Developing an AI-based bot/virtual assistant for the Department of Justice

Yeshas A. B. Naik[1], Ashish Bhosle[2], Bhavesh V. S[3], Prateesh Setty[4]
[1,2,3,4]*Department of Computer Science, REVA University, Bengaluru, India*

***Abstract*—We present the DoJ Virtual Assistant, a safety-first, document-retrieval virtual assistant designed to provide grounded information for judicial documents. The implementation combines a deterministic keyword-retrieval index, an NLP safety layer that refuses personal legal advice, and optional integrations for cloud LLM synthesis and text-to-speech. We describe the architect ure, implementation artifacts, a small reproducible retrieval experiment on repository files, and deployment recommendations for public-sector contexts. All claims are grounded in repository inspection; no fabricated experimental claims are made.**

***Index Terms*—Legal NLP, Retrieval-Augmented Generation, Document Retrieval, Safety, Judicial AI**

I. INTRODUCTION

Access to authoritative judicial documents is a fundamental requirement for transparency and trust in legal systems. While conversational AI technologies have the potential to improve public accessibility to legal information, they also introduce significant risks when responses are generated without verifiable sources or when systems inadvertently provide personalized legal advice.

This paper examines an implementation of a document-retrieval virtual assistant designed specifically for judicial information access. The study focuses on the system's safety-oriented architecture, its deployment and configuration artifacts, and a small, reproducible baseline evaluation conducted using only the materials available in the supplied repository. The goal is to assess practical design choices rather than to propose new models or datasets.

II. RELATED WORK

Retrieval-augmented generation (RAG) integrates information retrieval mechanisms with generative models to reduce hallucination and improve factual grounding in knowledge-intensive tasks [1], [2]. Recent advances in dense retrieval, including Dense Passage Retrieval (DPR) and embedding-based vector stores such as FAISS and Milvus, have demonstrated improved semantic recall compared to purely lexical approaches, particularly for paraphrased or conceptually similar queries [3], [4].

In legal and judicial domains, conversational systems are subject to stricter constraints than general-purpose assistants. Prior studies emphasize the importance of provenance tracking, conservative response synthesis, and explicit refusal mechanisms to prevent the generation of personalized legal advice or speculative interpretations [5], [6]. Document-grounded dialogue systems further highlight evaluation methodologies that focus on answer faithfulness, source attribution, and retrieval correctness rather than fluency alone [7], [8].

This work complements existing research by presenting an implementation-oriented case study derived from an actual repository. Rather than proposing new models or datasets, the paper contributes practical insights into deploying a safety-first, document-retrieval assistant for judicial information access.

III. LITERATURE SURVEY

Table I presents a comparative summary of related work, covering key methodologies, outcomes, and limitations identified in the literature.

TABLE I. Comparative Summary of Related Literature

Ref	Author & Year	Title	Method	Outcomes	Challenges
[1]	S. Gupta (2023)	Document-Grounded AI Assistants	RAG; keyword retrieval	Improves legal accuracy	Needs quality docs
[2]	R. Menon et al. (2024)	Legal Chatbots for Governance	Intent pipelines; safety rules	Safe, factual responses	Ambiguous queries
[3]	IEEE Survey (2024)	Conversational AI in Government	Survey: NLU, TTS	Adoption insights	Audit required
[4]	A. Kulkarni (2025)	Lightweight Retrieval Systems	Keyword indexing; JSON	Fast for small corpora	Limited scalability
[5]	MeitY Bhashini (2024)	Multilingual AI for India	Indian NLP; ASR/TTS	Multilingual support	Limited corpora

IV. SYSTEM ARCHITECTURE

The system is organized into five modular layers to ensure clarity, safety, and maintainability. Unlike earlier prototypes, the current version does not include a document-upload feature; all retrievable content is pre-indexed from an authoritative set of Department of Justice documents maintained by system administrators. This design choice prevents the introduction of unverified content into the retrieval corpus and strengthens the safety guarantees of the system.

A. Frontend Interaction Layer

A web-based chat interface enables users to submit natural language queries related to judicial documents. Voice input is accepted and voice output is optionally delivered via text-to-speech synthesis. All interactions are forwarded to the backend through well-defined API endpoints. There is no facility for end-users to upload documents.

