Tesseract OCR
NLP	Python NLP libraries
AI	Retrieval-Augmented Generation

Table 1 lists the technologies used. The modular design allows additional legal analysis features to be integrated without modifying the entire application.

4 Implementation and results

The Legal AI prototype was implemented as a web-based application providing multiple document analysis modules from a single dashboard. The implemented prototype integrates OCR, NLP, and conversational AI into a unified platform for preliminary legal document understanding.

4.1 Dashboard interface

The dashboard serves as the central access point for all document analysis functions.

The dashboard displays:

- Uploaded documents
- Risk indicators
- Saved document count
- AI confidence
- Quick access to every analysis module

The card-based interface allows users to navigate directly to individual tools without repeatedly uploading documents.

4.2 Simplify text module

This module converts complex legal language into easier-to-understand sentences while preserving the original meaning. It assists users who may not be familiar with legal terminology.

4.3 AI summary module

Instead of reading lengthy agreements, users receive concise summaries that highlight the most important information.

4.4 Risk detection module

The risk detection module highlights potentially important contractual conditions requiring user attention.

Examples include:

- penalty clauses
- restrictive conditions
- liability statements
- payment obligations

4.5 Clause detection module

The NLP pipeline automatically identifies major legal clauses, including:

- Confidentiality
- Termination
- Notice Period
- Payment Terms
- Liability
- Arbitration
- Governing Law

4.6 Timeline extraction

Important dates and deadlines are automatically extracted from uploaded documents whenever they appear.

4.7 AI legal assistant

Unlike a generic chatbot, the assistant answers questions using information extracted from the uploaded document, improving contextual relevance.

4.8 Accessibility features

The application includes additional accessibility-oriented modules.

Voice-assisted explanations improve usability for users who prefer listening rather than reading lengthy legal text.

4.9 Compliance and financial analysis

The system further extends document understanding through structured analysis.

The Compliance Checklist organizes legal obligations into actionable items, while the Financial Analyzer extracts payment-related information from uploaded documents.

4.10 Feature evaluation

Rather than evaluating only one AI capability, Legal AI integrates multiple document intelligence functions into a unified platform.

5 Discussion

The implemented system integrates multiple AI techniques within a single interface, providing a continuous workflow for preliminary legal document analysis. Instead of requiring users to switch

between separate OCR tools, summarization services, and chat assistants, the proposed platform provides a continuous workflow from document upload to actionable insights.

Current limitations include dependence on OCR quality for scanned documents and the present focus on English-language legal documents. These limitations provide opportunities for future enhancement through multilingual support and improved document preprocessing.

6 Conclusion

The increasing complexity of digital legal documents creates significant challenges for individuals who lack legal expertise. Manual review of contracts, agreements, and official documents requires considerable time and attention, while overlooking important clauses may lead to financial obligations or legal disputes. This work presented Legal AI, a web-based legal document analyzer that combines Optical Character Recognition (OCR), Natural Language Processing (NLP), and Retrieval-Augmented Generation (RAG) to simplify preliminary legal document understanding.

The proposed system integrates multiple document intelligence capabilities within a single platform, including multi-format document upload, OCR for scanned documents, automatic clause detection, plain-language summarization, risk identification, compliance checklist generation, financial clause analysis, timeline extraction, AI-assisted conversational interaction, audio explanation, and report generation. Instead of requiring users to switch between different tools, Legal AI provides a continuous workflow from document upload to actionable insights through an intuitive dashboard interface.

Unlike enterprise-focused legal AI solutions, the proposed application emphasizes accessibility for students, professionals, and general users by presenting legal information in simplified language while preserving document context. The modular architecture also allows future integration of additional legal intelligence capabilities without redesigning the entire system.

Overall, the developed prototype provides a practical foundation for improving legal document accessibility and supporting preliminary document review by combining OCR, NLP, and document-grounded conversational AI.

