National College of Ireland

Embedding Model Evaluation in Legal RAG Systems for Automobile Law Question Answering

MSc Research Project
MSc in Data Analytics

Nithin Reddy Yelala
Student ID: 23416858

School of Computing
National College of Ireland

Supervisor: Barry Haycock

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Nithin Reddy Yelala
Student ID:	23416858
Programme:	MSc in Data Analytics **Year:** 2025
Module:	Research Practicum
Supervisor:	Barry Haycock
Submission Due Date:	11-12-2025
Project Title:	Embedding model Evaluation in Legal RAG systems for Automobile Law Question Answering
Word Count:	8475 **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nithin Reddy Yelala
Date:	11-12-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Embedding Model Evaluation in Legal RAG Systems for Automobile Law Question Answering

Nithin Reddy Yelala
Student ID: 23416858

Abstract

The exponential growth of complex legal documents in India has made it difficult for non-experts to access and interpret statutory information which is mainly in domains such as automobile law. This study addresses that challenge by developing a Retrieval-Augmented Generation (RAG) framework for legal question answering by focusing on the Motor Vehicles Act, 1988. The study aims to identify the most powerful embedding model for accurate semantic retrieval and contextual response generation. There are eight embedding models which were evaluated using metrics such as Recall, Precision, nDCG, F1, ROUGE-L, BLEU, and Cosine Similarity. All experiments in this study were conducted exclusively on English legal text. The dataset was preprocessed, chunked, and vectorized, with embeddings stored in ChromaDB and used via LangChain for retrieval and LLaMA-based LLMs for answer generation. This study results showed that multilingual-e5-large achieved the highest semantic similarity (0.9558) and balanced accuracy across all metrics, outperforming both general-purpose and domain-specific models. The findings theoretically advance the understanding of embedding alignment in legal semantic retrieval and, in practice, offer a scalable framework for automated legal assistance that enhances accessibility and transparency in the justice system.

Keywords: Retrieval-Augmented Generation, Embedding Models, Legal Question Answering, Semantic Retrieval

1 Introduction

1.1 Background

The multiplied volumes of legal records and regulatory systems have posed a great challenge of accessing and understanding of law especially by non-experts who are required to find particular responses to law questions (Williamson and Prybutok, 2024). The legality of the legal language, heavy terms, cross-referencing, and hierarchical nature of the clauses ensure that it is extremely hard to grasp the meaning of the words and extract the information in an efficient manner and that too by ordinary citizens. Search systems that rely on keywords do not get the semantic shades and contextual association in legal texts and in most cases, they are found not to be complete or relevant to the search query (Magesh et al., 2025). The latest developments in Natural Language Processing and Retrieval-Augmented Generation (RAG) have provided new opportunities to intelligent legal question answering systems. RAG is a semantic retrieval and generative language model that generates accurate and contextually-sensitive answers (Pingua et al., 2025). Nonetheless, the quality of the models in embedding

the legal texts into the semantic vector spaces is critically dependent on the incorporation of model quality.

1.2 Importance of the study

The creation of efficient systems of answering the law questions is incredibly important in terms of the democratization of access to legal information and reduction of the gap between complicated legal system and the knowledge people possess. In India, with a low level of legal literacy and legal advice being both costly and geographically inaccessible, automated legal QA systems can enable citizens to know their rights, responsibilities and legal redress without necessarily having access to high levels of legal knowledge. These systems can help vehicle owners, drivers, insurance companies, and police officers to be in a position to access appropriate information that is accurate and fast with respect to licensing specifications, traffic offenses, penalties systems and accident investigations. The quality of retrieval and accuracy of answers in RAG systems are basically dependent on the model of embedding that is used.

1.3 Research Question

Legal question-answering systems in India face various challenges in accurately retrieving and interpreting complex statutory information from extensive legal documents such as the Motor Vehicles Act, 1988. Traditional keyword-based search methods fail to capture semantic nuances, contextual relationships, and legal reasoning embedded within dense legal terminology, hierarchical clause structures, and extensive cross-referencing. These limitations result in incomplete or irrelevant responses that do not address users' actual legal queries. Retrieval-Augmented Generation (RAG) systems offer a promising solution by combining semantic retrieval with generative language models, but their effectiveness critically depends on the embedding model's ability to represent legal texts in meaningful vector spaces.

