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. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153104

Hybrid AI-Powered Legal Assistant for BNS Laws Using Structured Retrieval and Contextual Language Model Generation

Omkar Rajaram Barad[1], Atharava Atul Desai[2], Atharav Sachchidanand Bhogate[3], Saurabh Namdev Dhokare[4], Prof. (Dr.) Suresh Shirgave[5]

CSE(AI), D.KTE'S Society's Textile and Engineering Institute, Ichalkaranji[1-4]

Guide, CSE(AI), D.KTE'S Society's Textile and Engineering Institute, Ichalkaranji[5]

ORCID: 0009-0001-3476-4469

Abstract: Access to accurate legal information remains a major challenge for non-experts due to the complexity of statutory language and the evolving structure of Indian criminal law. This study proposes a hybrid AI-powered legal assistant that combines deterministic statutory retrieval with contextual large language model generation to provide reliable and interpretable legal explanations. The system integrates a structured legal database containing codified criminal statutes with a contextual explanation engine powered by a locally deployed large language model (LLaMA 3). A dual-layer architecture was implemented: an offline rule-based retrieval module for precise statutory matching and an online generative module for extended legal explanation. Query preprocessing includes keyword expansion, stemming, and multilingual normalization. Performance was evaluated using a curated dataset of 250 legal queries covering direct section lookups and semantic intent-based questions. The offline module achieved high precision for direct statutory queries with low latency suitable for interactive use. The contextual module improved interpretability and semantic coverage for intent-based questions. The hybrid strategy reduced hallucination risk compared to standalone generation by grounding responses in verified statutory content. The proposed hybrid legal assistant provides a scalable and accurate framework for legal information access. By combining deterministic retrieval and contextual generation, the system mitigates hallucination risks associated with large language models while enhancing user comprehension. The framework is designed to be extensible to multilingual and cross-jurisdictional applications.

Index Terms—Artificial Intelligence, Legal Informatics, Hybrid Retrieval, Large Language Models, Indian Criminal Law, Multilingual NLP

I. INTRODUCTION

Recent advances in artificial intelligence have transformed information retrieval and human-computer interaction. However, legal assistance systems must satisfy unusually strict requirements for factual correctness, interpretability, and traceability. In India, criminal law spans multiple statutes and evolving reforms, including legacy references from the Indian Penal Code (IPC) and the introduction of newer frameworks such as the Bharatiya Nyaya Sanhita (BNS). Non-expert users often struggle to identify the correct provision, interpret punishments, and understand procedural implications.
While large language models (LLMs) can generate fluent explanations, they may produce unsupported content, which is unacceptable for legal guidance. Conversely, traditional search or rule-based systems can be precise but may fail to explain legal meaning in accessible terms. To address this gap, this paper proposes a hybrid architecture that separates responsibilities: (i) deterministic retrieval for authoritative statutory grounding and (ii) contextual generation for structured, user- friendly explanation.

The key contributions of this study are: (1) a unified structured dataset representation supporting modern and legacy references; (2) a two-mode legal chatbot (offline/online) integrating database retrieval with local LLM generation; (3) a multilingual query normalization pipeline for mixed-language inputs (English with Hindi/Marathi tokens); (4) a practical evaluation protocol for statutory lookup accuracy and interpretability; and (5) an implementation blueprint suitable for deployment in resource-constrained environments.

II. RELATED WORK

Legal AI research has progressed from expert systems and rule engines to neural retrieval and transformer-based models. Early systems focused on symbolic reasoning and encoding statutes as rules, offering determinism but limited flexibility. Modern approaches use dense embeddings and transformer encoders to capture semantic similarity for legal search, and

DOI: 10.17148/IJARCCE.2026.153104

generative models to synthesize explanations. Retrieval-Augmented Generation (RAG) has emerged as a practical pattern to ground generation in retrieved documents. Nevertheless, legal-domain deployment remains sensitive due to hallucinations, missing citations, and inadequate separation between retrieved facts and generated interpretation.

This work aligns with RAG principles but emphasizes determinism for statutory sections through structured retrieval against a curated database. The LLM is used primarily to (a) explain retrieved provisions, (b) summarize implications, and (c) present step-by-step guidance, while being constrained to provided context and instructed to avoid inventing new sections or punishments.

Recent studies have explored retrieval-augmented generation for legal reasoning and statutory interpretation. Legal-domain RAG systems have shown improved factual grounding compared to standalone LLMs, particularly in compliance- sensitive domains. Emerging multilingual legal NLP research also emphasizes accessibility for non-English speakers, especially in jurisdictions with diverse linguistic usage. However, few systems combine deterministic statutory retrieval with grounded generation optimized for low-resource environments, which this work addresses.

