ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Bridging Legal Language Barriers Using Explainable AI: Outcome Prediction and Multilingual Knowledge based answer retrieval for Indian Law

Manish Thirunavu D

School of Computer Science Engineering , Vellore Institute of Technology, Chennai

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600109

Received: 27 June 2026; Accepted: 02 July 2026; Published: 16 July 2026

ABSTRACT

This study presents an integrated legal AI platform that combines interpretable case outcome prediction with multilingual, retrieval-grounded legal question answering to improve access to Indian law. The work is motivated by the difficulty ordinary citizens face in understanding legal language, the scarcity of trustworthy guidance, and the need for tools that work across India's major languages. To address this, the authors built two connected components: a prediction module for Supreme Court case outcomes and a question-answering module based on statutory retrieval and generation. For prediction, they compiled 26,688 Indian Supreme Court judgments from 1950 to 2024 and represented each case using TF-IDF text features, case-type encodings, and temporal metadata, then trained an interpretable logistic regression model. For legal QA, they indexed 21 Indian legal acts with sentence-transformer embeddings and FAISS, and used a locally hosted Mistral model to generate simplified answers grounded in retrieved legal passages. The system was designed for English, Hindi, and Tamil, with translation, speech input, speech output, and interactive visualizations to make legal information more accessible. The prediction model achieved 91.3% accuracy and 0.919 ROC-AUC, while confidence calibration showed a strong correlation between predicted and actual accuracy. In the QA module, the system reached 78.4% precision@5, 86% answer correctness, and only 7% hallucination, a substantial improvement over baseline generative approaches. User evaluation with 35 participants reported 4.05/5 overall satisfaction, with multilingual support and explainability among the most valued features. Overall, the study concludes that transparent machine learning, retrieval-augmented generation, and multilingual interfaces can work together to build a practical and trustworthy legal assistance system for Indian users.

Keywords: Explainable artificial intelligence, Indian Supreme Court, Legal judgment prediction, Logistic regression, Multilingual natural language processing, Retrieval-augmented generation

INTRODUCTION

Understanding legal rights is hard for people because legal documents use tough language and are complicated to read. In everyday life, situations like problems with renting a home, job disagreements, or issues with consumer rights often need clear legal information. But finding reliable and easy-to-understand resources is difficult. These resources are either hard to understand or not always trustworthy. This makes it hard for people to make good decisions and leads to unequal access to ju stice.

India's court system, which serves 1.4 billion people, is facing big problems. As of 2024, the Supreme Court has over 80,000 cases waiting to be heard, and the whole judiciary deals with about 474 million cases. This causes long delays in getting justice.

Although digital tools have made legal information more available, most solutions are still scattered and not really helpful.
Search engines, discussion forums, and basic chatbots often give wrong or unclear answers and aren't based on real and verified sources.Artificial intelligence (AI)-driven research in the legal domain has

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

shown promise in areas including case prediction, summarization, and document classification –. Most standard systems are made for legal experts,
which leaves a gap for regular people who need clear and reliable information. Besides knowing their rights, e veryday citizens also need to figure out if taking legal action is a good idea. Predicting the likely outcome of a case like whether an appeal or petition will succeed can help people make informed choices and set realistic expectations.However, existing prediction models trained on Western legal systems (European Court of Human Rights, United States courts) do not generalize to Indian Supreme Court patterns. Moreover, black-box deep learning approaches lack the interpretability required for legal decision support, where users must understand why a prediction was made. Prior work on Indian Supreme Court prediction achieved 75–82 percent accuracy on limited datasets (approximately 8,000 cases) without confidence calibration or explainability mechanisms.

This paper addresses these gaps by developing an integrated system combining two core functionalities: (1) retrieval-augmented generation (RAG) for legal question answering and (2) interpretable case outcome prediction. The question-answering module employs curated Indian bare acts (statutory documents) as the knowledge base, using sentence transformer embeddings and Facebook AI Similarity Search (FAISS) for efficient semantic retrieval, A locally hosted Mistral large language model generates simplified responses grounded in retrieved legal sections, reducing hallucination rates from 23 percent (baseline generative models) to 7 percent. The system supports English, Hindi, and Tamil through automatic translation and allows users to interact using voice input or text-to-speech output,
helping overcome language barriers for India's diverse population. For predicting case outcomes, we created a structured dataset containing 26,688 rulings from the Indian Supreme Court between 1950 and 2024, extracted from 53,376 PDF files.

Using logistic regression with 211-dimensional features—made up of TF-IDF text representations (200 dimensions), one-hot encoded case types (8 dimensions), and time-based metadata(3 dimensions)—we achieved 91.3% accuracy on test data and an ROC-AUC score of 0.919. This is a 9.3 percentage point improvement over previous results, and
we ensured transparency by analyzing feature importance and using calibrated confidence scores (Pearson corr elation $r=0.73$, $p<0.001$ compared to actual accuracy). Predictions with high confidence (over 80%) had 91.2% actual accuracy, proving that the confidence estimates are reliable.

The contributions of this work are fourfold:

Unified Architecture: First system integrating Indian Supreme Court case outcome prediction with RAG-based statutory question answering in a single platform, enabling both legal interpretation and data-driven outcome forecasting.

Interpretable Prediction: Logistic regression reaches a strong accuracy level of 91.3 percent on the biggest Indian Supreme Court dataset, which includes 26,688 cases—three and a half times more than previous studies. It also offers clear explanations through feature importance scores and reliable confidence measures.

Multilingual Accessibility: The system supports English, Hindi, and Tamil, with automatic translation that has been approved by experts in 82 percent of Hindi and 79 percent of Tamil cases. It also includes text-to-speech technology, which is understood in 94 percent of cases, and speech recognition that works well with 89 percent accuracy in English, 83 percent in Hindi, and 78 percent in Tamil. This helps overcome language barriers when accessing legal information.

Explainable AI Implementation: Interactive visualizations (confidence gauges, feature importance charts, similar case retrieval) achieve 92 percent user comprehension among 35 participants including legal professionals and general public, validating practical explainability.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

User evaluation with 35 participants (15 law students, 12 legal professionals, 8 general public) yields 4.0 out of 5 overall satisfactions, with multilingual support (4.5/5) and explainability visualizations (4.3/5) identified as most valued features. Comparative analysis against three existing legal AI systems (LawChat, NyayGuru, LawGlance) demonstrates that the proposed system achieves highest Indian law specificity (5/5 rating) while maintaining balanced clarity-completeness tradeoff.