Research Question: "Which embedding model architecture demonstrates superior performance in semantic understanding and answer accuracy when integrated within a Retrieval-Augmented Generation (RAG) framework for Indian automobile law question answering?"

1.4 Research Objectives

The research focuses on the following specific objectives:

1. To develop a domain-specific RAG pipeline for legal question answering by integrating embedding-based semantic retrieval with a generative language model for contextual answer synthesis.
2. To preprocess, chunk, and vectorize the Motor Vehicles Act, 1988 into dense embeddings and store them in a vector database for efficient similarity search and legal context retrieval.
3. To evaluate and compare multiple embedding models (including All-MiniLM-L6-v2, bge-large-zh-v1.5, UAE-Large-V1, Qwen-0.6B, InLegalBERT, KoSimCSE-roberta-multitask, all-mpnet-base-v2, and multilingual-e5-large) in terms of metrics.

1.5 Limitations of the study

Although this study gives a detailed analysis of embedding models to legal RAG systems, there are a number of limitations that are to be considered. The study is confined to the Motor Vehicles Act, 1988 and the results may not be completely applicable to other legal areas of law like criminal law or contract law that have different linguistic peculiarities. The test is also taken on the basis of English legal literature only and there is no attempt to measure the performance in regional languages. The query dataset will include 50 questions that will be carefully curated; a bigger dataset will also bolster statistical significance of the differences in observed performance. The Kaggle environment had computational resource constraints that did not support large-scale hyperparameter exploration and model exploration. The study provides a point-in-time view of legislation, and does not consider the problem of keeping legal information in production processes up-to-date.

1.6 Structure of the Report

This report is structured into seven chapters. Chapter 1 introduces the research background, importance, research question, objectives, limitations, and the overall report layout. Chapter 2 reviews existing literature on legal question answering, semantic retrieval, embedding models, and Retrieval-Augmented Generation (RAG) systems. Chapter 3 presents the adopted methodology, including dataset preparation, preprocessing and model selection. Chapter 4 outlines the design specifications, system architecture and technologies used in developing the RAG pipeline. Chapter 5 details the system implementation steps, covering chunking, vector storage, retrieval, and augmented answer generation. Chapter 6 evaluates the embedding models through retrieval metrics and end-to-end QA performance, followed by result analysis and discussion. Finally, Chapter 7 concludes the research, summarizes key findings, and proposes improvements and future research directions.

2 Related Work

2.1 Question Answering in the Legal Domain

Question Answering (QA) in legal field is concerned with accessing and interpreting statutory and case laws and legal precedents automatically and then giving the answers to questions with precision and in context (Abdallah et al., 2023). Legal QA has its own set of challenges as compared to open-domain QA because of the complex constructions of sentences, dense vocabulary, cross-referencing in sentences, and interpretation based on jurisdiction (Barron et al., 2025). Legal sources usually contain very long and conditional clauses and multiple levels of meaning where one word can mean different things in different contexts and therefore the standard method of finding information by keywords is insufficient (Katz et al., 2023). Consequently, initial legal QA systems used rule-based search and Boolean matching which often returned incomplete output or irrelevant results. As Natural Language Processing has developed, semantic embeddings and deep learning have greatly enhanced the possibility of understanding the contextual meaning in legal text (Quevedo et al., 2024). The current systems are now focused on going beyond the surface matching and to comprehend the purpose of a legal query and match it to the correct statute or judgment and produce grounded and citation-based responses. Nevertheless, there are still issues, such as legal language ambiguity, the necessity of a high level of accuracy, and the threat of

hallucination with generative models. This has given rise to growing interest in Retrieval-Augmented Generation (RAG) and domain-specific embeddings, where it is necessary to make sure that legal QA systems can find the most relevant passages and preserve the facts. It is in general seen that to achieve effective legal QA, one needs accurate retrieval, as well as legally reliable answer generation.