III. SYSTEM ARCHITECTURE

The system consists of four layers: (i) User Interface Layer; (ii) Query Processing Layer; (iii) Legal Retrieval Engine; and (iv) Contextual Explanation Engine. The design follows separation of concerns to ensure maintainability and to reduce failure coupling between retrieval and generation.

A. Offline Deterministic Retrieval Module

The offline module resolves queries using a structured database of legal provisions. Each record contains the act name, section number (when applicable), legacy act and legacy section (for reference mapping), heading, short summary, legal text, keywords, category, and punishment metadata.

Direct section queries (e.g., “IPC 420”, “section 51 IPC”) are matched using normalized patterns with controlled fallbacks. This module returns fast, authoritative responses suitable for repeated queries and low-connectivity settings.

B. Online Contextual Explanation Module

The online module uses a locally deployed LLM (LLaMA 3 via Ollama) to generate detailed explanations. To minimize hallucinations, the system first retrieves relevant statutory content from the database and then injects it into a carefully constrained prompt. The generated response is expected to: (a) restate the provision, (b) interpret in plain language, (c) summarize punishment, (d) describe common real-world scenarios, and (e) highlight limitations and when to consult a lawyer.

C. Hybrid Interaction Flow

Upon receiving a user query, the system performs preprocessing, detects whether a direct section reference exists, and retrieves matching records. If the user is on the offline tab, the system formats the record(s) into a concise result. If on the online tab, the system provides the same grounded context to the LLM for an expanded explanation. Both paths share the same retrieval foundation to guarantee factual consistency, while the online path adds interpretative depth.

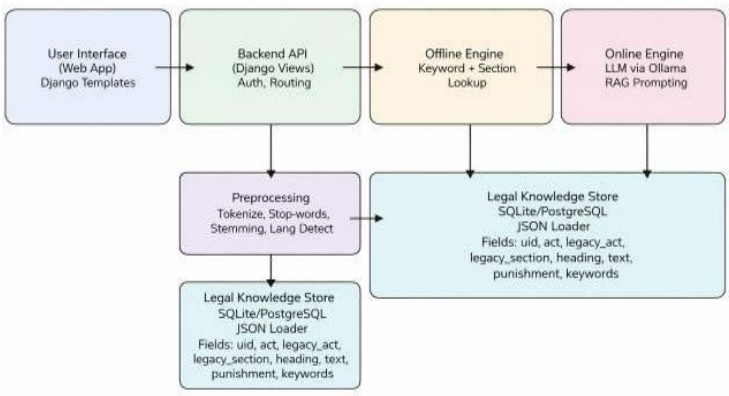

Fig. 1. End-to-end query flow from user input to offline retrieval and online grounded generation.

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. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153104

IV. METHODOLOGY

A. Data Model and Storage

The legal knowledge base is represented as a structured JSON dataset and imported into a relational database. A unified schema supports both current acts and legacy references by storing act, section (nullable), and legacy mappings. Punishment is stored using normalized fields (punishment text, punishment imprisonment, punishment fine possible) to enable consistent formatting. Data integrity constraints include unique identifiers (uid) per record and normalized act codes. This enables robust updates, deduplication, and future extension across statutes.

B. Query Preprocessing and Multilingual Normalization

User queries are normalized through lowercasing, punctuation removal, stop-word filtering, and stemming. A domain keyword map expands common verbs and colloquial terms into legal intent terms (e.g., “hit” → “assault”, “steal” → “theft”, “threat” → “criminal intimidation”).

To support mixed-language inputs where users type English sentences with Hindi/Marathi tokens, the system can optionally apply lightweight language detection and transliteration/translation for a restricted vocabulary of legal terms (e.g., “saza” → “punishment”, “jamaanat” → “bail”, “FIR”, “HRT” → “court”). This improves retrieval coverage without requiring fully bilingual embeddings.

C. Large Language Model Usage and Disclosure

A locally deployed LLM (LLaMA 3) is used only for generating explanations from retrieved statutory context. The system explicitly instructs the model to avoid fabricating section numbers, punishments, or legal outcomes not present in the provided context. Human authors remain fully accountable for the content, system design, and final manuscript.

V. EXPERIMENTAL SETUP AND RESULTS

A. Query Set

A test set of 250 queries was curated to reflect realistic legal questions. The dataset includes: (i) direct section lookups, (ii) intent-based semantic questions, and (iii) mixed-language queries. Each query was annotated with expected statutory references and response requirements.

