IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14573

LEGAL AI: An AI-Powered Legal Research and Case Prediction System for the Indian Judiciary

Pallavi Y[1], Amith M Shetty[2], R Bilwananda[3], Shalom Raj J[4], Shreyas M M[5], Suhas K M[6]

Assistant Professor, Department of CSE, MITM, Mysore, VTU Belagavi, India[1]

UG Student, Dept. of Computer Science and Engineering, Maharaja Institute of Technology, Mysore,

VTU Belagavi, India[2-6]

Abstract: This paper presents the design and implementation of LEGAL AI, an artificial intelligence-powered legal research and case prediction system customized for the Indian judicial context. It leverages a fine-tuned LLaMA-2 model and InLegalBERT using transfer learning and domain adaptation to provide functionalities such as case outcome prediction, legal explanation generation, and legal question answering (Legal QA). The system employs a Streamlit interface and FAISS-based vector search to retrieve relevant legal documents and provide contextual legal insights. With domain-specific fine-tuning and quantized models for CPU inference, LEGAL AI enhances accessibility, interpretability, and efficiency in legal research and decision-making.

Keywords: Legal AI, LLaMA-2, InLegalBERT, Legal Question Answering, Indian Judiciary, FAISS, Domain Adaptation , Retrieval-Augmented Generation.

I. INTRODUCTION

The Indian legal system is complex and vast, posing challenges to professionals and citizens in efficiently accessing and interpreting legal information. Existing keyword-based search tools lack the contextual depth required for nuanced legal analysis. LEGAL AI addresses this gap by introducing an AI-powered platform that performs legal prediction, reasoning, and question answering tailored to Indian law. The system fine-tunes LLaMA-2 and InLegalBERT using real-world Indian legal datasets and integrates with a Streamlit-based interface for intuitive usage. Additionally, the platform features a Legal Question Answering (Legal QA) module using Retrieval-Augmented Generation (RAG) that provides accurate and contextually grounded responses to user queries based on statutes such as IPC, CrPC, and the Constitution.

II. LITERATURE REVIEW

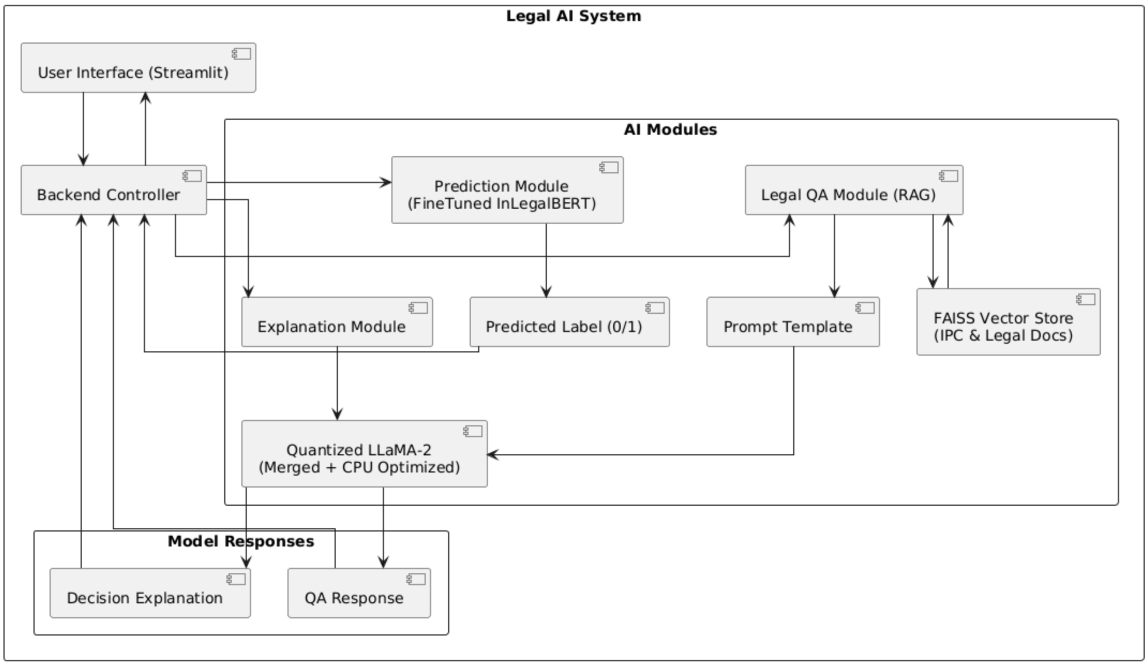
Recent literature shows the evolution of legal NLP from keyword-based systems to deep learning approaches. Zadgaonkar (2021) highlights NLP challenges in extracting structured data from legal documents. Naik (2022) illustrates the role of NER and summarization in simplifying legal judgment analysis. Quevedo (2022) traces the shift toward transformer-based models and emphasizes domain adaptation. LoRA (Hu et al., 2021) provides efficient fine-tuning with fewer parameters. Nigam (2023) demonstrates the importance of explanation-driven predictions in Indian courts. These insights validate LEGAL AI's architecture of using fine-tuned LLaMA-2, RAG-based Legal QA, and quantized models for efficiency.