2.2 Retrieval-Augmented Generation (RAG) Systems

The reviewed four studies demonstrate that Retrieval-Augmented Generation (RAG) is rapidly evolving and is used in the various fields of enhancing the performance of large language models (LLM), its dependability, and basing on facts. A systematic review of the development of RAG since its initial introduction to the open-domain question answering to enterprise-level applications and its role in reducing the use of hallucinations and obsolete information in LLMs, was carried out (Oche et al., 2025). Based on this premise, (Gupta et al., 2024) constructed a general survey that addressed the architectural design of the memories between retrieval and generation aspect, which aimed to enhance retrieval performance and RAG flexibility in responding to questions, summarizing and reasoning knowledge. The two sources stressed the advantage of RAG in the factual accuracy and relative contextual relevance but also stated that such concerns as scalability, bias, and ethical concerns remain with the procedure. The disparity in the assessment of RAG systems as it is observed in (Lyu et al., 2025) was addressed by including a CRUD-based benchmarking framework according to which RAG applications are divided into Create, Read, Update, and Delete scenarios and in which a range of datasets is created to assess the performance of each scenario. This international evaluation showed differences in the efficiency of RAG in activities suggesting that there should be adaptive processes of retrieval and production. (Rani et al., 2024) disclosed domain-specific performance of RAG through an example of knowledge graph-based RAG tools on the healthcare domain specifically diabetes-oriented solutions involving the application of the keyword, vector retrieval, and graph retrieval to gain improved explainability and authenticity. The framework has a score of 82.19 per ROUGE-1 which means that factual consistency and traceability have been significantly improved.

2.3 Review of Existing RAG-Based Legal Question Answering Studies

Recent developments in the Retrieval-Augmented Generation (RAG) have brought important benefits to the legal question answering systems. The case law of Sri Lanka was presented as domain-specific embeddings with the AnglE-BERT and the BM25 models to improve the retrieval accuracy of case law in legal texts (Jayawardena et al., 2024). The contextual alignment of the study was to enhance better case-based reasoning but limited in multilingual scalability and cross domain adaptability. In Adaptive RAG Pipeline (Keisha et al., 2025), the suggested flexible RAG model was based on dynamic SBERT and GTE embeddings, a query translator that is context-aware, and optimization of the retrieval depth and response form. Although it scored high Recall at K and Precision at K, it was not as automated as it depended on fixed corpora and manually configured query. (Lima, 2025) demonstrated a dual-embedding method that isolates both label and content-based vectors in enhancing retrieval of referential legal provisions contained in the Brazilian Constitution. It increased label centric query processing, however, it was not generalized to semantic, multi lingual and scale queries. Taken together, these works prove a step towards more interpretable and accurate legal RAG systems but underscore the importance of cross-lingual, semantically

balanced, and completely automated systems, which is well satisfied in the current study by the testing of a variety of embedding models and hybrid performance measures.

2.4 Embedding Models for Semantic Retrieval

Recent progresses in the semantic retrieval and language modeling has inspired different methods of enhancing retrieval accuracy, contextual comprehension, and robustness across domains. First study given by (Monir et al., 2024) who introduced VectorSearch, a high-performance retrieval model meant to overcome the drawback of the traditional retrieval system that has problem accessing semantic subtleties. The VectorSearch uses multi-vector search operations, semantic embeddings and efficient indexing to improve the contextual understanding and precision of retrieval albeit with the disadvantages of computational cost and the dependency on a model. Generalizing this thought into the educational field, (Sajja et al., 2025) suggested two open-source embedding models optimized to the educational question answering utilizing a synthetic dataset of 3,197 filtered and LLM-generated sentence pairs. These models, which were trained with MultipleNegativesRankingLoss and dual-loss, performed better than strong open-source baselines and made them similar to state-of-the-art performance, but due to the small size of the dataset, generalization is limited.

In the meantime, (Kong et al., 2023) introduced SparseEmbed, an embodiment of a hybrid retrieval model that incorporates the advantages of the sparse lexical and dense contextual representation to trade interpretability and semantic expressiveness. SparseEmbed has been shown to achieve competitive efficiency and accuracy compared to other models, such as SPLADE and ColBERT, through end-to-end training, at the increased computational complexity cost. Coupled with such studies, (Excoffier et al., 2024) investigated the quality of Large Language Model (LLM)-based embeddings in the medical field through rephrased ICD-10-CM code descriptions. The findings showed that the general-purpose models outperform domain-specific models, and sensitivity is a problem of specialization of the embedding, because of limited and non-diverse training data. All these papers collectively highlight the current development of the semantic retrieval systems as a general-purpose retrieval to domain-specific optimization by highlighting the novelty in embedding design, hybrid modeling, and domain adaptation. Nevertheless, they also highlight not only persistent limitations in the form of computational efficiency, diversity of the inference data, and generalizability of the models, but also provide a list of challenges to be considered in the future, including the development of more robust, interpretable, and domain-congruent retrieval systems in educational, clinical, and large scale text retrieval.

