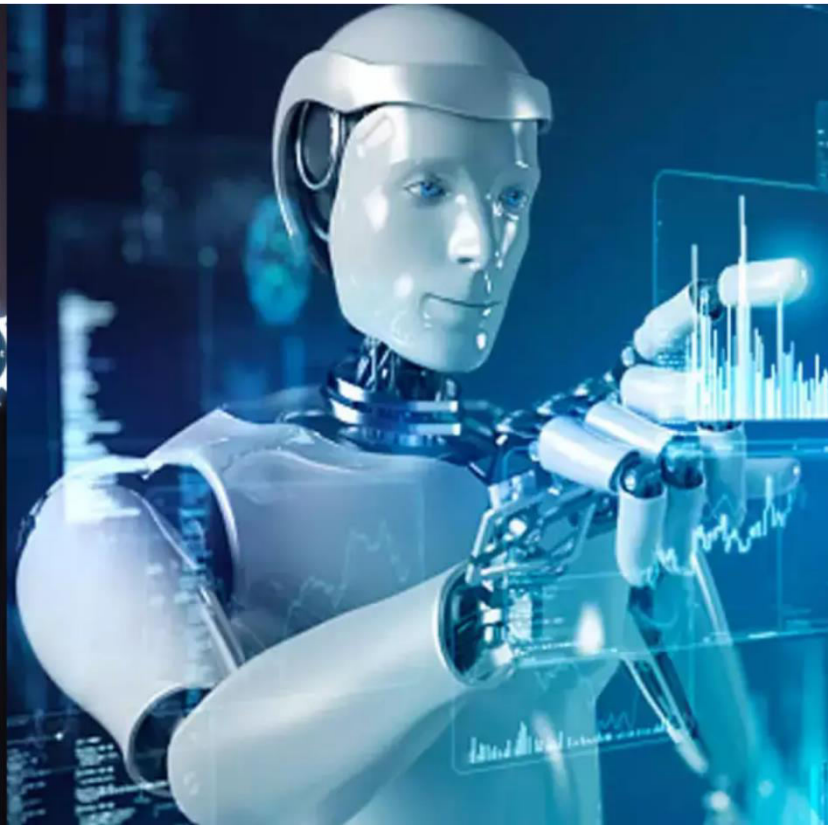
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

Intelligent Legal Documentation Analysis Framework for Entity Recognition and Context-Aware Data Highlighting

Chitoor Venkat Rao Ajay Kumar[1], Vakiti Sai Lohith Reddy[2], Korupu Manoj[3], Uppala Manikanta[4]

Assistant Professor, Department of CSE (AI & ML), ACE Engineering College Hyderabad, India[1]

Department of CSE (AI & ML), ACE Engineering College Hyderabad, India [2,3,4]

ABSTRACT: This project is a web application developed using Flask that aims to facilitate working with lengthy legal documents using LangChain and Open AI GPT-3.5 technologies. The application offers two features: text summarization and smart Q&A sessions. This application will enable users to upload documents or enter text and generate a summary that includes all the necessary information in a concise manner. Furthermore, users will also be able to pose context-specific questions related to their documents and obtain accurate answers using AI in real-time. This application will not only save users time but also increase their productivity while reducing manual reading time. This application is useful for students, researchers, and professionals in general who need to efficiently work with vast amounts of content in documents.

KEYWORDS: Legal Document Analysis, Entity Recognition, Context-Aware Summarization, LangChain, Flask Framework, OpenAI GPT-3.5, Text Summarization, Question Answering System.

I. INTRODUCTION

In the modern digital world, a tremendous volume of legal content is being produced every day in the form of contracts, case files, agreements, judgments, and other legal documents. It is quite challenging to interpret and analyze such voluminous legal content. For legal professionals, students, and researchers, it is always a challenge to go through the content of such voluminous legal documents to extract relevant entities, summarize the key content, and extract context-specific data. Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies have emerged as game-changers to overcome the existing challenges in the manual process of document understanding. Using the latest AI technologies like OpenAI's GPT and frameworks like LangChain, the project titled "Intelligent Legal Documentation Analysis Framework for Entity Recognition and Context-Aware Data Highlighting" has been proposed to make the process of analyzing complex legal content easier and more efficient. It is also possible to perform question-answer interactions using the proposed system, where context-specific questions can be asked and accurate answers can be obtained in real-time.

