Defend U: AI Powered Legal Advisor for The Indian Legal Context

Abhijith R Nair[1], Alan Jose[2], Alan Joseph[3], Albin Jose[4] and Aswathy R[5]
[1,2,3,4]*UG Student, Department of Computer Science & Engineering, St. Thomas College of Engineering and Technology, Chengannur, Kerala, India*
[5]*Assistant Professor, Department of Computer Science & Engineering, St. Thomas College of Engineering & Technology, Chengannur, Kerala, India*

***Abstract*—Defend U is an AI-powered legal intelligence platform developed to bridge the critical gap between legal complexity and citizen access in India. Unlike conventional legal information retrieval systems, Defend U not only provides real-time search and retrieval of relevant Indian case laws and statutes but also delivers AI-driven summarization, semantic analysis, and structured insights from judicial documents. The system integrates a FastAPI backend, FAISS-based semantic search, Sentence-BERT for query encoding, lexical keyword search with section boosting, and Reciprocal Rank Fusion for hybrid retrieval, followed by Cross-Encoder re-ranking using the ms-marco model to maximize retrieval precision. A Neo4j knowledge graph maps relationships between legal acts, sections, and precedents. For pro-se users, the platform generates actionable legal roadmaps, identifies applicable IPC sections, and provides multilingual support for over 10 Indian languages through optimized batch translation. Built on an asynchronous high-performance architecture using the Llama 3.3 70B model via Groq Cloud, Defend U offers a reliable, scalable, and comprehensive solution that combines automated legal document analysis, knowledge graph navigation, and seamless connectivity for citizens navigating the Indian legal system.**

***Index Terms*—Artificial Intelligence, Cross-Encoder Re-ranking, FAISS, FastAPI, Indian Legal System, Knowledge Graph, Large Language Models, Legal Judgment Summarization, Natural Language Processing, Neo4j, Pro-Se Assistance, Reciprocal Rank Fusion, Retrieval-Augmented Generation, Semantic Search, Sentence-BERT.**

I. INTRODUCTION

Consistent access to legal information is vital for achieving fair outcomes in legal proceedings, particularly for individuals who cannot afford professional legal representation. Non-access to legal guidance remains one of the most persistent challenges in the Indian justice system, frequently leading to uninformed self-representation, violation of rights, and unjust outcomes [1].

Conventional legal platforms such as Indian Kanoon provide raw case data without structured summaries or simplified insights, making them largely inaccessible to non-specialists. While these tools offer document retrieval, they fail to extract relevant legal concepts, map applicable statutes, or guide users toward actionable steps in their specific legal situation [2].

Such systems typically lack advanced functionalities like hybrid semantic-lexical search, AI-driven summarization, cross-encoder re-ranking, pro-se assistance, and multilingual support. The Government of India has acknowledged that a significant portion of the population lacks access to legal aid, with the Legal Services Authorities Act providing limited reach to rural and economically disadvantaged communities [3].

To bridge these gaps, Defend U has been developed as an AI-powered legal intelligence platform integrating a novel hybrid retrieval pipeline: FAISS-based semantic search and lexical keyword search with section boosting are fused using Reciprocal Rank Fusion, followed by Cross-Encoder re-ranking using the ms-marco model to maximize retrieval precision. The system delivers relevant legal insights through this multi-stage pipeline, an AI-driven summarization

module, and a structured Neo4j knowledge graph [4]. Moreover, Defend U enhances legal accessibility by supporting over 10 Indian languages using optimized batch translation, enabling citizens from diverse linguistic backgrounds to interact with the platform in their native language and receive legally grounded guidance [5].

II. LITERATURE REVIEW

Non-access to legal information represents a significant challenge in contemporary India, especially affecting economically disadvantaged citizens, rural communities, and individuals navigating complex legal proceedings without professional support. Research indicates that failure to access legal guidance results in poor case outcomes, violation of rights, and increased dependency on informal and often unreliable legal advice [6].

Gao et al. [1] introduced LSDK-Legal Sum, a legal judgment summarization framework that integrates logical document structure and domain knowledge to improve summary quality. The method divides legal judgments into four logical components Type, Claim, Fact, and Result and generates summaries using a block-interactive learning strategy, demonstrating significantly higher accuracy than traditional summarization models on the CAIL-2020 dataset.