Table 1: Comparison of Recent Semantic Retrieval and Embedding Studies

Study	Key Contribution	Methodology	Strengths	Limitations
(Monir et al., 2024)	Introduced VectorSearch, a high-performance retrieval model to overcome traditional retrieval systems' inability to capture semantic	Multi-vector search operations, semantic embeddings, and efficient indexing techniques	Improved contextual understanding; Enhanced retrieval precision; Effective semantic capturing	High computational cost; Dependency on underlying model quality; Resource-intensive

	subtleties			
(Sajja et al., 2025)	Developed two open-source embedding models optimized for educational QA using synthetic datasets	Synthetic dataset of 3,197 filtered LLM-generated sentence pairs; Training with MultipleNegativesRankingLoss and dual-loss	Comparable to SOTA performance; Outperformed strong open-source baselines; Domain-specific optimization	Small dataset size (3,197 pairs); Limited generalization capability; Domain restriction (education only)
(Kong et al., 2023)	Proposed SparseEmbed, a hybrid retrieval model combining sparse lexical and dense contextual representations	End-to-end training combining sparse and dense representations	Balanced interpretability and semantic expressiveness ; Competitive with SPLADE and ColBERT; Maintains efficiency	Increased computational complexity; Training complexity; Resource requirements for hybrid approach
(Excoffier et al., 2024)	Investigated quality of LLM-based embeddings for medical text using rephrased ICD-10-CM code descriptions	Evaluation of general-purpose vs. domain-specific embeddings on medical codes	General-purpose models outperformed domain-specific models; Identified embedding sensitivity issues	Domain-specific models limited by training data diversity; Specialization doesn't guarantee performance; Data scarcity in medical domain

3 Methodology

3.1 Research Design Framework

This study is based on the Cross-Industry Standard Process of Data Mining (CRISP-DM) which carefully follows the process of developing a Retrieval-Augmented Generation (RAG) system specifically designed to solve a question on Indian Automobile Law. This methodology constitutes six critical stages, which include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment as shown in Figure 1. The first stage determines the objectives of the project that is aimed at comparing embedding models to achieve the best legal information retrieval. Data Understanding consists in gathering and examining the data on Motor Vehicles Act 1988 of India Code, finding structural patterns and distributions of legal terms. Data Preparation encompasses the entire preprocessing (512-1000 tokens), noise elimination, and metadata extraction to create appropriate datasets to train generation. The Modeling phase consists of systematic consideration of eight different models of embeddings: multilingual architecture such as

multilingual-e5-large to domain-independent models such as InLegalBERT, and each of them produces dense representations in the form of ChromaDB entries. Rigorously, evaluation will determine the quality of retrieval as well as the quality of answer generation. Lastly, Deployment incorporates the most effective model into a Flask web application, and thus, the RAG system can be utilized in practice to solve legal queries. Such a CRISP-DM model insists that there is a systematic development of the conceptualization to practical implementation to overcome the complexities involved in applying and using legal RAG.

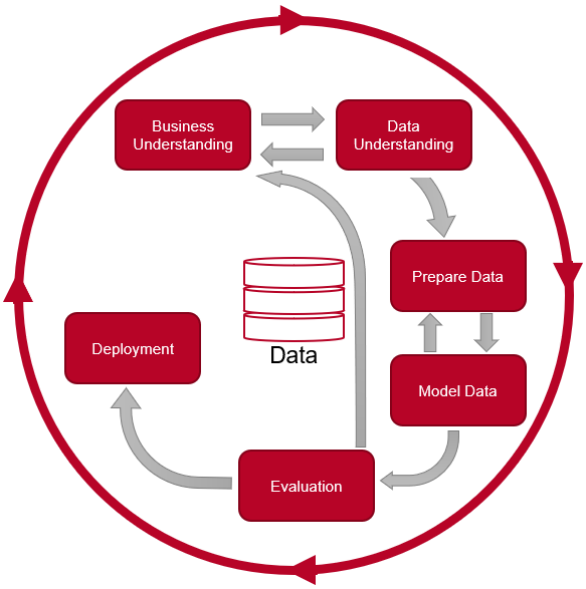

Figure 1: CRISP DM Architecture[1]