7 Future scope

Although the current implementation successfully demonstrates integrated legal document analysis, several enhancements can further improve the system's capabilities.

Future improvements include:

- Multilingual legal document support for Indian regional languages [9].
- Improved OCR accuracy for low-quality scanned documents.
- Advanced clause extraction using larger legal language models.
- Legal citation and precedent recommendation.
- Voice-based conversational interaction.
- Mobile application deployment.
- Integration with government legal databases for verified document validation.

These improvements would expand Legal AI from a preliminary document analyzer into a more comprehensive legal assistance platform.

References

- [1] Quevedo E, Cerny T, Rodriguez A, Rivas P, Yero J, Sooksatra K, et al. Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access*. 2024;12:145286–145317. <https://doi.org/10.1109/ACCESS.2023.3333946>
- [2] Tekade PM, Jadhav N, Chawda S, Bhosale V, Akashe Y. Legal Lens: Exploring NLP for document analysis in law. In: *Proc. 2024 IEEE Int. Conf. Blockchain and Distributed Systems Security (ICBDS)*. 2024. p. 1–5. <https://doi.org/10.1109/ICBDS61829.2024.10837237>
- [3] Elsayed M, Hassan A, Salah M. Enhancing legal document understanding and analysis using retriever-augmented generation (RAG). In: *Proc. 2025 Int. Conf. Intelligent Methods, Systems and Applications (IMSA)*. 2025.
- [4] Siino M, Falco M, Croce D, Rosso P. Exploring LLMs applications in law: A literature review on current legal NLP approaches. *IEEE Access*. 2025;13:18253–18276. <https://doi.org/10.1109/ACCESS.2025.3533217>
- [5] Niklaus J, Merane J, Nenadic L, Ahmadi S, Gao Y, Chevalley CAH, et al. SwiLTra-Bench: The Swiss legal translation benchmark. In: *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2025. p. 14894–14916. <https://doi.org/10.18653/v1/2025.acl-long.725>
- [6] Huber-Fliflet N, Zhang J, Gronvall P, Wei F, Spinelli P. Experimental study of in-context learning for text classification and its application to legal document review in construction delay disputes. In: *Proc. 2024 IEEE Int. Conf. Big Data (BigData)*. 2024. p. 2119–2125. <https://doi.org/10.1109/BigData62323.2024.10826061>
- [7] Lim EY, Park J. Agentic RAG for legal question answering in civil law: Evidence from the Korean bar examination. *IEEE Access*. 2026;14:124441–124458. <https://doi.org/10.1109/ACCESS.2026.3722717>
- [8] Sriram S, Vijayaraj N, Rajiv Krishna G, Vijay Bhanu S, Choudhary G, Murugan T. A reusable prompting framework for applying large language models to legal tasks. *IEEE Access*. 2026;14:3108–3129. <https://doi.org/10.1109/ACCESS.2026.3650941>
- [9] Paul A, Gupta AB, Bhat SY, Kumar PGS. Pre-training transformers on Indian legal text. In: *Proc. 2025 Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. 2025.
- [10] Fan A, Wang S, Wang Y. Legal document similarity matching based on ensemble learning. *IEEE Access*. 2024;12:33910–33922. <https://doi.org/10.1109/ACCESS.2024.3371262>
- [11] Smith R. An overview of the Tesseract OCR engine. In: *Proc. Ninth Int. Conf. Document Analysis and Recognition (ICDAR)*. 2007. p. 629–633. <https://doi.org/10.1109/ICDAR.2007.4376991>
- [12] Pallets. *Flask documentation (Flask 3.1.x)*. 2026.
- [13] Oracle Corporation. *MySQL 8.0 reference manual*. Oracle Corporation; 2026.
- [14] Python Software Foundation. *Python 3.11 documentation*. Python Software Foundation; 2026.
- [15] WHATWG. *HTML living standard*. Web Hypertext Application Technology Working Group; 2026.