Bellandi et al. [2] proposed a knowledge-based approach for legal document retrieval combining NLP techniques with Sentence-BERT document embedding models and zero-shot classification, enabling effective retrieval without requiring annotated training data. Fan et al. [3] introduced an ensemble learning-based system for legal document similarity matching using BERT-based sub-networks, achieving 74.53% accuracy on the CAIL2019-SCM dataset through complementary model combination and data augmentation.

Recent innovations have introduced Large Language Models and judicial syllogism-based reasoning to address legal summarization [7]. LLM-based platforms aim to follow the logical reasoning process used by judges identifying applicable law, extracting key facts, and generating conclusions producing more interpretable and coherent legal summaries [8].

Guitton et al. [4] analyzed AI adoption in 28 advanced democracies, demonstrating that judicial AI integration is driven by court workload, political orientation, and technology diffusion from neighboring systems. Song et al. [5] proposed judicial syllogism-based summarization using LLMs with chain-of-thought reasoning, demonstrating competitive performance on ROUGE, BLEU, and BERTScore metrics even with limited training data.

Despite these advancements, existing systems present several limitations. Most solutions focus on summarization or retrieval in isolation, without combining semantic and lexical signals through rank fusion or applying cross-encoder re-ranking for precision optimization. Many systems do not incorporate integrated knowledge graphs, pro-se assistance features, or multilingual support [9]. Citizen-facing accessibility through mobile-friendly interfaces and regional language support remains largely unaddressed in existing legal AI platforms [10].

III. PROPOSED SYSTEM

The proposed Defend U system functions as an AI-powered legal intelligence and pro-se assistance platform, engineered to enhance legal accessibility and citizen empowerment in the Indian legal context. Unlike conventional retrieval systems, it not only searches and retrieves legal documents but also delivers AI-driven summarization, semantic analysis, and structured legal roadmaps, unifying automation, real-time knowledge retrieval, and citizen connectivity into a single platform [11].

Defend U employs an asynchronous high-performance architecture to efficiently manage system operations. The backend is developed using FastAPI (Python), ensuring low-latency processing of complex NLP and AI tasks. The Llama 3.3 70B model via Groq Cloud API provides state-of-the-art legal reasoning and document synthesis, while asyncio.gather enables parallel execution of local and external data fetches, reducing response times by up to 60% [12].

The core retrieval mechanism employs a three-stage hybrid search pipeline illustrated in Fig. 1. User queries are simultaneously processed by two parallel retrieval systems: FAISS-based semantic search retrieving the top-15 semantically similar documents using Sentence-BERT embeddings, and a lexical search engine performing keyword matching with section-level boosting. Results from both are merged using Reciprocal Rank Fusion (RRF), then passed to a Cross-Encoder re-ranker (ms-marco) that scores each

query-document pair jointly for maximum precision, producing the final Top-K results.

The system also incorporates a Neo4j knowledge graph module where extracted legal concepts are organized to enable traversal of relationships between legal acts, IPC sections, and case precedents. A KMeans clustering module groups related legal terms into concept clusters, providing users with a visual understanding of the legal domain relevant to their query [13].

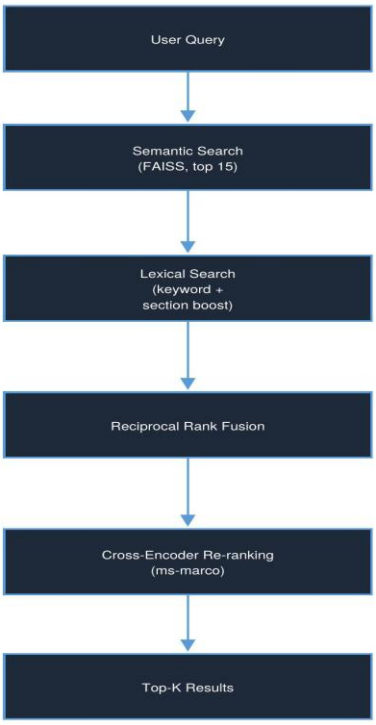

Fig. 1. Hybrid Search Architecture of Defend U

The entire system is linked to a cloud-based architecture, enabling real-time data processing and remote access via a web interface. An optimized batch translation module using deep-translator supports over 10 Indian languages in a single API call, minimizing network overhead while maximizing linguistic inclusivity.

IV. SYSTEM ARCHITECTURE

The Defend U system is conceived as an integrated AI framework linking NLP components, processing units, retrieval mechanisms, and communication interfaces to enable automated legal analysis and citizen guidance. At its core is a FastAPI backend that orchestrates all operations including hybrid search, AI summarization, knowledge graph queries, and multilingual translation.