III. SYSTEM ARCHITECTURE AND DESIGN

Architecture Overview:

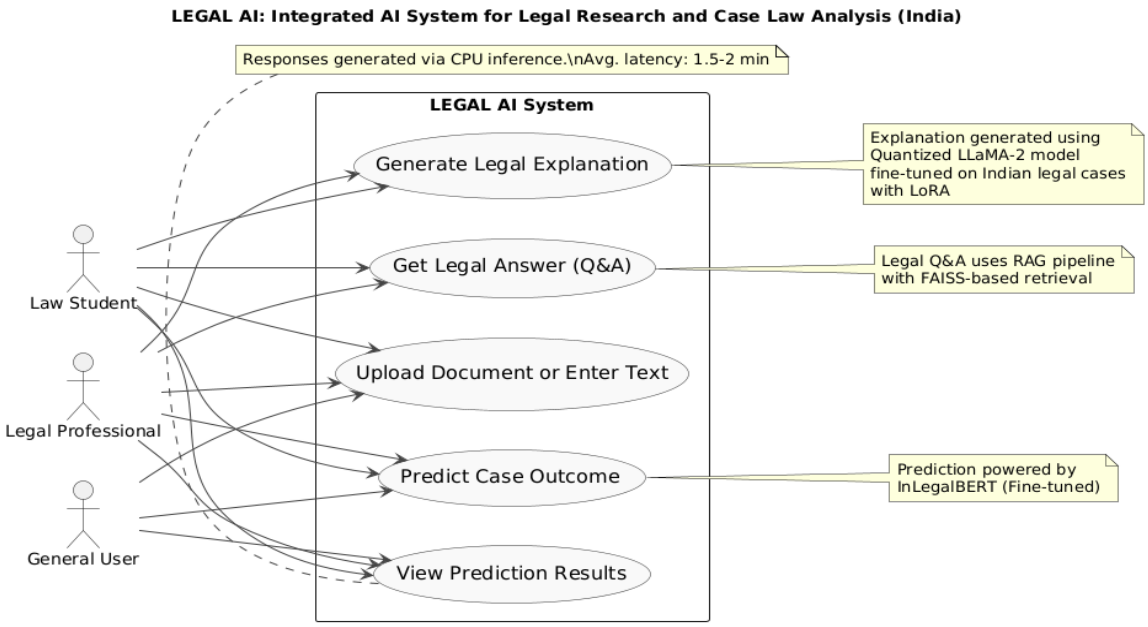
LEGAL AI is built on a modular architecture with three main components. The **user interface** is a Streamlit-based, chat-driven platform that facilitates user interaction. The **backend controller** manages and routes user queries to the appropriate processing modules. The **AI modules** include InLegalBERT for binary classification tasks, a LoRA-tuned LLaMA-2 model for generating legal explanations, and a Retrieval-Augmented Generation (RAG) pipeline integrated with FAISS for efficient and accurate Legal Question Answering (QA).

DOI: 10.17148/IJARCCE.2025.14573



B. Use Case Scenarios

LEGAL AI serves three primary user groups—**law students**, **legal professionals**, and **general users**—with key functionalities tailored to their needs. The platform enables **case outcome prediction** (accepted or rejected), **legal explanation generation** to understand the rationale behind decisions, and **legal question answering** based on Indian laws such as the IPC and CrPC. Additionally, it offers a feature to **export results as PDFs**, particularly useful for legal professionals during documentation and case preparation.