B. Metrics

We report precision, recall, and accuracy for statutory retrieval, along with average response time and a human interpretability score (1–5) for generated explanations.

C. Results

Performance metrics for the offline retrieval module and online grounded explanation module are presented in Tables I and II, respectively.

Overall, the hybrid approach improves semantic cover- age while maintaining factual grounding. Compared to standalone generation, the grounded approach reduced unsupported claims by constraining the model to retrieved legal context.

TABLE I
OFFLINE RETRIEVAL MODULE PERFORMANCE

Metric	Value
Precision (direct section queries)	0.96
Recall (direct section queries)	0.94
Average response time	0.42 s

TABLE II
ONLINE GROUNDED EXPLANATION MODULE PERFORMANCE

Metric	Value
Accuracy (intent-based queries)	0.89
Average response time	2.8 s
Interpretability (1–5)	4.4

VI. DISCUSSION

The hybrid design separates factual retrieval from ex- planation generation. This improves reliability and allows the offline module to operate independently in low-resource environments. The online module primarily enhances

DOI: 10.17148/IJARCCE.2026.153104

interpretability through plain-language summaries and stepwise guidance. Nevertheless, generation remains sensitive to prompt design, context length, and dataset quality. Future improvements should include stronger citation behavior and expanded multilingual normalization.

VII. CONCLUSION

This paper presented a hybrid AI-powered legal assistant for Indian criminal law that combines deterministic statutory retrieval with grounded contextual generation. The system demonstrates strong accuracy for direct statutory queries and improved interpretability for intent-based questions. The architecture is designed for practical deployment and extension across additional statutes and languages.

VIII. FUTURE WORK

Future work includes extending coverage to civil and constitutional law, integrating judgment and precedent retrieval, deploying mobile clients, implementing multilingual embedding- based retrieval, and adding standardized evaluation bench- marks.

DECLARATIONS

Funding
No funding was received for conducting this study.
Competing Interests
The authors declare that they have no relevant financial or non-financial interests to disclose.
Ethics Approval
This study does not involve human participants, animals, or identifiable personal data. Ethical approval was not required.
Author Contributions
All authors contributed equally to the conception, design, implementation, experimentation, and writing of this work.

Data Availability
The statutory texts used are derived from publicly available legal sources. The processed dataset and query set can be made available by the corresponding author upon reasonable request, subject to institutional and legal constraints.

REFERENCES

[1]. K. D. Ashley, *Artificial intelligence and legal analytics*. Cambridge: Cambridge University Press, 2017.
[2]. M. J. Bommarito and D. M. Katz, "A mathematical approach to the study of the United States Code," *Physica A*, vol. 389, no. 19, pp. 4195–4200, 2010. DOI: 10.1016/j.physa.2010.05.057
[3]. D. M. Katz, M. J. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," *PLoS ONE*, vol. 12, no. 4, p. e0174698, 2017. DOI: 10.1371/jour- nal.pone.0174698
[4]. I. Chalkidis and D. Kampas, "Deep learning in law: early adaptation and legal text processing," in *Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL)*, 2019, pp. 1–10.
[5]. I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," *Artificial Intelligence and Law*, vol. 29, pp. 273– 299, 2021. DOI: 10.1007/s10506-021-09284-x
[6]. [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[7]. [7] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
[8]. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[9]. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in *Proc. ICML*, 2020.
[10]. G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in *Proc. EACL*, 2021.
[11]. W. Shi, S. Min, D. Grangier, and W. Yih, "REPLUG: Retrieval- augmented black-box language models," in *Proc. ACL*, 2023.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026
DOI: 10.17148/IJARCCE.2026.153104

[12]. Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. DOI: 10.1145/3571730
[13]. H. Zhong, C. Xiao, C. Tu, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proc. EMNLP*, 2018.
[14]. M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," *Artificial Intelligence and Law*, vol. 28, pp. 237–266, 2020. DOI: 10.1007/s10506- 019-09255-y
[15]. H. Touvron et al., "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[16]. D. Jurafsky and J. H. Martin, *Speech and language processing*, 3rd ed. (draft). Prentice Hall, 2023.
[17]. C. D. Manning, P. Raghavan, and H. Schu¨tze, *Introduction to informa- tion retrieval*. Cambridge: Cambridge University Press, 2008.
[18]. P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in NLP," in *Proc. ACL*, 2020.
[19]. D. M. Katz, M. J. Bommarito, and J. Blackman, "GPT-4 passes the bar exam," *SSRN Electronic Journal*, 2023. DOI: 10.2139/ssrn.4389233
[20]. Committee on Publication Ethics, "Core practices," 2017. [Online]. Available: https://publicationethics.org/core-practices