II. LITERATURE SURVEY

1. **Current Trends and Advances in Extractive Text Summarization — Maryam Azam, Shah Khalid, Sulaiman Almutairi (2025)** proposed a multi-layered extractive summarization framework using machine learning and deep learning techniques. Their study found that combining statistical, clustering, and deep learning methods improves summary accuracy and quality.
2. **Quality Requirement Documentation Guidelines for Agile Software Development — Woubshet Behutiye, Pilar Rodriguez, Markku Oivo (2022)** introduced the Agile QR-Doc framework using Design Science Research Methodology. The framework improves clarity, usefulness, and consistency in documenting quality requirements without affecting agility.
3. **Abstractive Summarization of Historical Documents — Keerthana Murugaraj, Salima Lamsiyah, Christoph Schommer developed HistBERTSum-Abs**, a domain-specific abstractive summarization method for historical documents. It outperformed standard baselines and matched advanced large language model performance.
4. **Context-Aware Recommender Systems for Social Networks — Areej Bin Suhaim, Jawad Berri** presented a

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

systematic literature review on recommender systems for social networks. The work identified major techniques, research gaps, challenges, and future opportunities.

5. **News Text Summarization Based on Multi-Feature and Fuzzy Logic — Yan Du, Hua Huo proposed a summarization model** using multi-feature analysis, fuzzy logic, and genetic algorithms. Results showed better performance than MS Word summarizer, System19, and Ranking SVM.

OBJECTIVES:

The aim of this project is to design an intelligent document analyzer assistant that is able to assist users in the analysis of text documents such as summarization, entities, and question answering in one interface.

The intelligent document assistant is meant to provide concise summaries, identify entities such as persons, dates, organizations, and locations, and answer questions based on the document. This is meant to assist users in understanding documents more efficiently by eliminating the need to read the documents.

The intelligent document assistant uses language processing techniques that are able to interpret the text content correctly and provide users with the ability to interact with documents efficiently.

III. METHODOLOGY

The methodology of the project includes several stages:

Document Upload:
Users upload legal documents through the web interface.

Text Preprocessing:
Cleaning and formatting of the document.Removing unnecessary characters and structuring the text.

Text Chunking:
The document is divided into smaller sections using LangChain for efficient processing.

Embedding Generation:
Each text chunk is converted into vector embeddings for semantic understanding.

Entity Recognition:
Named Entity Recognition identifies important legal entities in the document.

Semantic Retrieval:
Relevant document sections are retrieved based on user queries.

Context-Aware Question Answering:
The system uses GPT models to generate accurate answers from the document context.

Result Display:
Summaries, highlighted entities, and answers are shown to the user through the interface.

IV. PROPOSED SYSTEM

The proposed system is an AI-powered legal document analysis platform built using NLP and Large Language Models. The system works as follows:

Users upload legal documents through a Flask-based web interface.The system processes the document using LangChain and OpenAI GPT models.The text is divided into smaller chunks for efficient processing.(NER).The system generates automatic summaries of long legal documents.Users can ask questions related to the document and receive context-aware answers.This system improves efficiency compared to traditional manual analysis methods and provides faster and more accurate results.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

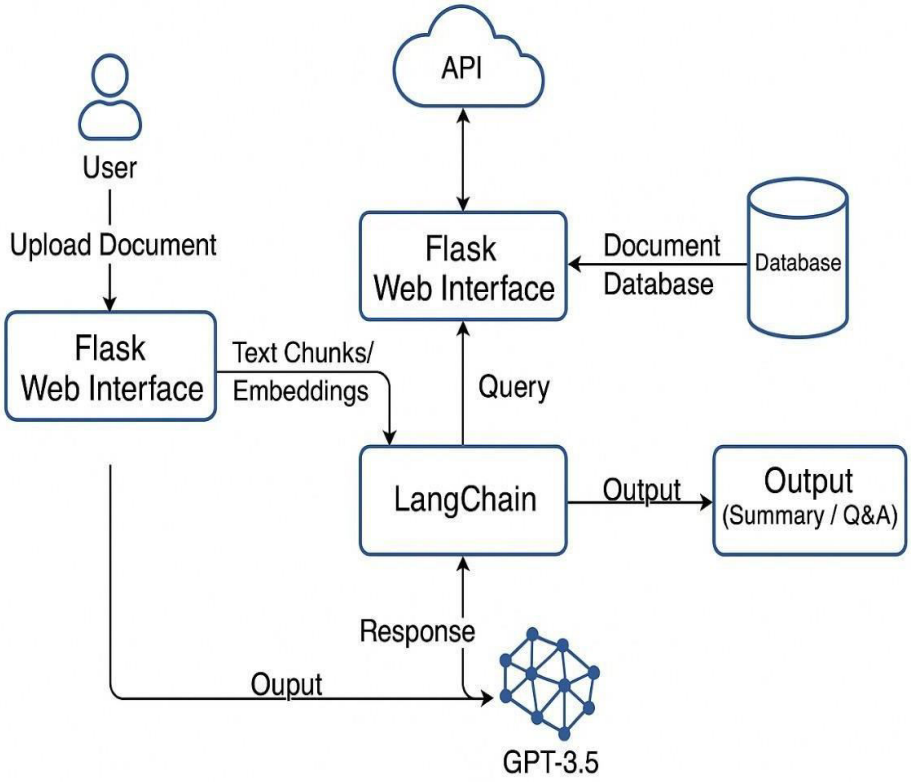


Fig.1. System Architecture

Architectural Flow Limitations in Existing Tools:

- **Integration of Advanced AI Models:** Recent AI technologies like GPT and modern large language models remain unused, revealing a gap in leveraging them to improve summarization and Q&A performance.
- **Domain-Specific Adaptation:** Many summarization and Q&A systems fail to perform well on technical, legal, or scientific documents that require specialized understanding.
- **User Control and Customization:** There is limited flexibility for users to adjust summary length, focus, or tone according to their specific needs.

Hardware Components:

Processor: Intel i3 or above RAM: 8 GB minimum Storage: 256 GB HDD/SSD
System: 64-bit Computer/Laptop Internet Connection

Software Components:

Operating System: Windows / Linux Programming Language: Python Framework: Flask
Libraries: LangChain, NLTK / SpaCy, Transformers Tools: VS Code / PyCharm
Web Technologies: HTML, CSS, JavaScript API: OpenAI GPT Model

V. APPLICATIONS

The proposed system can be used in multiple real-world scenarios such as:

Legal Research – Helps lawyers and researchers quickly analyze large legal documents and extract important information.

Law Firms & Courts – Automates document review, reducing manual effort and saving time.

Students & Academics – Assists law students in understanding long legal texts through summaries and highlighted entities.

Contract Analysis – Identifies key legal entities like names, dates, agreements, and clauses from contracts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

Document Question Answering Systems – Allows users to ask questions related to uploaded documents and receive context-based answers.

VI. ALGORITHMS

The project uses a combination of NLP and AI algorithms:

1. Named Entity Recognition (NER)
Used to identify:
Person names Locations Dates Organizations Legal terms

2. Text Chunking Algorithm
Splits large documents into smaller chunks for efficient processing and retrieval.

3. Embedding Algorithm
Converts text into vector representations using embedding models to understand semantic meaning.

4. Semantic Search Algorithm
Finds the most relevant parts of the document related to the user's query.

5. Transformer-based Language Model (GPT)
Generates:
Document summaries Context-aware answers
Intelligent responses to user queries.

6. Retrieval-Augmented Generation (RAG)
Combines document retrieval with AI-generated responses to improve accuracy.

VII. RESULTS

The test was aimed at checking how effective the Document Summarizer was in processing the uploaded documents and generating significant output. The process of summarizing the documents, identifying significant entities in the documents, and generating responses using the input documents was effective. The generation of output in a structured manner and identifying significant entities such as persons, dates, and organizations was effective. The generation of responses to questions using the input documents was also effective. The output generated was significant and related to the input documents. The execution of the AI Document Assistant was effective in uploading documents smoothly and generating responses using the documents uploaded.