C. Design Decisions

The design of LEGAL AI emphasizes efficiency and maintainability. **Quantization using GGUF** was implemented to enable **CPU-friendly inference**, making the system more accessible on low-resource machines. **FAISS** was chosen for **fast and efficient document retrieval**, crucial for the Legal QA module. Additionally, the architecture was built with **modular components** to ensure ease of maintenance, scalability, and future upgrades.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁐ Peer-reviewed & Refereed journal ⁐ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14573

IV. IMPLEMENTATION AND METHODOLOGY

This Legal AI system tailored to the Indian legal domain, integrating fine-tuned transformer models, instruction-based prompting, and retrieval-augmented generation (RAG) for robust performance across distinct legal tasks: classification, judgment prediction with explanation, and legal QA.

A. Fine-Tuning InLegalBERT for Legal Judgment Classification

We fine-tuned the InLegalBERT model using two strategies—**shallow tuning (freezing all encoder layers and training only the classification head)** and **deep tuning (unfreezing the top 4 encoder layers: layers 8–11, along with the classification head)**—for a binary classification task (Accepted/Rejected). The dataset was tokenized using InLegalBERT's tokenizer with truncation and padding. A classification head was appended to the pre-trained model. Both configurations were trained using the HuggingFace Trainer API with identical hyperparameters: 6 epochs, learning rate 2e-5, batch size 16, AdamW optimizer, and linear warm-up scheduling. Evaluation metrics included accuracy, macro/weighted F1 score, precision, recall, and confusion matrix. Deep tuning outperformed shallow tuning in final accuracy and F1 score.

B. Instruction-Tuned Judgment Prediction with Explanation Using LLaMA + LoRA

For interpretable legal judgment prediction, we used the **PredEx dataset**, formatted for instruction-based supervision. The model was fine-tuned using **Low-Rank Adaptation (LoRA)** on LLaMA in 4-bit precision (load_in_4bit=True) to enable CPU-efficient training. Only LoRA parameters were trained, significantly reducing memory and computation requirements. Tokenization combined instruction, case summary, and response fields. Training utilized HuggingFace's Trainer, and logs were managed via Weights & Biases. The model was quantized to gguf format (Q4_K_M) using llama.cpp for inference. Final outputs included both a binary prediction and a rationale, enhancing transparency.

C. Legal QA via Retrieval-Augmented Generation (RAG)

For Legal QA, we implemented a **RAG pipeline** using LangChain. Key legal documents (e.g., Constitution, IPC) were ingested from PDFs via PyPDFLoader and split using RecursiveCharacterTextSplitter. Embeddings were generated using sentence-transformers/all-MiniLM-L6-v2 and stored in a **FAISS vector store**. Queries were processed via LangChain's RetrievalQA, retrieving relevant documents and generating answers using domain-specific prompts. The language model backend was powered by CTransformers, optimized for CPU inference.

D. Interface and Evaluation

A **Streamlit-based interface** allows real-time querying and document transparency. The system is continuously evaluated using legal metrics (accuracy, citation correctness), user feedback, and model retraining as the corpus evolves.

V. RESULTS AND DISCUSSION

A. InLegalBERT For Legal Prediction

Performance Metrics: Shallow vs. Deep Fine-Tuning (InLegalBERT - Prediction Task)

This chart compares key classification metrics—Accuracy, Precision, Recall, F1 Score, ROC AUC, and PR AUC—between shallow fine-tuning (classifier head only) and deep fine-tuning (last 4 layers unfrozen) of InLegalBERT.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ∺ Peer-reviewed & Refereed journal ∺ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14573

Confusion Matrix Comparison: Shallow vs. Deep Fine-Tuning (InLegalBERT - 6 Epochs)