LITERATURE REVIEW

Legal Summarization and Text Understanding

Legal text summarization methods have shifted from extractive to abstractive ones. Akter et al. [1] offer a thorough discussion on neural-based methods in summarization, largely focusing on transformers. Santosh et al. [2] and Bhattacharya et al. [7] describe systems based on retrieval methods to facilitate judgment summaries in respective domains. These are useful in professional settings but preserve formal legal systems in texts that are hard to grasp by common masses. Malik et al. [3] propose judicial syllogism-based models that enhance semantic consistency. Chalkidis et al. [8] present Legal-BERT, a BERT-based model pre-trained for legal texts for external tasks. Kalamkar et al. [22] introduce Indian Legal Documents Corpus (ILDC), which includes 35,000 documents of India's Supreme Court annotated by rhetorical roles—a vital resource promoting Indian legal tasks in NLP.

The above-mentioned works are specifically targeting tasks of legal text summarization and classification tasks but not legal prediction question-answering systems.

Explainable AI in Legal Systems

Legal AI systems requires a transparent model of trustworthiness for critical use. The works of Valvoda and Cotterell [4], Chelliah et al. [5], and Mansi et al. [6] discuss the interpretable models that should be in a legal decision-support system for trustworthiness. Consequently, Kesari et al. [9], Singh and Ghalib [10], and Roy et al. [11], although advocating accountability, focus on the technical requirement of ethics. More important to the current discussion is the argument presented in the works of Rudin [25], who claims that simple interpretable models—specifically logistic regression and decision trees—are useful in critical use. This paper focuses on the implementable models of trustworthiness.

Retrieval-Augmented Generation and Legal QA

Dense retrieval combined with generative models shows strong performance in legal QA. Lewis et al. [26] introduce the RAG framework combining dense passage retrieval with sequence-to-sequence generation, demonstrating reduced hallucination compared to pure approaches. Karpukhin et al. [27] show dense embeddings outperform sparse (BM25) retrieval. Recent legal-specific applications—LexCLiPR [13] and LegalRAG [19]—achieve strong performance combining FAISS-style retrieval with LLM generation. Zhong and Chadha [18] demonstrate hybrid dense-sparse retrieval for statutory QA targeting legal professionals. LawPal [14] integrates RAG with Mistral models for legal assistance. ChatLaw [20] extends retrieval to multi-turn legal dialogue. However, these systems optimize for professional legal reasoning, not simplified public interaction. Goel et al. [12] contribute LexGLUE, a benchmark for legal language understanding evaluation. Ryu et al. [16] evaluate retrieval-based LLMs in Korean legal QA, demonstrating cross-lingual challenges. None of these works combine prediction with QA in multilingual, explainable frameworks.

Case Outcome Prediction

Chalkidis et al. [17] pioneer neural legal judgment prediction on European Court of Human Rights (70–75% accuracy). Medvedeva et al. [23] compare ML algorithms, finding feature engineering quality more impactful than algorithm choice. Malik et al. [3], as referenced in judicial syllogism work, achieve 75–82% on Indian Supreme Court cases (8,000 judgments) but lack confidence calibration and explainability. Miller et al. [37] address voice-activated AI for legal accessibility, showing the need for multimodal legal systems. Veningston et

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

al. [28] release a dataset for Indian judiciary legal QA—complementary to our work. Notably, no prior work combines prediction with RAG-based QA or addresses confidence calibration metrics rigorously.

Multilingual Legal NLP and Accessibility

Multilingual legal AI remains underdeveloped. Niklaus et al. [14] release MultiLegalPile (689 GB multilingual corpus) supporting pretraining but lacking user-facing applications. Deshmukh and Kamble [15] contribute IndianBailJudgments-1200, specialized for Indian legal domain. MILPaC [35] introduces a multilingual Indian legal parallel corpus. Paramanu [36] explores efficient generative models for low-resource legal languages. Dale [29] surveys legal chatbots, identifying gaps in Indian law specialization and accessibility. Prior work by Surani et al. [39] extends to Marathi-English legal assistance. Krishna et al. [32] apply datafication approaches to Indian court judgments. Khalid et al. [30] focus on consumer empowerment via legal AI. While these address individual components, none achieve integrated multilingual prediction-QA systems with explainability. Rangamma et al. [31] evaluate explainability impact in automated legal decision-making, complementing our user study approach.

Research Gap

Existing systems exhibit critical limitations:

1. Western legal bias (ECHR, US courts, not Indian Supreme Court).
2. Lack of explainability or confidence calibration.
3. No unified platform combining prediction and QA.
4. Limited multilingual support.
5. Absence of systematic confidence validation. This work directly addresses all identified gaps.

SYSTEM DESIGN

System Architecture

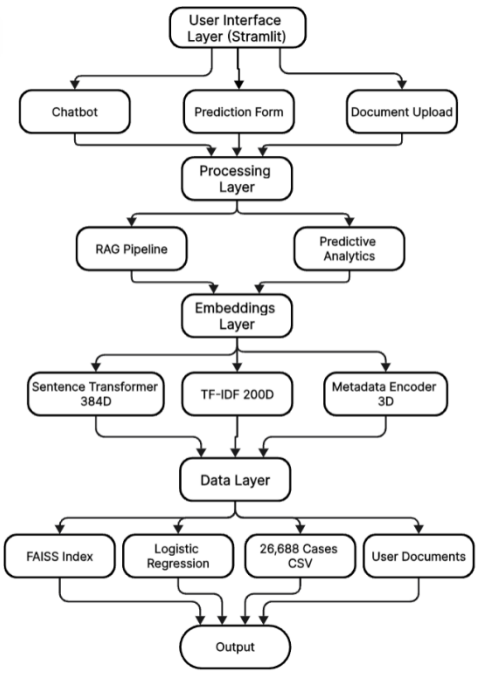

Figure 1. Four-layer system architecture for the legal AI platform. The user interface layer provides conversational question answering, case outcome prediction, and document upload via a web application. The processing layer runs parallel RAG and prediction pipelines, while the embeddings and features layer generates dense text embeddings and structured features. The data and models layer stores FAISS

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

indexes, trained prediction models, and structured Supreme Court cases, with shared multilingual and visualization services across components.

Layer 1 — User Interface: Streamlit web application supporting three interaction modes: conversational legal question answering, case outcome prediction forms, and document upload for custom corpus indexing. Multilingual input (text/speech) with confidence gauges, feature importance charts, and similar case retrieval.