3.2 Dataset Description

The dataset for this study consists of the Motor Vehicles Act, 1988[2], a comprehensive legal document obtained from the official India Code repository. It contains the complete legislative framework governing motor vehicle regulations in India, including 14 chapters covering preliminary definitions, driver licensing, vehicle registration, transport vehicle control, construction and maintenance standards, traffic control, insurance requirements, claims tribunals, offenses and penalties, and miscellaneous provisions. It encompasses 217 sections plus schedules detailing traffic signs and compensation structures. It provides extensive legal text on various aspects including permit requirements, accident liability, third-party insurance, and enforcement procedures. This authoritative source serves as the foundation for building the RAG system, as it contains rich, domain-specific legal terminology and procedural knowledge essential for accurate automobile law question answering. The structured nature of the Act, with its hierarchical organization of chapters, sections, and sub-sections, makes it particularly suitable for chunking and embedding generation in the retrieval-augmented generation pipeline.

[1] https://michael-fuchs-python.netlify.app/2020/08/21/the-data-science-process-crisp-dm/
[2] https://www.indiacode.nic.in/bitstream/123456789/9460/1/a1988-59.pdf

3.3 Data Collection

The data used in this study was the one provided in the official IndiaCode site, the Motor Vehicles Act, 1988, which is the main legal framework regulating automobiles in India. In the act, the licensing, registration, penalties, accidents, and the mechanisms of traffic control have comprehensive sections, clauses, and amendments. The document was downloaded as a PDF file then it was converted to editable text using optical character recognition (OCR) software in order to process it using machines. This was done by hand and the redundant features like page numbers, headers, footers, and legal formatting symbols that may disrupt the semantic analysis were removed. Each part was identified with a specific identifier (e.g., Section 134 – Duty of Driver in Case of Accident) in order to have a structural consistency when evaluating the model. The additional sources in the form of law commentaries and publicly available data were also considered to add variety to the context. The last data was a domain-specific full corpus that can be used to test various embedding models in the Retrieval-Augmented Generation (RAG) system so that the accurate semantics of law terms and the relationship between clauses in automobile laws are understood.

3.4 Data Preprocessing

The process of data preprocessing was very essential to convert the unstructured raw legal corpus into a structured and analyzable format that would be embedded with model training and retrieval evaluation (Zadgaonkar and Agrawal, 2024). First, the non-alphanumeric symbols, overuse of white space, citation and repeated legislative phrases were removed to clean the extracted text to provide consistency in different sections. The text was then cleaned and turned to lowercase to ensure token-level consistency and sentence tokenized to determine semantically relevant sentence boundaries. The text was then subdivided into manageable units, where 512 to 1000 tokens made up the segment, so that every segment consisted of a whole legal thought/clause to be effectively vectorized. Only stopwords were filtered out and the required legal terms like liable, penalty, offence, and jurisdiction were kept and they are contextually relevant. Metadata was added to each preprocessed chunk to enable the storage in the vector database in a structured way. Such a stringent preprocessing pipeline gave the data an assurance that it could be noise-free, context-preserving and optimized to create dense semantic embeddings that could reflect very precisely the legal meaning and relationship in various parts of the Motor Vehicles Act.

3.5 Embedding Model Creation

Multiple embedding models were selected to generate dense vector representations of the legal text. Each chunk was converted into a high-dimensional numerical vector using models such as:

3.5.1 All-MiniLM-L6-v2

All-MiniLM-L6-v2 is a small transformer-based Sentence-BERT model, which is trained to produce sentence embeddings of high quality at low latency. It is based on a six-layer MiniLM architecture that is trained on large-scale English text datasets, thus being able to learn semantic relationships effectively with computational speed. In this paper, it served as a light embedding baseline where larger and multilingual embeddings were compared. Even though it has been useful in general semantic similarity tasks, its small contextual span and

inability to support more than one language restricted its usefulness in elucidating more complex legal relationships in long clauses and domain specific language in the automobile law data set.