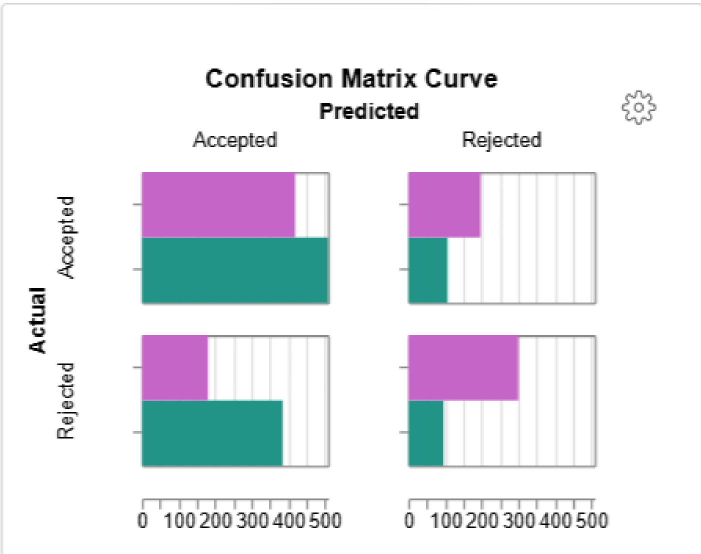


Chart 2: Confusion Matrix Curve

The confusion matrix shows that deep fine-tuning (pink) achieves higher recall for the 'Accepted' class by capturing more true positives, though with a slight increase in false positives. In contrast, shallow fine-tuning (green) is more conservative, reducing false positives but missing some true cases.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14573

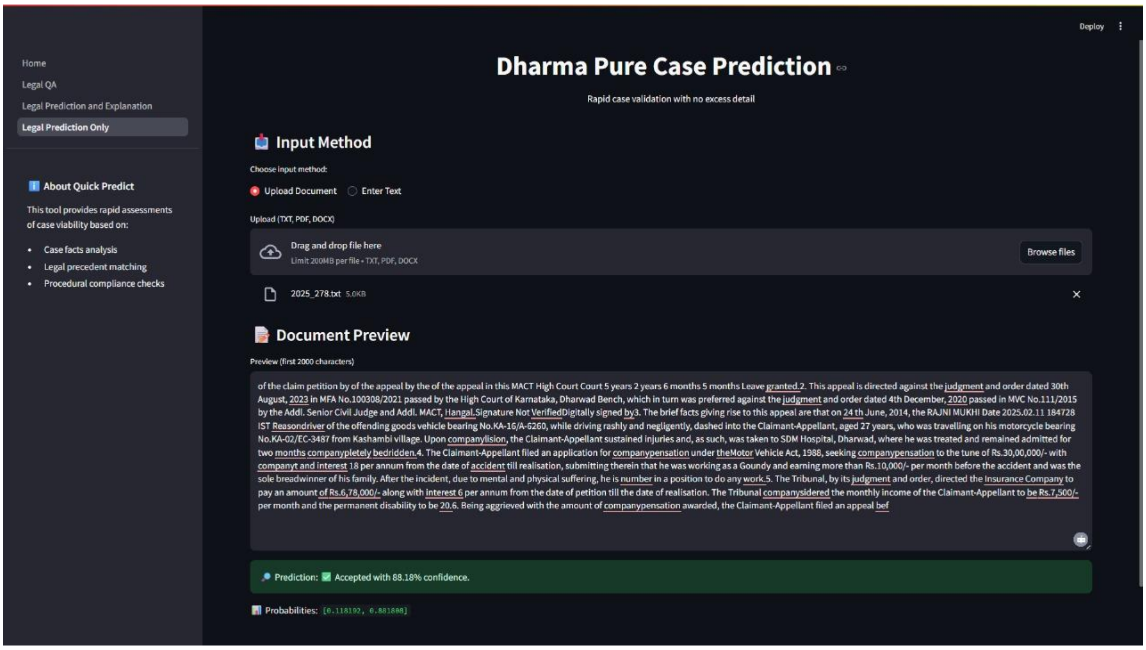


Figure: InLegalBERT prediction “Accepted”

