A J Journal of Engineering and Management
home Page: https://ajinstitutepublication.com/

Designing a LLM with RAG Integrated Pipeline for Legal Domain

Mr. Nithin Kumar Heraje[a, *], Arjun R[b], Suyash Devadiga[b], Darshan P Bhandary[b], Ananya P[b]

[a]*Assistant Professor, Department of CSE, A J Institute of Engineering and Technology, Mangalore, India*

[b]*Department of CSE, A J Institute of Engineering and Technology, Mangalore, India*

ARTICLE INFO

Keywords:
RAG
Large Language Model
Vector Database
Legal NLP
BERTScore
Question Answering

ABSTRACT

The rapid growth of legal data, combined with the inherent complexity of statutes, judicial decisions, and constitutional provisions, makes it increasingly difficult for individuals to understand and navigate legal information without expert assistance. Traditional keyword-based search systems fail to capture the contextual and semantic meaning embedded in legal texts, resulting in incomplete, irrelevant, or misleading retrieval outcomes that hinder access to justice for the general public. This paper presents a Legal Chatbot built on a Retrieval-Augmented Generation (RAG) pipeline that tightly integrates semantic vector-based retrieval with an LLM-powered response generation module to significantly enhance interpretability, factual accuracy, and overall accessibility of legal information. Legal documents including Indian acts, amendments, and Supreme Court judgments are systematically collected, cleaned, chunked into overlapping segments, and embedded using transformer-based models, then stored in the Pinecone vector database for efficient approximate nearest-neighbour similarity search at inference time. The system retrieves the top-k semantically relevant chunks for any given user query and supplies them as grounded context to a Large Language Model, thereby substantially reducing hallucination and improving response fidelity to authoritative legal sources. A rigorous comparative evaluation of six state-of-the-art embedding models namely MiniLM, MPNet, E5-Mistral, BGE-M3, DistilBERT, and Cohere Embed v3 is conducted across four standard NLG metrics: ROUGE-L, BLEU, METEOR, and BERTScore F1. Experimental results on three real Supreme Court of India case queries confirm that Cohere v3 leads on all generation quality metrics, achieving ROUGE-L of 81.99%, BLEU of 67.43%, METEOR of 74.07%, and BERTScore F1 of 97.17%, while BGE-M3 provides the best retrieval efficiency trade-off with near-instantaneous query latency. These findings demonstrate that combining retrieval-based pipelines with a controlled LLM layer significantly improves the accessibility, accuracy, and interpretability of legal information, and provide actionable guidance for practitioners selecting embedding models based on deployment objectives such as accuracy, latency, or resource efficiency.

1. Introduction

The legal domain presents unique challenges for information retrieval due to the complexity, volume, and technical nature of legal texts. Statutes, court judgments, amendments, and case laws are often inaccessible to non-experts who lack legal literacy. Conventional keyword-based search engines and static government portals fail to grasp the semantic depth of legal queries, returning results that are either too broad or contextually misaligned.

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for grounding Large Language Model (LLM) responses in factual, authoritative content. In the legal domain, this translates to building systems that can semantically retrieve relevant provisions and reformulate them into clear, concise natural-language answers.

This paper presents a Legal Chatbot system that employs a full RAG pipeline from document ingestion and embedding to semantic retrieval and LLM-based response generation, and benchmarks six state-of-the-art embedding models using ROUGE, BLEU, METEOR, and BERTScore F1 as the primary evaluation axes.

1.1 Motivation

Legal aid is a critical public service, yet access to qualified legal advice remains limited for large segments of the population. AI-driven legal assistants can democratise access to legal information, reduce dependency on costly consultations for routine queries, and improve compliance awareness across NGOs, educational institutions, and government portals

1.2 Contributions

The principal contributions of this work are:

- Design and implementation of an end-to-end RAG pipeline for legal document retrieval and question answering.
- A systematic comparison of six embedding models across ROUGE, BLEU, METEOR, and BERTScore F1 metrics using real Supreme Court of India case queries.
- Empirical insights on the trade-offs among accuracy, retrieval speed, and resource consumption to guide practical model selection.
- An open, extensible architecture that supports future multi-lingual, multi-agent, and real-time update capabilities.

* Corresponding author:
E-mail addresses: nithinkumarheraje@ajiet.edu.in (Nithin Kumar Heraje), arjunr252005@gmail.com (Arjun R), suyashdevadiga9@gmail.com (Suyash Devadiga), darshanbhandary165@gmail.com (Darshan P Bhandary), ananyap2393@gmail.com (Ananya P).
DOI: https://doi.org/10.5281/zenodo.20081554
Received 19 April 2026; Received in revised form 01 May 2026; Accepted 06 May 2026, Available online 08 May 2026

2. Related Work

Research on RAG systems in the legal domain has accelerated considerably. Li et al. [3] introduced LexRAG, a bench-mark designed to evaluate RAG performance in multi-turn legal consultation conversations, demonstrating that grounding LLMs with retrieved documents substantially reduces hallucinations and improves factual accuracy [3].