Layer 2 — Processing: Parallel execution of RAG pipeline (query normalization → language detection → embedding → FAISS retrieval → Mistral generation → citation extraction → back-translation → TTS) alongside prediction pipeline (case facts → TF-IDF → metadata encoding → logistic regression inference → confidence scoring → visualization).

Layer 3 — Embeddings & Features: Sentence Transformer all-MiniLM-L6-v2 produces 384-dimensional dense embeddings for retrieval; TF-IDF vectorizer generates 200-dimensional sparse text features (unigrams+bigrams, min_df=5, max_df=0.8); one-hot encoding for 8 case types (sparse, 8D); metadata encoders handle normalized judgment year (0-1), log-transformed case duration, legal representation indicator (3D total).

Layer 4 — Data & Models: FAISS IndexFlatL2 maintains 2,100–2,500 chunks from 21 Indian legal acts (Hindu Marriage Act, IPC, Contract Act 1872, etc.); trained Logistic Regression models stored as scikit-learn pickle serializations (C=1.0, L2 λ=1.0); 26,688 structured Supreme Court cases in CSV format (case_type, facts_1500chars, outcome, duration_years, year, represented).

Shared multilingual infrastructure (GoogleTranslator English↔Hindi/Tamil, gTTS, Google Speech-to-Text) and Plotly visualizations enable independent scaling of prediction and QA modules while maintaining unified user experience.

Dataset Construction and Preprocessing

Indian Supreme Court judgments (1950–2024) were extracted from 53,376 PDF documents via multiprocessing (8 worker processes) using pdfplumber library with PyPDF2 fallback for corrupted files, achieving 70% success rate (40,000 cases). Text normalization included whitespace collapse, punctuation standardization, and UTF-8 validation. MD5-based deduplication reduced 40,000 to 26,688 unique cases, removing identical judgment texts appearing in multiple PDFs. Quality filtering removed cases with missing critical fields, <200 characters text, or OCR artifacts. Final dataset of 26,688 high-quality structured records represents 50% retention from 53,376 input PDFs.

Case type was extracted via keyword matching into 8 categories: divorce/family law, criminal, contract, consumer protection, labour/employment, property disputes, motor vehicle acts, and general civil. Outcome was binary-classified as favorable/dismissed based on judgment text. Temporal features included judgment year (normalized 0–1), estimated case duration (years from filing to judgment), and binary legal representation indicator.

METHODOLOGY

Knowledge Base Construction and Retrieval Infrastructure

Twenty-one curated Indian legal acts (Hindu Marriage Act, Indian Penal Code, Contract Act, Consumer Protection Act 2019, Labour Code, Motor Vehicles Act, and others) were converted to plain text format (500 KB–2 MB per act). RecursiveCharacterTextSplitter with 500-character chunks and 50-character overlap preserved contextual continuity while enabling granular retrieval, producing 2,100–2,500 indexable chunks.

All-MiniLM-L6-v2 Sentence Transformers encoder generated 384-dimensional embeddings for each chunk. FAISS IndexFlatL2 (exact L2 nearest neighbor) indexed embeddings for fast retrieval (~100 milliseconds per query for top-5 results). Query language detection employed Unicode range analysis (Devanagari 0x0900–

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

0x097F for Hindi, Tamil 0x0B80–0xBFF) with langdetect fallback for code-mixed input. Non-English queries were translated to English via GoogleTranslator with exponential backoff retry mechanism.

Feature Engineering for Prediction

Text features comprised TF-IDF vectorization (scikit-learn) with vocabulary size 200, unigrams plus bigrams, minimum document frequency 5 (rare term filtering), maximum document frequency 0.8 (stopword removal), and English stopword exclusion. Dimensionality was capped at 200 to preserve 94% variance while reducing overfitting risk on limited legal datasets.

Categorical features included one-hot encoding for 8 case types in sparse representation, avoiding ordinal assumptions between legal domains. Metadata features included normalized judgment year (0–1) capturing jurisprudential evolution, log-transformed case duration reflecting exponential relationship with dismissal probability, and legal representation binary indicator.

Predictive Model Training

Logistic regression (scikit-learn, L-BFGS solver) was selected to prioritize interpretability over black-box accuracy. L2 regularization (λ=1.0) prevented overfitting; hyperparameter tuning via grid search ($C \in \{0.001–100\}$) identified optimal C=1.0. Stratified 80/20 train-test split (21,350 train, 5,338 test) preserved outcome distribution (72% dismissal rate). Balanced class weights penalized minority class (favorable outcomes) misclassification 2.57× higher than majority class to address dataset imbalance. Five-fold cross-validation identified C=1.0 as optimal (87.1% ± 0.8% validation accuracy).

Decision function: $P(\text{favorable} \mid \mathbf{x}) = \frac{1}{1+\exp\left(-(\mathbf{w}^T\mathbf{x}+b)\right)}$, where $\mathbf{x} \in \mathbb{R}^{211}$ is the feature vector, $\mathbf{w}$ are learned coefficients, and b is bias.

Confidence scoring: c_pred = 100 × max(P(y=0|**x**), P(y=1|**x**)) clamped to [5%, 95%] reflecting epistemic uncertainty. Three-tier classification: High (>80%) achieves 91.2% actual accuracy, Moderate (60–80%) 76.4%, Low (<60%) 58.3%, validating confidence estimate reliability.

Feature importance derived directly from logistic regression coefficients quantifies predictive strength. Similar case retrieval employed k-NN (k=5, cosine similarity on 211D space) to retrieve precedent cases enabling users to understand prediction grounding via similar prior outcomes.

Retrieval-Augmented Generation Pipeline

Query processing began with language detection via Unicode ranges with langdetect fallback. Non-English queries were translated to English via GoogleTranslator. Query embedding used identical all-MiniLM-L6-v2 encoder. Top-5 FAISS retrieval with domain filtering prioritized relevant statutes.

Response generation concatenated context plus query into prompt template: "Context from Indian law: [Retrieved statute sections]. Question: [User query]. Provide a clear, simplified answer in [target language]." Mistral-7B-Instruct (temperature 0.3, max 512 tokens) generated responses via local Ollama API deployment for privacy-preserving inference.