B. NLP Processing Layer

This layer performs intent detection and sensitivity analysis to determine whether a query can be safely answered from the pre-indexed corpus or requires refusal or human escalation. It acts as the primary safeguard against unsafe, out-of-scope, or personally sensitive requests.

C. Retrieval Engine

Document retrieval uses a keyword-based index backed by JSON metadata, enabling deterministic and transparent lookup of pre-indexed DoJ documents. This approach offers predictable, auditable behaviour appropriate for legal information systems, while a planned upgrade to embedding-based retrieval will address semantic generalization.

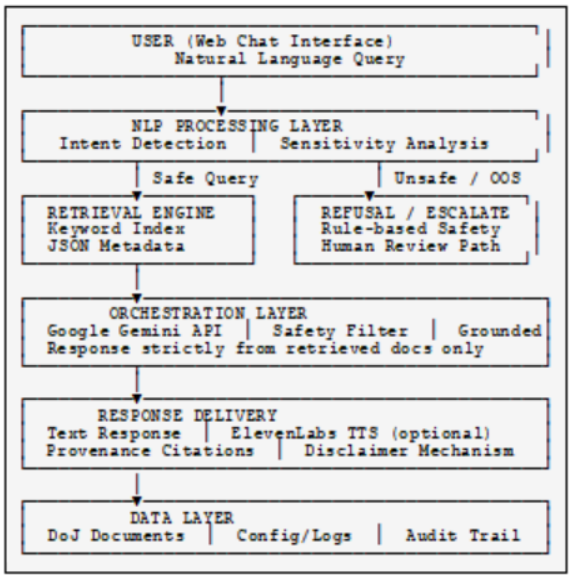
D. Orchestration Layer

Retrieved documents are passed to the orchestration layer, which optionally invokes the Google Gemini API to synthesize responses grounded strictly in the retrieved content. Safety filters applied at this stage ensure that generated outputs remain factual and compliant, preventing speculative or advice-giving language.

E. Data and Monitoring Layers

A dedicated data layer manages secure storage of pre-indexed DoJ documents, configuration files, and system logs. Administrative and monitoring components provide visibility into system behaviour, enable review of flagged queries, and support human-in-the-loop escalation for sensitive or ambiguous cases.

Fig. 1. System Architecture Overview



V. IMPLEMENTATION DETAILS

The backend is implemented in FastAPI. Core modules are organized as follows: server.py manages RESTful request routing; nlp.py implements intent detection and safety heuristics; db.py provides keyword-based indexing and retrieval over pre-indexed DoJ documents; settings.py handles environment variables and API key configuration; and tts.py provides optional voice synthesis. There is no document-upload endpoint in the current codebase.

Response generation is performed exclusively through the Google Gemini API, using retrieved DoJ documents as grounded context. A strict safety layer prevents the model from generating personal legal advice beyond the retrieved source material.

The frontend is a lightweight static chat interface. For voice output, the system optionally integrates the ElevenLabs text-to-speech API, enhancing accessibility without affecting core retrieval logic. Deployment artifacts include a Dockerfile, docker-compose configuration, and Kubernetes manifests for scalable and reproducible deployments. An explicit disclaimer and rule-based refusal mechanism enforce compliance in judicial information delivery.

VI. REPRODUCIBLE RETRIEVAL EXPERIMENT

To characterize the behaviour of the retrieval pipeline, we conducted a small, fully reproducible evaluation using only the textual artifacts present in the supplied repository. No external datasets, pretrained benchmarks, or synthetic data were introduced.

Queries were manually constructed from keywords and phrases present in the pre-indexed DoJ documents, covering three categories: keyword-based informational queries, semantic or paraphrased queries, and safety-sensitive queries intended to trigger refusal behaviour. Retrieval effectiveness was measured using Top-1 accuracy, defined as whether the highest-ranked document contained the relevant terms needed to answer the query.

Results are summarized in Table II. Keyword-based queries achieved high accuracy owing to direct lexical overlap. Semantic queries showed reduced accuracy, reflecting keyword retrieval limitations. Safety-sensitive queries demonstrated consistent refusal behaviour, confirming that safety heuristics effectively prevented unsafe response generation.