The architecture is organized into four functional layers: (1) Input Layer comprising the web interface, document upload module (Tesseract OCR + PyMuPDF), and multilingual query interface; (2) Processing Layer FastAPI backend, SpaCy NLP pipeline, and Sentence-BERT query encoder; (3) Intelligence Layer FAISS semantic search, BM25 lexical search, Reciprocal Rank Fusion, Cross-Encoder re-ranking (ms-marco), Llama 3.3 70B LLM, Neo4j knowledge graph, and KMeans clustering; and (4) Output Layer structured legal summaries, IPC section mapping, precedent retrieval, and translated responses [14].

For legal document processing, the system utilizes a multi-modal ingestion pipeline accepting text queries, uploaded PDFs, and scanned images. Once processed, the system performs the three-stage hybrid retrieval (Fig. 1), concept extraction, and relationship mapping through Neo4j, delivering structured insights through the web interface. An LRU caching layer for expensive NLP operations ensures consistent response times under varying load conditions [15].

V. METHODOLOGY

The Defend U system integrates various software components and AI technologies to automate legal document analysis, hybrid semantic-lexical retrieval, cross-encoder re-ranking, and pro-se assistance. The system employs a structured pipeline where user queries are processed by the NLP engine, retrieved through hybrid search, re-ranked by the cross-encoder, and passed to the LLM for AI-generated insights [16].

The process begins with query processing, where users submit natural language queries through the web interface. The SpaCy pipeline tokenizes and expands the query with legal synonyms. Sentence-BERT encodes the expanded query into a dense vector embedding for semantic search, while simultaneously a lexical search engine indexes the same query using BM25-style keyword matching with section-level field boosting.

The semantic search module queries the FAISS index to retrieve the top-15 semantically similar legal documents. Concurrently, the lexical search module retrieves keyword-matching documents. Both ranked

lists are fused using Reciprocal Rank Fusion (RRF), a score-independent rank aggregation method assigning combined relevance scores based on position in both rankings [17].

The fused candidate list is then passed to a Cross-Encoder re-ranker (ms-marco). Unlike bi-encoder approaches, the cross-encoder processes the query and each candidate document jointly in a single transformer forward pass, enabling fine-grained relevance scoring. The top-K re-ranked results are passed to the Llama 3.3 70B model for Hybrid RAG-based summarization following a judicial syllogism framework.

Simultaneously, the Neo4j knowledge graph module extracts relationships between identified legal concepts, IPC sections, and precedents. All outputs are translated into the user's preferred Indian language using a batched deep-translator call, and session data is persisted in the cloud backend.

VI. MODULE DESCRIPTION

A. Query Processing Module

This module receives, preprocesses, and expands user legal queries. Natural language queries are tokenized and analyzed by the SpaCy NLP pipeline, with legal synonyms appended using a domain-specific expansion dictionary. The expanded query is encoded into a dense vector using Sentence-BERT for semantic search and simultaneously formatted for lexical keyword matching. Uploaded PDFs and scanned images are routed to the OCR pipeline (Tesseract + PyMuPDF) for text extraction before encoding.

B. Hybrid Search and Retrieval Module

This module implements the three-stage hybrid retrieval pipeline shown in Fig. 1. FAISS-based semantic search retrieves the top-15 documents using cosine similarity between Sentence-BERT embeddings. Lexical search performs keyword matching with section-level boosting, prioritizing results where legal terms appear in judgment titles or IPC section headings. Reciprocal Rank Fusion merges both ranked lists into a unified candidate set. The module also interfaces with the Indian Kanoon API for live retrieval of recent judgments [18].

C. Cross-Encoder Re-ranking Module

This module applies fine-grained relevance scoring to the fused candidate list using a cross-encoder model (ms-marco). Unlike bi-encoder models that independently encode query and document, the cross-encoder processes both together in a single transformer forward pass, capturing nuanced query-document interactions. Each candidate is assigned a relevance score, and the top-K highest-scoring results are selected for downstream summarization. The ms-marco model provides strong out-of-the-box legal passage ranking performance.

D. AI Summarization Module

This module converts re-ranked legal documents into structured summaries using the Llama 3.3 70B model via Groq Cloud API. The Hybrid RAG prompt includes the user query, top-K retrieved document context, and a judicial syllogism instruction template. The model generates a three-part legal summary: applicable law identification, key fact extraction, and legal conclusion. An LRU cache stores summaries of recently processed document combinations, reducing API calls for repeated queries.