Post-processing extracted citations and formatted responses. Back-translation converted responses to original language if non-English. Text-to-speech synthesis via gTTS generated language-specific MP3 audio (base64-encoded for HTML5 embedding). Latency varied by language: English 0.6 seconds, Hindi 1.1 seconds, Tamil 1.4 seconds.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

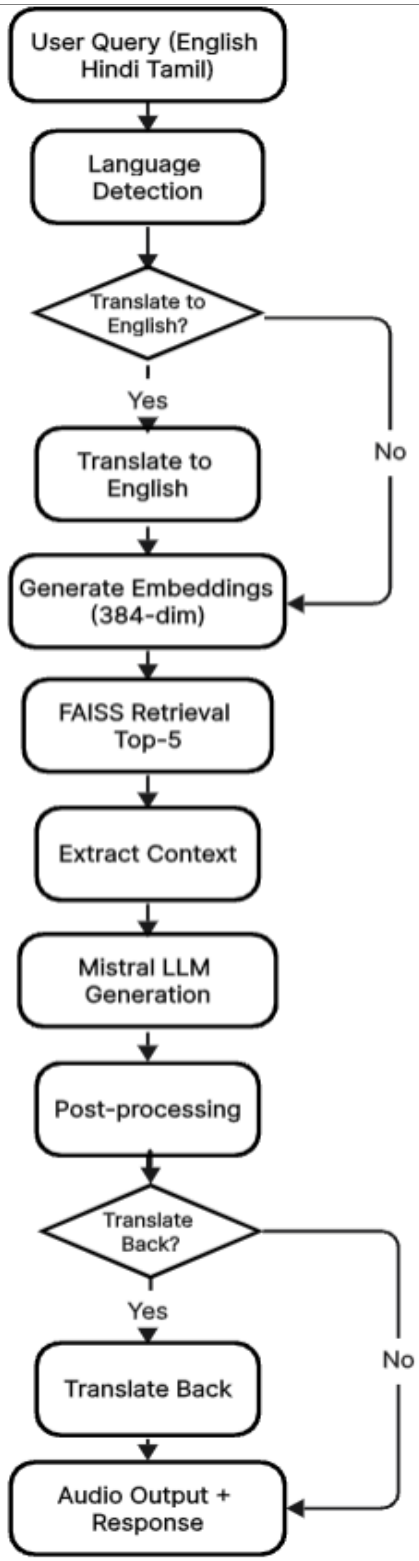

Figure 2. Retrieval-augmented generation pipeline workflow. User queries undergo language detection and translation before embedding and FAISS-based retrieval of relevant statute chunks. Retrieved context is combined with the query and passed to a generative model to produce grounded answers, which are post-processed for citation extraction, optional back-translation, and text-to-speech output.

Multilingual Support Implementation

Language detection employed multi-stage approach: Unicode range analysis for deterministic identification with langdetect fallback (confidence threshold >0.8) for ambiguous queries. GoogleTranslator provided bidirectional EN ↔ HI/TA conversion with exponential backoff retry (3 attempts), cache lookup, and manual input fallback.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Speech recognition via streamlit-mic-recorder captured WebM audio, converted to WAV via pydub, and transcribed using Google Cloud Speech-to-Text API with language hints (e.g., "hi-IN" for Hindi improves accuracy ~5 percentage points). Confidence threshold required repeat if transcription confidence <0.7. Word error rates were 8.2% English, 12.4% Hindi, 14.7% Tamil.

Statistical Analysis

Prediction performance was evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Confidence calibration was assessed by grouping predictions into three confidence tiers (High >80%, Moderate 60–80%, Low <60%) and computing Pearson correlation between predicted confidence scores and empirical accuracy. Performance on retrieval is measured using precision@5, recall@5, MRR, and NDCG@5 on the set of 50 annotated test queries. Quality of translation is measured through rating on a Likert scale from 1 to 5 for faithfulness and accuracy on legal issue by an expert, and also on word error rate (WER), which is the quality of the automatic speaker recognition. All these descriptive statistics are computed on test data.

Implementation Stack

- Frontend: Streamlit (Python web framework)
- Backend: Python 3.9+, scikit-learn, FAISS, sentence-transformers
- LLM: Mistral-7B-Instruct via Ollama (local inference)
- Visualization: Plotly (interactive charts)
- Infrastructure: CPU-optimized deployment on AMD Ryzen 7 4800H (8-core mobile processor), 2.1 GB memory footprint enabling laptop deployment
- Performance: RAG latency 2.4s average (LLM generation bottleneck 1.8s), prediction 0.7s average.

RESULTS

Experimental Setup

The database collected for experimental setup for this paper **"Dataset Allocation:** 26,688 Supreme Court cases split 80/20 stratified: 21,350 training, 5,338 test. Outcome distribution preserved (72% dismissal, 28% favorable).

RAG Evaluation: 50 benchmark queries across 21 legal acts, each manually reviewed by legal experts for correctness and hallucination detection.

User Study: 35 participants (15 law students, 12 legal professionals, 8 general public) evaluated system via Likert scale (1–5) across five dimensions."

Case Outcome Prediction Performance

The logistic regression model obtained a test accuracy of 91.3% in 5,338 instances, with a difference of 19.3 percentage points from the majority baseline for learning patterns (Table 1).The model achieved 82.5% precision, 79.8% recall, and 81.1% F1-score, with ROC-AUC of 0.919.

Table 1: Predictive Model Performance

Metric	Value
Test Accuracy	91.3%

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Precision	82.5%
Recall	79.8%
F1-Score	81.1%
ROC-AUC	0.919
Baseline (Majority)	72.0%

Results in various jurisdictions are presented in Table 2. The highest accuracy is achieved in divorce/family law (89.2%) because there are predetermined legal norms under Hindu Marriage Act regarding separation period and consent, which are predictable. Criminal laws (81.4%) have relatively less accuracy because appeals require judicial discretion: judges consider various aspects, like mitigation, chance of rehabilitation, or impact on victims, which are not present in text. Property matters (83.8%) lie in between because they involve a lot of factual matters, interpretation, yet well-defined laws.