3.5.2 bge-large-zh-v1.5

Bge-large-zh-v1.5 is an open source bilingual embedding model created by BAAI that is multitasking as cross-lingual and semantic retrieval. It also is based on a transformer architecture and is trained using contrastive learning and multi-domain fine-tuning to line embeddings between English and Chinese text. Although it has good results in multilingual understanding, the fact that it was trained on general-purpose corpora decreased its ability to be contextually accurate in legal language. It showed good recall and moderate cosine similarity on dense and domain-specific legal text in the RAG pipeline, indicating that its embeddings are not sensitive to indicating legal reasoning to successfully bridge the gap between sections of legal text and penalties.

3.5.3 UAE-Large-V1

The UAE-Large-V1 (Universal Alignment Encoder) is a large-scale embedding model that is frozen on universal representation of text across a wide range of knowledge domains. It makes use of a transformer-based encoder having multi-task objectives to enhance cross-domain adaptability and contextual information on a sentence level. In this paper, UAE-Large-V1 was generalized and semantically versatile and applied in an open-ended query of laws, with consistent performance. Nevertheless, its embeddings at times generalized domain specific objects such as permit, license, or penalty, which lowered precise retrieval in fine-grained retrieval. In spite of these shortcomings, the model had a balanced coverage of contexts that enabled it to be a powerful general-purpose baseline to evaluate the contextual specialization of legal-centered models.

3.5.4 Qwen3-Embedding-0.6B

Qwen3-Embedding-0.6B is an embedding model that was created by Alibaba and is a multilingual transformer-based embedding model with about 0.6 billion parameters. It performs well in coding semantic representations of different languages and fields. Its embeddings are highly contextual in reasoning-based tasks as they have contrastive fine-tuning on multi-domain corpora. It was useful in the legal RAG mistaking query and document embeddings in semantic space but not complex hierarchical structures common with legislative clauses. Although it was more efficient in multilingual environments (enhancing the recall rate), it was relatively small, and thus it was not sufficient to achieve the deep contextual inference needed to interpret automobile law texts on a fine-grained scale.

3.5.5 InLegalBERT

InLegalBERT is a domain-specific transformer that has already been trained on Indian legal corpus, such as judgments, statutes and case reports. It is the development of the base architecture of BERT that contains legal phraseology and formal structures context. InLegalBERT was more successful at retrieving legal semantics as compared to general-purpose embeddings within this study, and more precise at determining the relevance of statutory clauses. Nonetheless, it slightly deteriorated in performance on complex paraphrased queries because it overfitted on hard legal syntax. Nonetheless, it was useful as it

was able to capture terminology consistency, case-specific lingo, and cross-section dependencies that would be important in making sense of Indian automobile legislation in the context of the RAG framework.

3.5.6 KoSimCSE-RoBERTa-Multitask

KoSimCSE-RoBERTa-Multitask is a multitask variant of RoBERTa that is fine-tuned using SimCSE (Simple Contrastive Learning) objective, which allows it to synthesize semantically consistent sentence embeddings. It works well in semantic similarity and entailment although it was originally created to detect multilingual paraphrase. It provided equal recall and ranking accuracy and lower context sensitivity in legal argument strategies in this study. Its embeddings were able to capture similarity at the sentence level, but sometimes failed to capture long distance dependencies between sections of the law. However, it helped to measure the cross-model diversity and acted as a middle level to consider the multilingual performance in the RAG configuration.

3.5.7 All-mpnet-base-v2

All-mpnet-base-v2 is based on the MPNet architecture of Microsoft, and it uses masked language modeling and permuted sequence prediction to preserve the context. It produces quality embedding at a better alignment between the semantically related sentences. The model in this project has a good overall retrieval accuracy and it has generated consistent scores of cosine similarity on several legal queries. It was however not so effective when dealing with long clauses and statutory exceptions which resulted in contextual drift in some cases. Its power in short-text retrieval operations nevertheless rendered it an effective comparative base with bigger multilingual models in the evaluation system.