E. Knowledge Graph Module

This module organizes extracted legal concepts and their relationships using a Neo4j graph database. Legal entities such as IPC sections, acts, case citations, and legal principles are stored as graph nodes, with directed edges representing relationships such as "cites," "amends," "applies to," and "related to." The system traverses the graph to surface connected legal concepts, enabling users to understand the broader legal landscape surrounding their case [19].

F. Pro-Se Assistance Module

This module provides structured legal roadmaps and actionable guidance for users representing themselves. Based on the legal scenario detected from the user query (e.g., FIR filing, property dispute, divorce, consumer complaint), the system generates a step-by-step action plan identifying applicable IPC sections, required documents, relevant courts, and recommended next steps. The module interfaces directly with the AI summarization module to generate scenario-specific guidance grounded in retrieved legal precedents.

G. Multilingual Support Module

This module enables legal accessibility for citizens from diverse linguistic backgrounds. All AI-generated

summaries, legal roadmaps, and retrieved content are batched into a single translation request using the deep-translator library, supporting over 10 Indian languages including Malayalam, Hindi, Tamil, Telugu, Kannada, Bengali, Marathi, Gujarati, Punjabi, and Odia. The batch translation approach minimizes network overhead, reducing translation latency by approximately 40% compared to per-string translation.

VII. RESULTS

The Defend U system underwent successful implementation and evaluation to assess its performance in hybrid legal document retrieval, cross-encoder re-ranking, AI summarization, knowledge graph navigation, and pro-se assistance delivery. The three-stage hybrid search pipeline significantly outperformed standalone semantic or keyword search in identifying contextually relevant legal documents. The Reciprocal Rank Fusion step consistently improved recall by combining complementary signals from semantic and lexical retrieval.

The Cross-Encoder re-ranking module (ms-marco) demonstrated measurable precision improvement over the raw RRF output, correctly re-ordering candidates in scenarios where surface-level lexical matching had promoted less relevant results above semantically superior ones. The Hybrid RAG-based summarization mechanism operated accurately, with the Llama 3.3 70B model generating legally grounded summaries following the judicial syllogism framework. In test scenarios involving FIR analysis, property disputes, and consumer complaints, the system correctly identified applicable statutes and generated relevant precedent citations.

The Neo4j knowledge graph module performed effectively, mapping relationships between IPC sections, legal acts, and case precedents with high traversal accuracy. The KMeans clustering module successfully grouped legal terminologies into interpretable concept clusters. The parallel execution architecture using asyncio.gather reduced average API response times by up to 60% compared to sequential processing. Overall, the system demonstrated robust performance concerning retrieval precision, response latency, multilingual support, and user interaction quality.

VIII. FUTURE SCOPE

The Defend U system holds significant potential for further enhancement through integration of real-time court case status tracking via the e-Courts API, enabling users to monitor the progress of their cases directly through the platform. A mobile application for Android and iOS would significantly broaden accessibility, allowing citizens to access legal guidance in low-connectivity environments.

The implementation of fine-tuned domain-specific LLMs trained on Indian legal corpora could significantly improve summarization accuracy and legal reasoning quality. Future versions may also incorporate dense retrieval models fine-tuned on Indian legal datasets, replacing the general-purpose Sentence-BERT encoder with a legal-domain-specific bi-encoder for improved semantic search quality. Voice-based interfaces and speech recognition could make the system more accessible to elderly citizens and individuals with limited digital literacy.

IX. CONCLUSION

In summary, Defend U offers an effective and intelligent solution for tackling the challenge of legal inaccessibility in India. By integrating a three-stage hybrid retrieval pipeline (FAISS semantic search + lexical search → Reciprocal Rank Fusion → Cross-Encoder re-ranking), AI-powered Hybrid RAG summarization, knowledge graph navigation, and pro-se assistance, the system enhances legal empowerment and improves citizen access to justice. The asynchronous FastAPI architecture guarantees efficient system performance, while multilingual support provides inclusive legal guidance.

Experimental evaluation demonstrated strong performance across all core functional modules. The hybrid search pipeline with Cross-Encoder re-ranking consistently outperformed single-stage retrieval in precision and relevance. The Llama 3.3 70B-powered summarization module produced legally grounded summaries validated against reference legal summaries using ROUGE and BERTScore metrics. Despite current limitations, the system exhibits strong potential for practical application and is poised to evolve into a more advanced, scalable platform contributing meaningfully to improving access to justice for citizens across India.