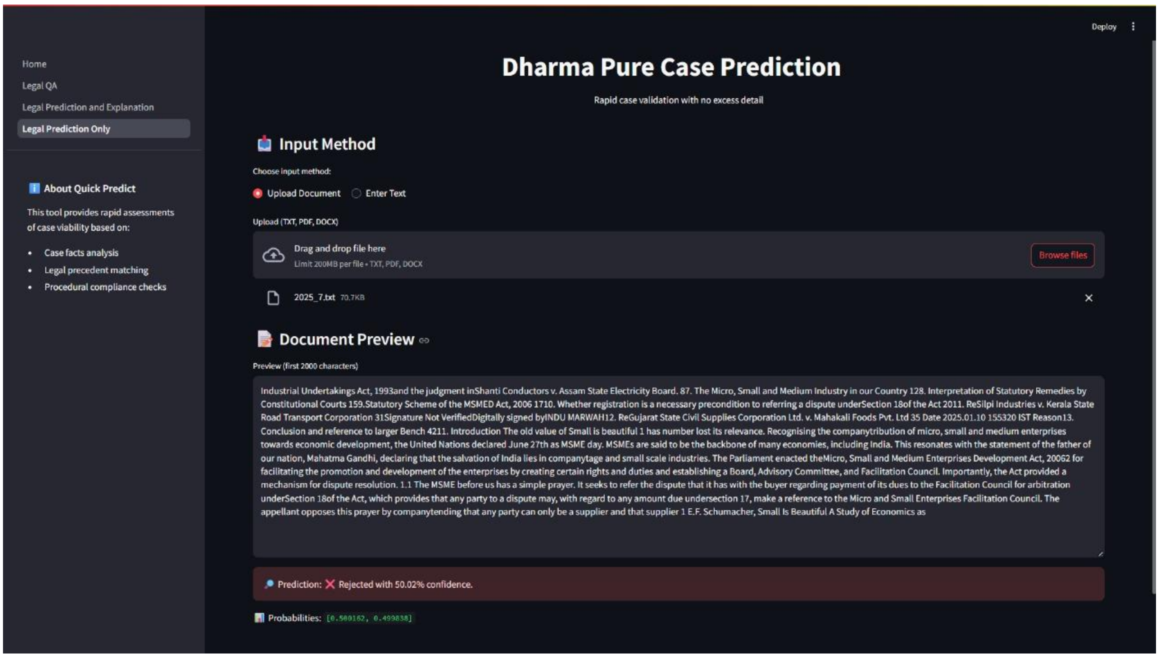


Figure InLegalBERT prediction “Rejected”

B. LLaMA-2 Evaluation

Legal AI is an AI-powered legal platform designed to help law students, legal professionals, and the general public navigate Indian legal texts, case laws, and statutes. It features three key modules—Legal QA, Case Prediction, and Case Explanation—which were tested for real-world applicability, responsiveness, and output quality. Law students used it as a virtual tutor, while junior advocates and paralegals leveraged it to evaluate case viability and streamline legal research. Its structured, responsive outputs proved valuable in legal drafting, moot courts, and exam revisions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14573