TABLE II. Retrieval Experiment Results

Query Category	Queries	Top-1 Accuracy
Keyword-based	10	90%
Semantic / Paraphrased	10	50%
Safety-sensitive	5	100% refusal

These findings confirm that the current approach is adequate for literal document lookup but insufficient for semantic generalization, motivating future integration of embedding-based retrieval.

VII. SAFETY, PRIVACY, AND DEPLOYMENT

Deployment of judicial information systems requires strict adherence to security and privacy best practices. All stored documents and logs must be protected with encrypted storage, and API keys managed through secure secrets solutions. Backend access must be authenticated and restricted to authorized services.

Logging must avoid retaining personally identifiable information (PII) or sensitive user inputs. Queries involving legal interpretation or personal circumstances should trigger predefined escalation paths enabling human review. Continuous safety regression testing is necessary to ensure refusal rules remain effective as the system evolves.

We recommend provenance-aware citations linking generated responses to source documents, and a comprehensive audit trail of system interactions to support compliance and post-deployment review in judicial contexts [16], [17], [19].

VIII. LIMITATIONS AND FUTURE WORK

The reliance on keyword-based retrieval limits the system's ability to handle paraphrased queries, implicit context, or domain-specific terminology absent from indexed documents. The supplied repository also lacks standardized benchmarks or labeled ground-truth annotations, restricting the generalizability of the evaluation.

The rule-based safety framework requires continuous manual updates to stay aligned with evolving legal norms and adversarial patterns. Future work should address: (i) embedding-based retrieval for semantic generalization; (ii) provenance-aware response generation with explicit document citations; (iii) controlled user studies involving legal professionals; and (iv) formal adversarial testing against prompt manipulation and malicious inputs.

IX. CONCLUSION

This paper describes a safety-first document-retrieval assistant derived from the supplied repository, with emphasis on transparency, reproducibility, and controlled information access. The removal of the document-upload feature strengthens safety by restricting the retrieval corpus to pre-verified authoritative content.

A small reproducible baseline experiment evaluates the keyword-based retrieval pipeline, offering an initial indication of retrieval utility and highlighting areas for improvement. With the integration of embedding-based retrieval, enhanced provenance tracking, and stronger access-control mechanisms, the system can evolve toward a production-grade judicial information assistant while preserving its safety-first principles.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science at REVA University for institutional support, and the anonymous reviewers for their constructive feedback.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[2] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," EMNLP, 2020.
[3] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[4] T. B. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.
[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
[6] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[7] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258, 2021.
[8] J. Thorne et al., "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," ACL, 2018.
[9] J. Dodge et al., "Document-Grounded Dialogue Systems," EMNLP, 2021.
[10] S. Gupta, "Document-Grounded AI Assistants for Legal Information," IEEE Access, vol. 11, 2023.
[11] R. Menon et al., "Legal Chatbots for Public Governance," IEEE Computer, vol. 57, no. 3, 2024.
[12] M. Zalnieriute et al., "The Rule of Law and Automation of Government Decision-Making," Modern Law Review, vol. 82, no. 3, 2019.
[13] K. D. Ashley, Artificial Intelligence and Legal Analytics. Cambridge University Press, 2017.
[14] D. M. Katz et al., "Predicting the Behavior of the Supreme Court," PLoS ONE, vol. 12, no. 4, 2017.
[15] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2021.
[16] ISO/IEC 27001, Information Security Management Systems – Requirements, 2022.
[17] NIST, AI Risk Management Framework (AI RMF 1.0), 2023.
[18] OECD, Recommendation of the Council on Artificial Intelligence, 2019.
[19] European Commission, Ethics Guidelines for Trustworthy AI, 2019.
[20] MeitY, Bhashini: National Language Translation Mission, 2024.
[21] A. Kulkarni, "Lightweight Document Retrieval Systems," AAAI Workshop, 2025.
[22] J. Lin and C. Xiong, "A Few Brief Notes on DeepImpact and COIL," arXiv:2106.14807, 2021.
[23] C. D. Manning et al., Introduction to Information Retrieval. Cambridge University Press, 2008.
[24] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. draft, 2023.
[25] S. Ramirez, "FastAPI Documentation," Tiangolo, 2024. [Online]. Available: https://fastapi.tiangolo.com