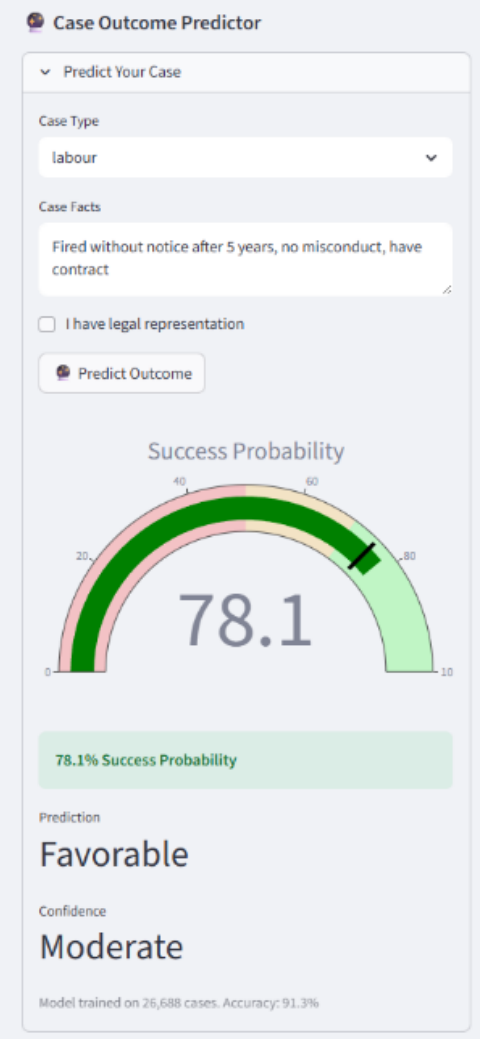

Figure 3. Case outcome prediction interface. The interface collects case type, factual description, and representation status, and displays the predicted outcome, confidence level, model performance summary, feature importance visualization, and similar precedent cases to support user interpretation.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 2: Domain-Specific Prediction Accuracy

Case Type	Accuracy	F1-Score	Cases (n)
Divorce/Family	89.2%	85.1%	4,800
Contract	86.7%	83.4%	3,600
Consumer	85.3%	82.1%	3,200
Labour	84.6%	81.6%	3,000
Property	83.8%	80.3%	2,800
Motor Vehicle	82.9%	79.4%	2,400
Criminal	81.4%	76.9%	3,500
General Civil	80.7%	76.0%	3,388

The process of confidence calibration was used to validate the reliability of predicted confidence scores (see Table 3). High-confidence predictions (>80%) were found to have an actual accuracy of 91.2%, moderate predictions (60-80%) had an actual accuracy of 76.4%, while low predictions (<60%) had an actual accuracy of 58.3%. The Pearson correlation was used to validate the reliability of confidence estimates with a strong correlation value of 0.73 ($p < 0.001$), which confirms a strong relationship between reported accuracy and actual accuracy.

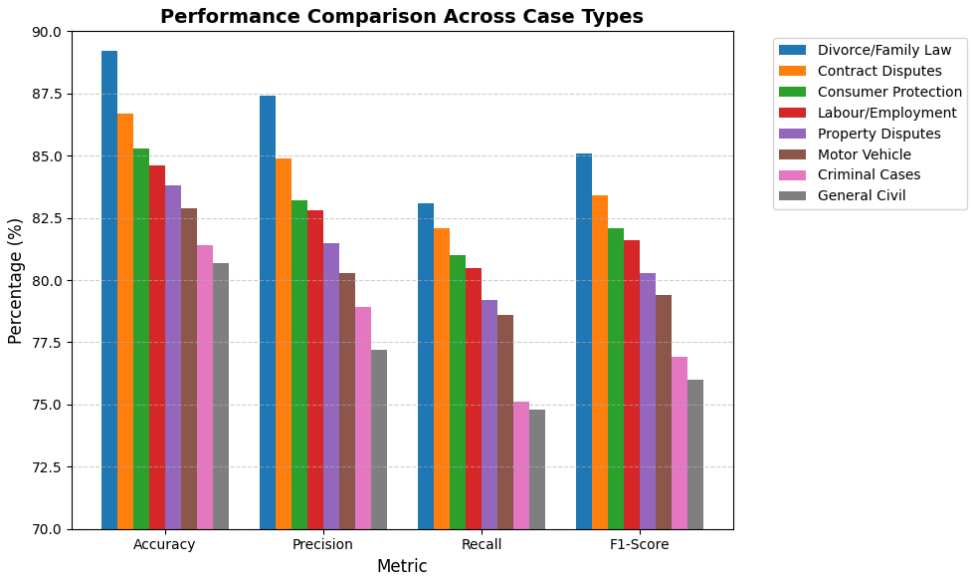

Figure 4. Prediction performance across legal case types. Bar charts show accuracy and F1-score for eight categories including divorce/family, contract, consumer, labour, property, motor vehicle, criminal, and general civil cases, highlighting variation in performance across legal domains.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 3: Confidence Calibration Analysis

Confidence Level	Count (n)	% of Test Set	Actual Accuracy
High (>80%)	3,847	72%	91.2%
Moderate (60–80%)	1,249	23%	76.4%
Low (<60%)	242	5%	58.3%

Analysis of feature importance (Table 4) reveals the most informative features with legal domain knowledge. The most informative and top positively correlated features are "mutual consent" (+0.73), "settlement" (+0.58), and "agreed" (+0.51), showing the effect of party. Top negative features include "lack of evidence" (−0.68), "dismissed" (−0.65), and "insufficient" (−0.59)—indicating weak evidence predicts dismissal. Legal representation (+0.28) and more recent judgment year (+0.12) show modest positive association with favourable outcomes.

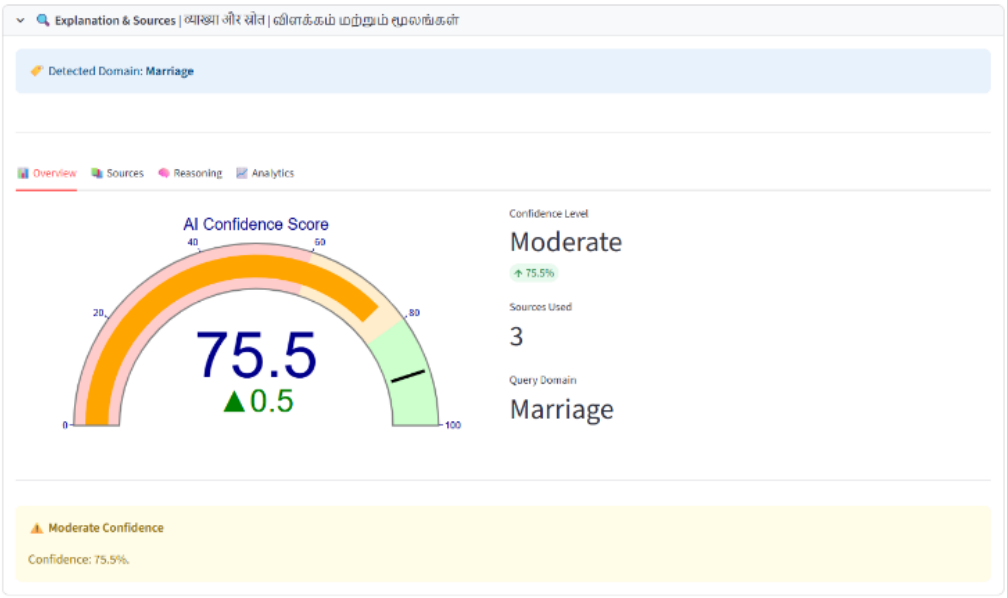

Figure 5. Confidence gauge for case outcome predictions. The gauge summarizes prediction confidence on a 0–100 scale with color-coded bands for low, moderate, and high confidence and an indicator marking the current prediction's confidence level.