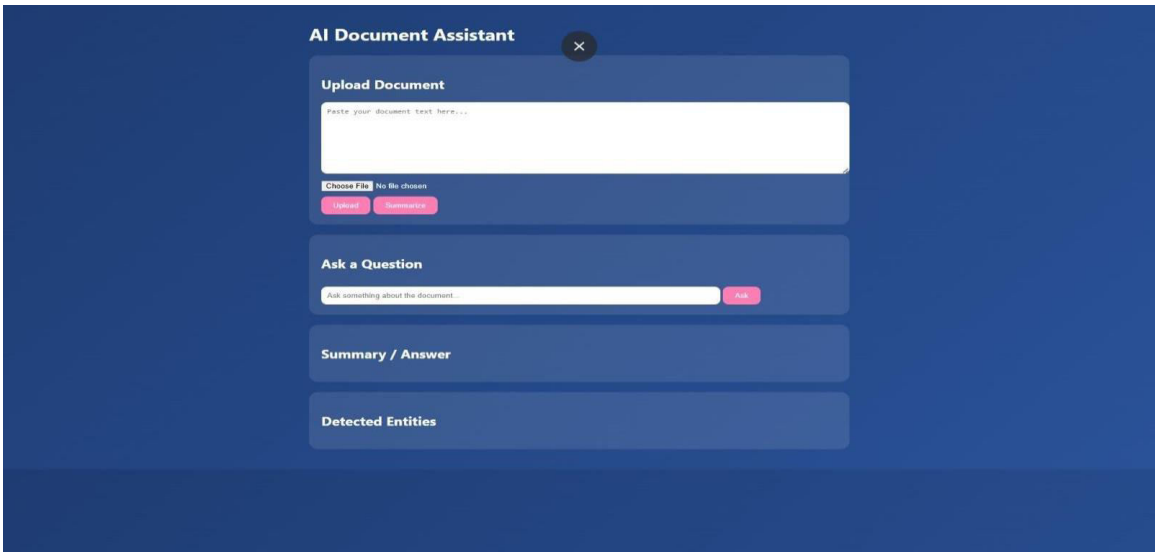


Fig.2. Document Summarizer Main interface

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

Fig.3. Summary and Entity Recognition Output

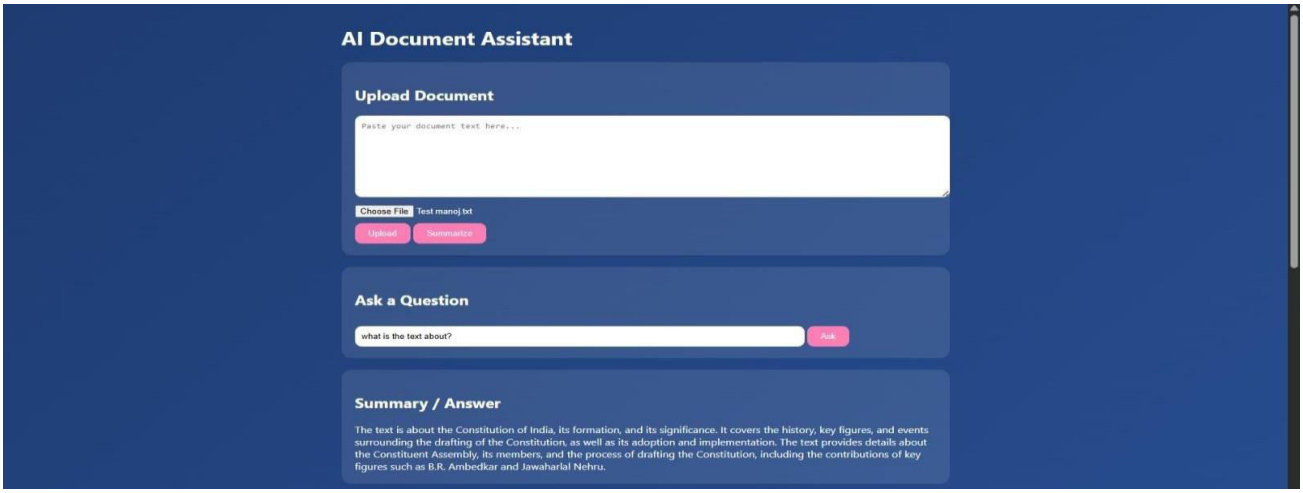


Fig.4. Document Question Answering Output

VIII. CONCLUSION

The AI-Based Document Question Answering System with API Validation showcases a good use case for artificial intelligence in intelligent document processing and question answering. The system utilizes natural language processing, semantic search, vector embeddings, and large language models to generate accurate and contextual answers based on uploaded documents. The addition of a vector database to efficiently retrieve relevant content and API validation to improve the accuracy and trustworthiness of generated answers is a good approach. This project can save considerable time and effort in reading documents manually and can significantly improve productivity by providing exact information instantly. Although there are a few limitations to this system, such as dependency on embedding quality

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503317|

and computational capabilities, it can be easily scaled up to improve its functionality. With further improvements such as adding multilingual support, improving fact-checking capabilities, and deploying the system on a cloud platform, this system has a strong potential to be implemented in a real-world environment. Overall, this project has been successfully implemented to show the potential of AI-based systems to transform traditional document interaction into a smarter, faster, and more efficient digital experience.

IX. FUTURE ENHANCEMENT

The AI-Based Document Question Answering System with API Validation has a strong potential for further expansion and improvement to make it more powerful, scalable, and industry-ready. Some of the future enhancements to this system include: The implementation of multiple document intelligence, where the system can process, analyze, and cross-reference multiple documents simultaneously. This will enable users to compare information contained in various documents such as reports, research papers, contracts, and business documents. Another important enhancement to this system will include the implementation of multiple language capabilities. This will enable users to use regional and international languages to perform question answering, where users can ask questions in one language and get answers based on documents written in a completely different language.

REFERENCES

[1] A. Vaswani *et al.*, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6000–6010.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
[3] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.
[4] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[5] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.
[6] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details