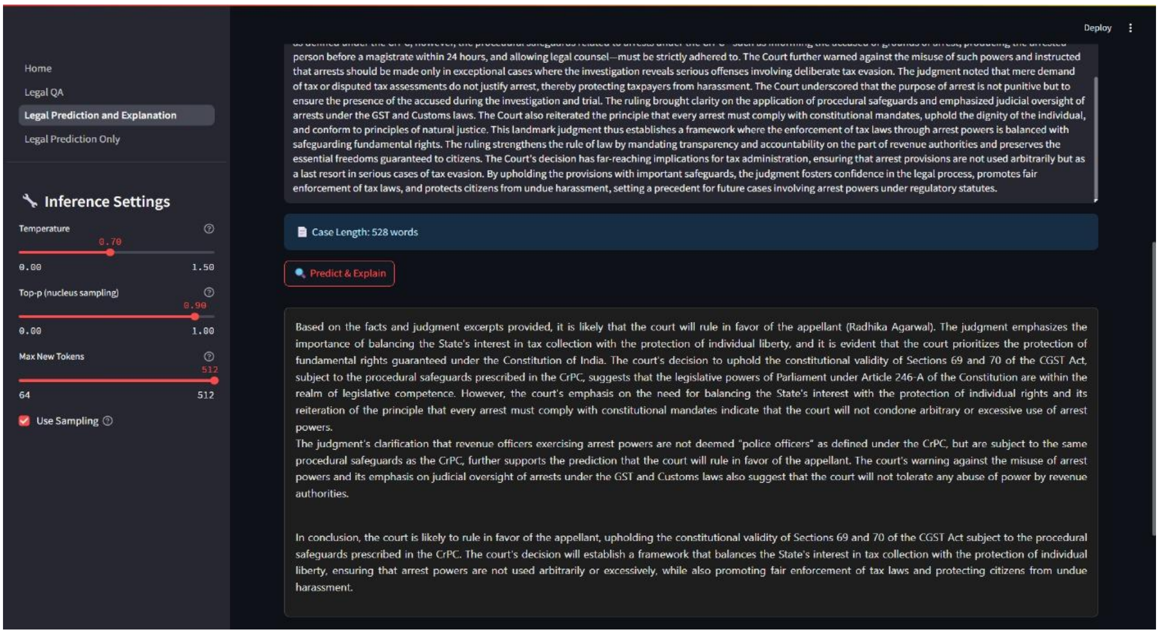
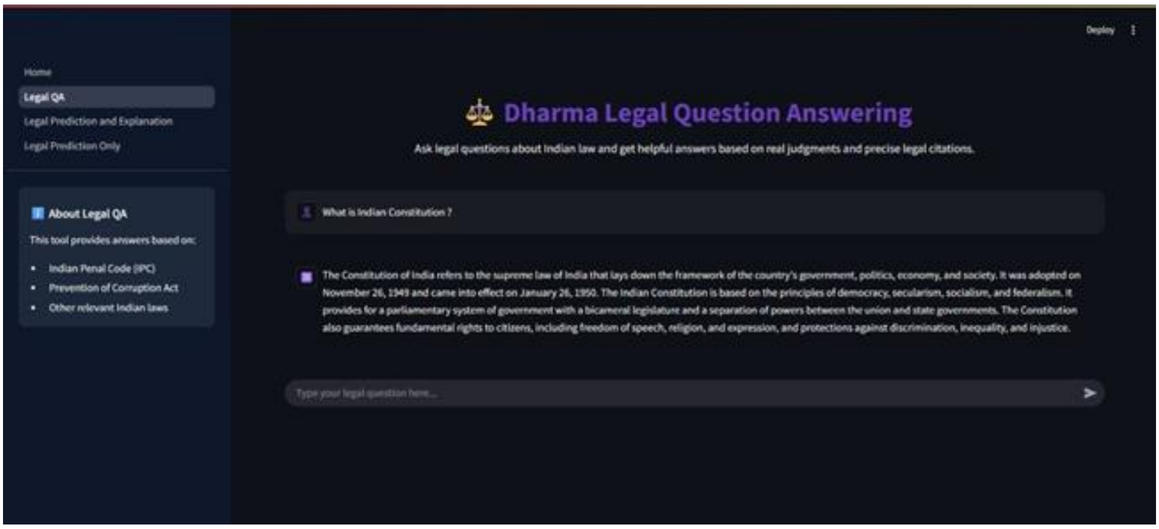


Figure Prediction and Explanation

C. Legal QA Evaluation

The RAG pipeline returned legally grounded responses for complex queries. Retrieval performance was consistent with cosine similarity and embedding quality.



D. Real-World Deployment

CPU-based deployment ensured the tool worked efficiently even in law colleges and NGOs. Streamlit interface made it user-friendly for non-technical users.

VI. CONCLUSION

This project introduces an AI-powered legal research and case law analysis system tailored to the Indian legal domain. Using transfer learning and domain adaptation, a LLaMA-2 model is fine-tuned on legal texts to understand complex legal language and context. It combines FAISS-based vector search with advanced NLP for efficient document retrieval, case prediction, and explanation generation. Unlike traditional keyword-based systems, it offers higher accuracy and contextual understanding. The system is scalable, adaptable to various legal use cases, and accessible to both professionals and non-experts, significantly improving the speed, accuracy, and accessibility of legal research and decision-making.

DOI: 10.17148/IJARCCE.2025.14573

REFERENCES

[1] A. V. Zadgaonkar, "An Overview of Information Extraction Techniques for Legal Document Analysis and Processing," Int. J. of Computer Applications, 2021.
[2] V. Naik, "An Effective Search Algorithm for Analyzing and Extracting Indian Legal Judgments using NER and Document Summarization," ICISDP, 2022.
[3] E. Quevedo et al., "Legal NLP from 2015–2022: A Systematic Mapping Study," J. of Info. Science, 2022.
[4] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
[5] S. K. Nigam, "PredEx and the Rise of Intelligent AI Interpretation in Indian Courts," Indian J. of Law and Tech., 2023.
[6] A. Farahani et al., "A Brief Review of Domain Adaptation," ACM Computing Surveys, vol. 54, no. 11, pp. 1–34, 2021.