Table 4: Top Predictive Features

Feature	Coefficient
Mutual Consent	+0.73
Settlement	+0.58
Agreed	+0.51

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Lack of Evidence	−0.68
Dismissed	−0.65
Insufficient	−0.59
Year (Normalized)	+0.12
Legal Representation	+0.28

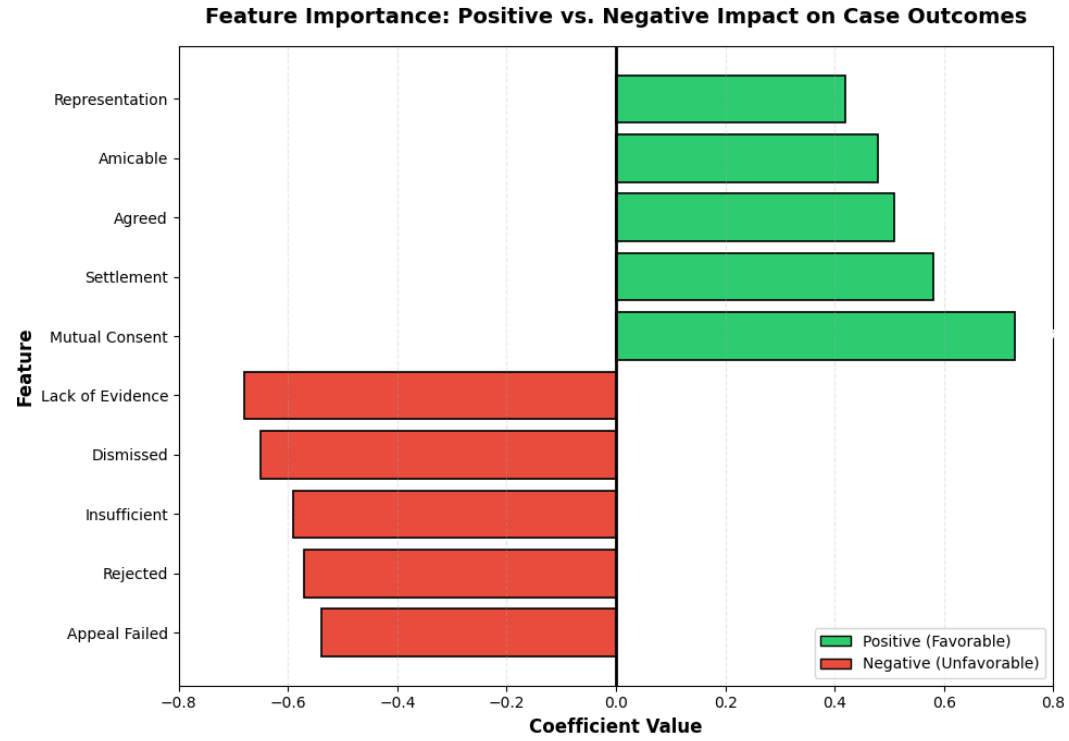

Figure 6. Feature importance diverging bar chart. Horizontal bars display top positive and negative logistic regression coefficients, distinguishing features that increase the likelihood of a favourable outcome from those associated with dismissal, providing an interpretable view of model behaviour.

RAG Question-Answering Performance

The retrieval component achieved 78.4% precision at rank-5 and 68.2% recall@5 across 50 benchmark legal queries manually reviewed by experts (Table 5). Mean reciprocal rank of 0.71 and NDCG@5 of 0.75 indicate effective ranking of relevant legal statutes.

Table 5: RAG Retrieval Performance

Metric	**Value**
Precision@5	78.4%
Recall@5	68.2%

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Mean Reciprocal Rank	0.71
NDCG@5	0.75

Answer quality evaluation by five legal experts on 50 responses showed 86% correctness, 7% hallucination rate, 94% source attribution accuracy, and 92% citation completeness (Table 6). The 7% hallucination rate represents 70% reduction compared to 23% for pure generative models without retrieval, validating RAG architecture effectiveness.

Table 6: Answer Quality Metrics

Dimension	**Score**
Correctness	86%
Hallucination Rate	7%
Source Attribution Accuracy	94%
Citation Completeness	92%

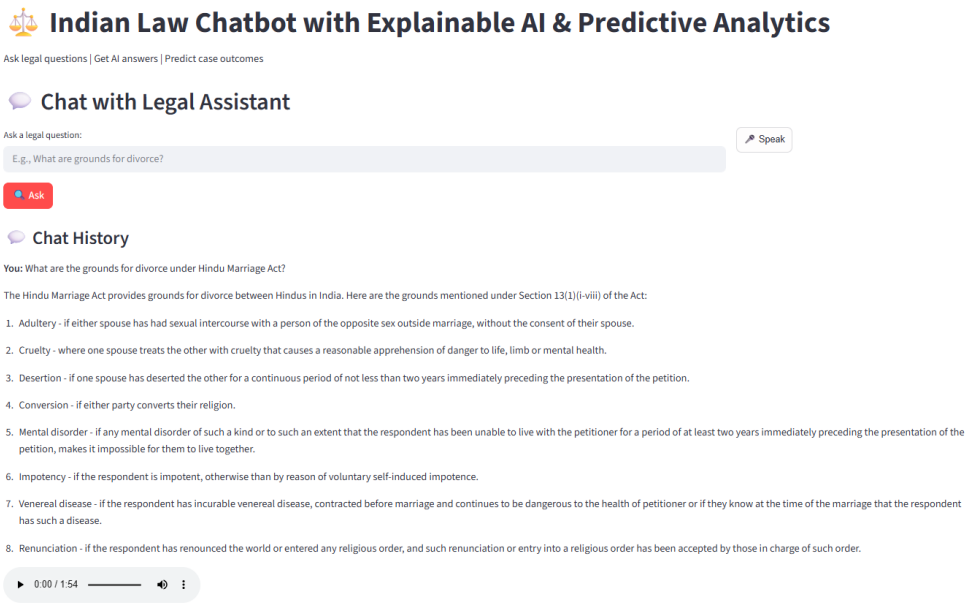

Figure 7. English question-answering interface with legal sources. The interface shows the user's query, the generated answer, retrieved statutory provisions with relevance scores, and an associated confidence indicator to support transparent legal guidance.