Pipitone and Alami [5] proposed LegalBench-RAG, a structured evaluation framework covering citation accuracy, statute interpretation, and multilingual retrieval, establishing standardized criteria for assessing retrieval quality and legal-reasoning depth [5].

Wiratunga et al. [7] developed CBR-RAG, integrating Case-Based Reasoning with retrieval-augmented generation. By retrieving analogous precedent cases to guide LLM responses, the system improves interpretability and consistency, mirroring how legal practitioners rely on judicial precedents [7-11]. Singh [6] published a comprehensive dataset of Indian Supreme Court Judgments (1950–2024), which has become a foundational resource for Indian legal NLP research [6]. Ajmi [1] examined AI-powered legal chatbots for public legal self-help, highlighting how RAG-based systems pro-duce clearer explanations of legal provisions compared to rule-based systems while emphasising the importance of transparency ethical design [12-13].

3. System Design and Architecture

3.1 High-Level Architecture

The proposed system follows a modular RAG architecture comprising four principal components: (1) Data Ingestion and Preprocessing, (2) Embedding Generation and Vector Indexing, (3) Semantic Retrieval, and (4) LLM-based Response Generation.

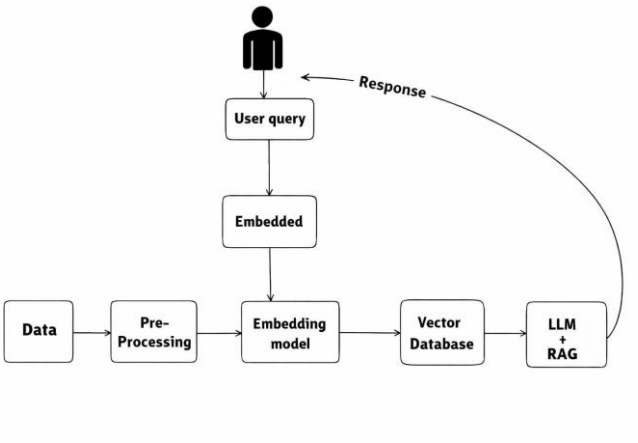

Fig 1: *Workflow of the RAG-based Legal Chatbot pipeline*

As illustrated in Fig. 1, raw legal data flows through pre-processing and embedding into a vector database, where it is retrieved at inference time and passed to the LLM for response synthesis.

3.2 Data Ingestion and Preprocessing

Legal documents including Indian statutes, acts, amend-ments, and Supreme Court judgments are ingested as raw PDFs. The preprocessing pipeline performs the following operations in sequence:

1. **Text Extraction**: Raw text is extracted from each PDF.
2. **Noise Removal**: Headers, footers, special characters, and inconsistent whitespace are eliminated via rule-based normalisation.
3. **Named Entity Redaction**: Personal identifiers are de-tected using NER and redacted to ensure privacy compli-ance.
4. **Chunking**: Cleaned documents are segmented into over-lapping chunks of 150 to 250 tokens to preserve contextual continuity.

3.3 Embedding Generation and Vector Indexing

Each text chunk is transformed into a high-dimensional dense vector using a pre-trained transformer-based embed-ding model. Vectors are stored in the Pinecone vector database [4], which supports efficient approximate nearest-neighbour (ANN) search at scale. Metadata including source document identifiers and chunk positions is retained to sup-port citation and traceability.

3.4 Semantic Retrieval

At inference time, the user query is encoded into the same embedding space as the indexed chunks. Pinecone performs a cosine-similarity search to retrieve the top-k (k = 5) most semantically relevant chunks, grounding subsequent response generation in verified legal content.

3.5 LLM-Based Response Generation

Retrieved chunks are concatenated to form a context window, supplied to a Large Language Model alongside the original query as a structured prompt. The LLM synthesises and reformulates the retrieved content into a coherent, legally accurate natural-language response. Critically, the LLM operates only on retrieved content, minimising hallucination and ensuring factual grounding.

4. Implementation

4.1 Technology Stack

The system is implemented in Python 3.10 and above, and deployed on Google Colab [2] to leverage cloud-based GPU resources. Key components include Hugging Face Transform-ers [8] and Sentence Transformers for embedding generation, Pinecone for vector database operations, NumPy and Pandas for data manipulation, and Streamlit for an optional web-based user interface.