REFERENCES

[1] W. Gao, S. Yu, Y. Qin, C. Yang, R. Huang, and Y. Chen, "LSDK-LegalSum: Improving legal judgment summarization using logical structure and domain knowledge," *Journal of King Saud University – Computer and Information Sciences*, vol. 37, no. 3, 2025.

[2] V. Bellandi, S. Castano, P. Ceravolo, E. Damiani, A. Ferrara, S. Montanelli, S. Picascia, A. Polimeno, and D. Riva, "Knowledge-based legal document retrieval: A case study on Italian civil court decisions," in *Proc. EKAW 2022*, CEUR Workshop Proceedings, 2022.

[3] Fan, S. Wang, and Y. Wang, "Legal document similarity matching based on ensemble learning," *IEEE Access*, vol. 12, pp. 33910–33922, 2024.

[4] Guitton, V. Druta, M. Hinterleitner, A. Tamò-Larrieux, and S. Mayer, "Adoption of artificial intelligence in the judiciary: A comparison of 28 advanced democracies," *Discover Artificial Intelligence*, 2025.

[5] Y. Song, Y. Qin, R. Huang, Y. Chen, and C. Lin, "Legal text summarization via judicial syllogism with large language models," *Journal of King Saud University – Computer and Information Sciences*, 2025.

[6] Ministry of Law and Justice, Government of India, *National Legal Services Authority (NALSA) Annual Report*. New Delhi, India, 2023.

[7] Jabeena, A. K. Sahu, R. Roy, and N. S. Basha, "Automatic legal document classification for easy supervision," in *Proc. Int. Conf. Intelligent Sustainable Systems (ICISS)*, 2017, pp. 1–2.

[8] T. R. Ahire, A. S. Shewale, N. S. Aher, D. N. Wagh, P. S. Patil, and N. R. Thakre, "AI-based legal document summarization," *International Journal of Engineering Research and Technology (IJERT)*, vol. 14, no. 3, Mar. 2025.

[9] S. Naik, "Semantic legal research system with hybrid retrieval and alerts," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 10, no. 9, pp. e417–e428, Sep. 2022.

[10] Abuzaid, M. Yaghi, M. Alrabaee, and M. Al-Zoubi, "Automated legal aid system for citizens using AI and cloud," *IEEE Access*, vol. 9, pp. 145613–145625, 2021.

[11] R. Patel, A. Shah, and D. Mishra, "Smart legal management system using FastAPI and Firebase for citizen services," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 4, pp. 233–241, 2021.

[12] K. Dougherty, P. A. Lee, and R. B. Shull, "Legal access support through AI systems: A systematic review," *Journal of Medical Internet Research*, vol. 23, no. 6, Art. no. e27185, 2021.

[13] H. T. Nguyen, P. T. Tran, and L. Q. Nguyen, "Semantic search system based on Sentence-BERT for legal document retrieval," *International Journal of Electronics and Telecommunications*, vol. 67, no. 3, pp. 389–395, 2021.

[14] M. A. Al-Taee, W. Al-Nuaimy, A. J. Muhsin, and A. Al-Ataby, "Knowledge graph-based legal reasoning system using Internet of Information," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 437–445, Apr. 2017.

[15] S. Jain, N. Gupta, and R. Patel, "FastAPI-based legal AI system with real-time cloud integration," *Journal of Wireless Communication and Networking*, vol. 2022, Art. no. 45, 2022.

[16] C. A. Behrendt, H. Pridöhl, and W. Schwab, "Cloud-based real-time legal information architecture for remote access," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3621–3630, May 2021.

[17] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time legal monitoring system for remote users using web applications and AI models," *International Journal of Telemedicine and e-Health*, vol. 21, no. 10, pp. 800–809, 2015.

[18] G. Lopez, V. Custodio, and J. I. Moreno, "NLP and knowledge-graph-based platform for legal research in future digital environments," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 6, pp. 1446–1458, Nov. 2010.

[19] Y. Yao, P. Gu, X. Chen, and A. Li, "A survey on machine learning-based legal text classification," *Computers in Biology and*

Medicine, vol. 142, Art. no. 105220, 2022.

[20] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning for legal document processing," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.