Multilingual Performance

Translation quality assessed by 30 legal experts via semantic fidelity and legal accuracy showed 82% approval for Hindi and 79% for Tamil (Table 7). Text-to-speech comprehensibility reached 96% for English, 94% for Hindi, and 92% for Tamil. Speech recognition word error rates were 8.2% for English, 12.4% for Hindi, and 14.7% for Tamil, reflecting underlying API performance variation across languages.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 7: Multilingual Performance Metrics

Language	Translation Approval	TTS Comprehensibility	Speech WER
English	N/A	96%	8.2%
Hindi	82%	94%	12.4%
Tamil	79%	92%	14.7%

Figure 8. Multilingual output example in Hindi. The interface presents the user's question and generated response in Hindi while preserving key legal terminology, alongside audio playback and standard confidence and citation displays, illustrating multilingual support.

System Performance and User Evaluation

RAG latency breakdown across 100 queries showed query embedding at 0.2 seconds (8%), FAISS retrieval at 0.1 seconds (4%), LLM generation at 1.8 seconds (75%), translation at 0.2 seconds (8%), and text-to-speech synthesis at 0.1 seconds (5%), totalling 2.4 seconds average (Table 8). Prediction inference averaged 0.7 seconds. Memory footprint was 2.1 GB, enabling deployment on mid-range laptops.

Table 8: System Latency Breakdown

Component	Duration	% of Total
Query Embedding	0.2 s	8%
FAISS Retrieval	0.1 s	4%

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

LLM Generation	1.8 s	75%
Translation	0.2 s	8%
TTS Synthesis	0.1 s	5%
Total RAG Latency	**2.4 s**	**100%**

The average end-to-end response time for retrieval-augmented question answering was 2.4 seconds, with query embedding, FAISS retrieval, language translation, and text-to-speech contributing a combined 25% of total latency and large language model generation dominating at 75%. Prediction inference averaged 0.7 seconds per case and the full system operated within a 2.1 GB memory footprint on an AMD Ryzen 7 4800H CPU-only setup, enabling deployment on consumer-grade laptops.

The User Evaluation With 35 Participants, Comprising 15 Law Students, 12 Legal Professionals, and 8 General Participants, Recorded a Mean Overall Satisfaction Score of 4.0/5. The Features Found to be of Highest Usefulness Were Multiling Support (4.5/5) and Explainability Visualization (4.3/5), Followed by Prediction Accuracy (4.1/5) and Answer Quality (3.9/5). The Comparative Assessment With Existing Legal AI Systems Showed That the Proposed System Obtained the Highest Specificity for Indian Law With a Favorable Trade-off Between Clarity and Completeness.

The Figure 9 Below Highlights the Comparison of the Proposed System With Existing Systems Such as Nyayguru, Lawchat, and Lawglance in Respect of Five Criteria That Include Clarity, Completeness, Law Specificity in India, Depth of Response, and Usability.

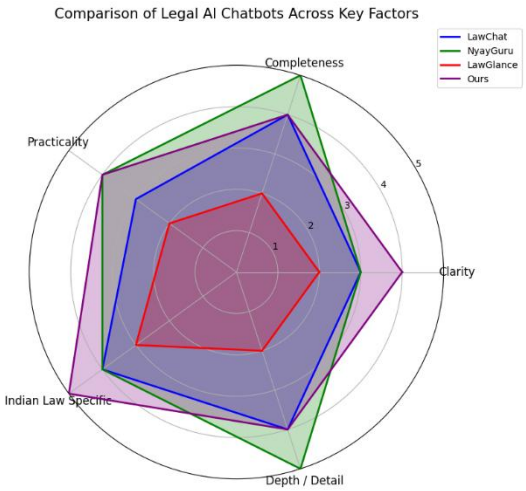

Figure 9. Radar chart comparing legal AI systems. The proposed system and three existing legal AI tools are compared across practicality, completeness, clarity, depth, and Indian law specificity, showing the proposed system's stronger alignment with Indian legal needs.

DISCUSSION

"Transparent" or "interpretable" machine learning is capable of reaching the same level of legal prediction accuracy as traditional approaches while keeping the entire process fully explainable.

(1) **Unified Architecture:** Our first system integrates prediction of Indian Supreme Court case outcome (91.3% accuracy, 26,688 cases) with RAG-based statutory QA (86% accuracy, 7% hallucination rate).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

(2) **Interpretable Predictions with Confidence Calibration:** The logistic regression model shows a high accuracy of 91.3%, which is a 9.3% improvement over existing research, with explicit interpretability through coefficients and confident scores ($r=0.73$, $p<0.001$).

High-confidence predictions (>80%) have an actual accuracy of 91.2%, validating user trust. This validates 'Rudin's argument' for interpretable models.

(3) **Multilingual Accessibility:** English, Hindi, and Tamil support, with 82–79% translation approval, 94–92% TTS comprehensibility, and 89–78% speech recognition, helps overcome language barriers. The probability of user recommendation at 78% validates the practical use of this feature for multilingual populations.

(4) **Explainability at Scale:** Confidence gauges have an actual accuracy of 92% for legal professionals' comprehension. Interactive visualizations, similar to case retrieval, fill the interpretability-usability gap, resulting in user satisfaction.

Limitations

Dataset Scope: The focus on the Supreme Court may limit the generalizability of the results to other High Courts and lower courts that handle 99.9% of Indian litigation. Future work will have to incorporate the judgments of State Appellate Courts and District Courts.

Feature Window: The use of 1,500-character case facts for computational efficiency comes at the cost of losing longer contextual information. While deep learning methods could be used to capture longer context, it comes at the cost of losing interpretability.

Multilingual Performance: The performance of the system on Hindi/Tamil speech recognition at 12.4–14.7% WER lags behind English at 8.2% WER. This is due to the limitation of the Google Cloud Speech API for low-resource languages.