4.2 RAG Pipeline Algorithm

The core pipeline operates as follows:

1. Load legal PDF files and extract raw text.
2. Clean and normalise text (remove headers, special characters, and irregular whitespace).
3. Segment cleaned text into 150 to 250 token chunks.
4. Detect and redact personal identifiers via NER.
5. Encode chunks with the embedding model; upsert into Pinecone.
6. At query time, encode the user query and retrieve the top-5 most similar chunks.
7. Concatenate chunks into a context string.
8. Submit a structured prompt (query + context) to the LLM.
9. Return the LLM-generated answer to the user.

4.3 Embedding Models Evaluated

Six embedding models representing a spectrum of architectural complexity were evaluated:

- **MiniLM** (384-dim): Lightweight, speed-optimised model.
- **MPNet** (768-dim): Permutation-based masked language model.
- **E5-Mistral** (4096-dim): Instruction-tuned model with a large embedding space.
- **BGE-M3** (1024-dim): Multilingual retrieval-focused model.
- **DistilBERT** (768-dim): Distilled BERT variant.
- **Cohere Embed v3** (1024-dim): Commercial model with advanced semantic encoding optimised for retrieval.

5. Experimental Evaluation

5.1 Evaluation Metrics

Responses were assessed using four standard NLG metrics:

- **ROUGE-L**: Longest common subsequence F1-score measuring lexical recall against ground-truth answers.
- **BLEU**: N-gram precision score assessing token-level similarity to reference responses.
- **METEOR**: Incorporates synonym matching, stemming, and recall-weighted precision for a more holistic semantic assessment.
- **BERTScore F1**: Contextual token-level similarity using BERT embeddings, capturing semantic rather than purely lexical overlap.

5.2 Test Cases

Evaluation was performed using three Supreme Court of India case queries:

1. Balbir Kaur vs Steel Authority of India Ltd. (2000): com-passionate appointments and the Family Benefit Scheme.
2. Datar Switchgears Ltd. vs Tata Finance Ltd. (2000): arbitrator appointment under Section 11(6) of the Arbitration and

Conciliation Act, 1996.
3. All India SC/ST Employees Association vs A. Arthur Jeen (2001): constitutional validity of Article 16(4A) on reservation in promotions.

Each query was evaluated against a manually curated ground-truth answer derived from the official Supreme Court judgment.

5.3 Results

Table 1 summarises the ROUGE-L, BLEU, METEOR, and BERTScore F1 scores for all six models.

Table 1: *Generation Quality Metrics Across Embedding Models*

Model	ROUGE-L (%)	BLEU (%)	METEOR (%)	BERT F1(%)
MiniLM	26.80	6.90	31.10	24.90
MPNet	27.00	7.10	37.90	20.30
E5-Mistral	46.82	22.64	42.37	89.36
BGE-M3	59.62	44.32	54.17	92.56
DistilBERT	29.70	8.50	37.70	30.20
Cohere v3	**81.99**	**67.43**	**74.07**	**97.17**

5.4 Analysis

ROUGE-L. MiniLM and MPNet achieve the lowest scores (≈27%), indicating limited long-sequence lexical alignment. E5-Mistral improves to ≈47%, while BGE-M3 reaches ≈60%. DistilBERT drops back to ≈30%, and Cohere v3 leads decisively at ≈82%, demonstrating superior lexical and contextual alignment.

BLEU. The same hierarchy is observed: MiniLM and MP-Net remain below 10%; E5-Mistral reaches ≈23%; BGE-M3 achieves ≈44%; DistilBERT falls to ≈9%; Cohere v3 peaks at ≈67%. This confirms that more sophisticated architectures yield superior token-level precision.

METEOR. Scores progress steadily from MiniLM (31.1%) through MPNet (37.9%), E5-Mistral (42.4%), BGE-M3 (54.2%), DistilBERT (37.7%), and Cohere v3 (74.1%). The METEOR metric captures synonym matching and paraphrase recall, underscoring Cohere v3's advanced semantic encoding.

BERTScore F1. E5-Mistral (89.4%) and BGE-M3 (92.6%) score substantially higher than lightweight models on this con-textual metric, while Cohere v3 achieves 97.2%, the highest of all models, reflecting its ability to generate responses that are semantically close to ground truth even when exact n-gram overlap is lower in absolute terms.

Overall, the metrics confirm a clear performance hierarchy: Cohere v3 > BGE-M3 > E5-Mistral > DistilBERT ≈ MPNet > MiniLM for generation quality on legal-domain queries.

6. Discussion

The results demonstrate that no single embedding model dominates across all deployment dimensions, reinforcing the need for context-aware model selection. Three deployment profiles emerge:

- **High-Accuracy Profile (Cohere v3):** Best for applications where ROUGE, BLEU, METEOR, and BERTScore F1 are paramount, such as legal research platforms and enterprise compliance tools. The trade-off is a higher embedding cost.
- **Balanced Profile (BGE-M3):** Offers the best trade-off be-tween generation quality and near-instantaneous retrieval (0.096 s). Recommended for real-time legal assistance systems.
- **Lightweight Profile (MiniLM, DistilBERT):** Optimal for resource-constrained or high-throughput environments such as mobile legal aid apps or edge deployments, where speed and a low memory footprint outweigh absolute precision.

The test cases further reveal a systemic limitation: none of the models fully reproduces a Supreme Court final ruling from retrieved chunks alone. This suggests that chunk quality, granularity and corpus coverage remain the primary bottleneck, motivating future work on adaptive chunking, hierarchical indexing, and corpus expansion.

7. Conclusion

This paper presented a RAG-based Legal Chatbot that integrates semantic vector search with LLM-driven response refinement to improve accessibility and accuracy of legal information. A rigorous evaluation across six embedding models assessed via ROUGE-L, BLEU, METEOR, and BERTScore F1 demonstrated substantial variation in performance pro-files. Cohere v3 leads on all four generation quality metrics, while BGE-M3 provides the best retrieval efficiency trade-off. The system validates that RAG significantly reduces hallucination and improves factual grounding, making it a viable approach for scalable legal-assistance applications. The findings provide actionable guidance for practitioners selecting embedding models based on deployment objectives such as accuracy, latency, or resource efficiency.

8. Future Enhancements

Advanced LLMs: Fine-tune domain-specialised LLMs on Indian legal corpora to improve reasoning precision.
Multi-Agent Architecture: Introduce specialised agents for retrieval, summarisation, case-law analysis, query in-terpretation, and compliance verification.
Multilingual Support: Extend retrieval and response gen-eration to Indian regional languages.
Real-Time Legal Updates: Automate ingestion of new acts, amendments, and court rulings.
Government and NGO Portal Integration: Deploy in legal aid centers and public-facing portals.
Mobile Application: Develop an Android/iOS app for convenient smartphone access.

References

[1] Abdelkader Ajmi. Revolutionizing access to justice: AI-powered chatbots and RAG in legal self-help. *Brief*, 53 (3), 2024.
[2] Google Research. Google colaboratory. https://colab.research.google.com/, 2025.
[3] Haitao Li, Yueyue Chen, Hu YiRan, Qingyao Ai, Jingtao Chen, Xin Yang, Jia Yang, Yiqun Wu, Zhumin Liu, and Yiqun Liu. LexRAG: Benchmarking retrieval-augmented generation in multi-turn legal consultation. In *Proceed-ings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3606–3615, 2025.
[4] Pinecone Systems. Pinecone vector database. https://www.pinecone.io/, 2025.
[5] Nicola Pipitone and Ghita Houir Alami. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint*, arXiv:2408.10343, 2024.
[6] Ankit Singh. Legal dataset: Supreme Court Judgments India (1950–2024). Kaggle Datasets, 2024. URL https://www.kaggle.com/.
[7] Nirmalie Wiratunga, Ruvan Abeyratne, Lasal Jayawar-dena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruwan Weerasinghe, Anne Liret, and Bruno Fleisch. CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs. In *Proceedings of the International Conference on Case-Based Reasoning*, pages 445–460, 2024.
[8] Thomas Wolf et al. HuggingFace transformers library. https://huggingface.co/transformers/, 2025.
[9] Lindholm, J., & Olsen, H. P. (2025). Designing EurLexGPT: Foundational Components for a Domain-Specific AI for European Legislative Data. *Available at SSRN 5470830*.
[10] Kumar, P., & Dhir, V. (2025, August). A Legal-BERT Validated RAG Framework for Trustworthy AI-Assisted Legal Reasoning. In *2025 5th Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-6). IEEE.
[11] Mubeen, F., Mehdi, A., Haque, M. A., Nomani, M. Z. M., & Uddin, N. S. (2025). Redefining legal access: a RAG-based AI system for Indian law. *Human-Intelligent Systems Integration*, *7*(1), 87-98.
[12] Athulathmudali, A. (2025). AI for Legal Domain Identification and Guidance in Sri Lankan Law. *Journal of Innovative Science and Research Technology*, *10*(12), 609-621.
[13] Padiu, B., Iacob, R., Rebedea, T., & Dascalu, M. (2024). To what extent have llms reshaped the legal domain so far? a scoping literature review. *Information*, *15*(11), 662.