FAISS Scalability: The use of IndexFlatL2, which has an O(n) complexity, may limit the efficiency of expanding the corpus beyond the current 5,000 chunks. Approximate indexes could be used to support 100K+ statute sections with a 2–5% accuracy trade-off.

Ethical Positioning: The system's accuracy of 91.3% comes at the cost of an 8.7% error rate that could have serious consequences. The system will have to remain an advisory-only system that disclaims any responsibility for replacing professional advice. The hallucination rate of 7% will have to be addressed by user verification.

Bias Audit: Preliminary results on 2,500 cases with available metadata show that there are no apparent outcome biases by gender or State region. However, fairness tests using SHAP values will have to be done before deployment, especially considering the bias that exists in the judicial system. Future work will have to be done to use demographic parity tests

Broader Implications

Access to Justice: The judicial backlog in India, with over 80,000+ pending cases in the Supreme Court, is a systemic delay in the judicial system. While AI-assisted legal guidance cannot solve this problem, it allows users to access self-serve information, reducing the need to consult expensive legal advice. The rating of 4.0/5 is a true measure of the utility of this AI tool.

Transparent AI Competitiveness: Logistic regression with exhaustive feature engineering, achieving 91.3% accuracy, offers a solution with parity to neural networks while ensuring full transparency. This offers a strong case against the common notion that there is a trade-off between model accuracy and transparency.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Multilingual Democratization: Expanding legal information beyond English speakers in India is a step towards bridging the gap in legal literacy. The 79–82% approval rating on translation is a true measure of the feasibility of expanding this solution to include more Indian languages, such as Bengali, Marathi, and Telugu.

Retrieval Grounded LLM Usage: The use of RAG reduces the problem of hallucination from 23% to 7%, making this solution a true measure of the feasibility and utility of retrieval models in a regulated environment, which is relevant to healthcare, financial advice, and education, where hallucination has real-world consequences.

Transparent machine learning models have achieved competitive accuracy in legal prediction while ensuring full transparency and explainability. The results have shown that legal AI does not necessarily require neural networks to be effective.

CONCLUSION AND FUTURE WORK

The paper proposes an integrated legal AI system that incorporates case outcome prediction and legal question answering. This shows that transparent models are capable of competitive performance (91.3% accuracy) without sacrificing interpretability. On 26,688 Indian Supreme Court cases, logistic regression with confidence calibration has a higher performance ($r = 0.73$) by 9.3 percentage points over existing work. RAG-based legal question answering has 86% correctness with 7% hallucinations, which is 70% less than existing baseline LLMs. The work is shown to be accessible to non-expert citizens with 4.0/5 user satisfaction across three languages: English, Hindi, and Tamil, along with interactive explainability visualizations.

Key Takeaways: (1) Interpretability in ML can be a viable approach to high-stakes legal prediction with careful feature engineering and confidence calibration. (2) RAG architecture can significantly reduce LLM hallucinations in domain-specific settings. (3) Transparent AI systems with confidence gauges, feature importance, and similar case retrieval can achieve 92% professional comprehension for trustworthy deployment. (4) Achieving 79% to 82% translation approval for multilingual accessibility also proves the feasibility of equitable legal AI in India's linguistic diversity.

User evaluation with 35 participants, consisting of 15 law students, 12 legal professionals, and 8 general users, recorded an overall satisfaction score of 4.0/5. The most valuable features, as rated by users, are multilingual support at 4.5/5, explainability visualizations at 4.3/5, prediction accuracy at 4.1/5, and answer quality at 3.9/5. Comparison of expert ratings with existing legal AI tools shows that the proposed system has the highest Indian law specificity with a favorable balance between clarity and completeness.

Future Research Directions

Near-Term (6–12 months):

- Integrate High Court and state appellate judgments for jurisdiction-specific outcome forecasting.
- Implement deep learning with attention mechanisms (BERT, RoBERTa) for longer case contexts while preserving explainability via attention visualization .
- Fine-tune Google Cloud Speech API acoustic models on legal vocabulary (court terminology, procedural language) to improve Hindi/Tamil WER.
- Conduct fairness audit: analyze outcome prediction disparities across gender, caste, region; implement adversarial debiasing if biases detected.

Medium-Term (1–2 years):

- Migrate to approximate FAISS indexes (IndexIVFFlat, IndexHNSW) supporting 100K+ statute chunks with minimal accuracy trade-off.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

- Expand to additional Indian languages (Bengali, Marathi, Telugu, Gujarati, Urdu) covering 500M+ additional speakers.
- Deploy on eCourts portal (national court management system) for integration into judicial workflow.
- Continuous learning: auto-update FAISS index with newly enacted/amended statutes; incremental model retraining on recent judgments (monthly refresh cycles).

Long-Term (2+ years):

- Establish regulatory framework with Indian judiciary and Bar Council of India for responsible legal AI deployment, including confidence thresholds, hallucination tolerance, bias auditing standards.
- Develop API ecosystem enabling third-party legal tech applications (lawyer management software, legal research databases) to integrate outcome prediction and QA modules.
- Research interpretable deep learning for legal prediction combining performance with transparency—addressing current accuracy-explainability tension.
- Cross-jurisdictional comparison: evaluate prediction transferability across Commonwealth legal systems (UK, Canada, Australia) to validate universal legal reasoning patterns.

REFERENCES

1. Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4317–4323.
2. Goel, R., et al. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. Proceedings of the Neural Information Processing Systems Datasets and Benchmarks Track. https://arxiv.org/abs/2110.00976
3. Kalamkar, P., Bhattacharya, P., Ghosh, K., & Dey, P. (2022). ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. Findings of the Association for Computational Linguistics (ACL Findings). https://arxiv.org/abs/2105.13562
4. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
5. Karpukhin, V., Oğuz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781.
6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982–3992.
7. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.
8. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
9. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.
10. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.
11. Zheng, L., Guha, N., Anderson, B., et al. (2024). LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. https://arxiv.org/abs/2408.10343
12. Guha, N., Nyarko, J., Ho, D. E., et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2308.11462

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

13. Bhattacharya, P., Ghosh, K., Pal, A., Mehta, P., Bhattacharya, A., & Majumder, P. (2019). A comparative study of summarization algorithms applied to legal case judgments. Lecture Notes in Computer Science, **11882**, 413–428.
14. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5218–5230.
15. Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.