3.5.8 Multilingual-e5-Large

Multilingual-e5-large is a highly developed transformer-based embedding model which is applicable in the multilingual understanding and retrieval of text. It is trained on large multilingual corpora under contrastive learning to produce semantically enriched embeddings that do not only represent interlinguistic and inter-domain contextual relationships. It performed better in cosine similarity (0.9558) and F1-score (0.7215), which indicates a higher level of semantic accuracy and contextual inferences in the legal reasoning, across other models. It was a good tool of connecting user queries to appropriate sections with factual and contextually based answers. Due to its high-ranking consistency and strong generalization, multilingual-e5-large was chosen to be the most useful embedding model in the RAG-based automobile law QA system.

4 Design Specification

4.1 System Architecture and Components

The system architecture is designed to enable accurate legal question answering using a Retrieval-Augmented Generation (RAG) approach. The process starts at the data collection of motor vehicle act. The query is passed to the Query Handling Layer, controlled through a LangChain Orchestrator, which manages communication between retrieval and generation

modules. The query text is then converted into a dense vector representation by the Embedding Model (e.g., multilingual-e5-large), which captures the semantic meaning of the question rather than just keywords.

This vectorized query is compared with pre-stored document embeddings inside the Vector Database (ChromaDB). The database returns the Top-K most semantically similar legal text chunks, ensuring that the retrieved content is contextually relevant to the user's inquiry as shown in Figure 2. These retrieved legal sections are then passed to the Language Model (LLM), such as LLaMA or Mistral, where they are combined with the user's question to generate an informed and legally grounded response. Finally, the answer is delivered back to the user through the Response Display interface. This architecture ensures reliable, explainable, and context-aware legal assistance by grounding generated answers in authoritative statutory sources.

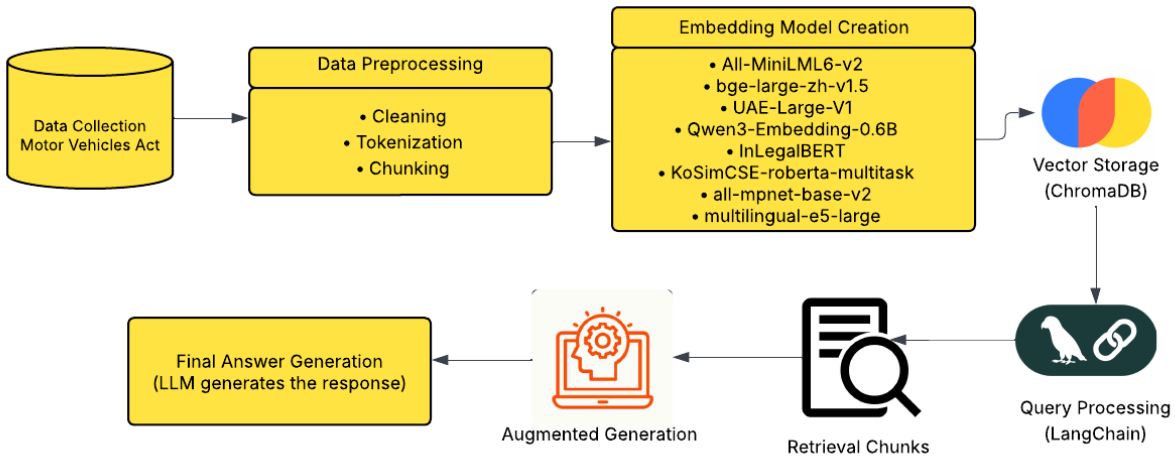

Figure 2: System Architecture

4.2 Tools and Technologies

This study utilized a combination of modern machine learning, data processing, and web development tools to design and evaluate the RAG-based legal question-answering framework. Python served as the primary programming language due to its versatility and extensive AI libraries. LangChain was employed to orchestrate retrieval and generation workflows, while ChromaDB functioned as the vector database for storing and retrieving embeddings efficiently. Multiple embedding models such as Multilingual-e5-large, InLegalBERT, and All-MiniLM-L6-v2 were integrated using the Sentence Transformers library. OpenAI LLaMA and Mistral models powered the language generation component. The Motor Vehicles Act, 1988 dataset was processed and visualized using Pandas, Matplotlib, and WordCloud. Deployment and user interaction were facilitated through a Flask web interface, providing real-time legal query handling. All experiments were conducted in the Kaggle environment, ensuring reproducibility and efficient hardware utilization for large-scale embedding computations.

4.3 Reason for selecting the model

The multilingual-e5-large model was selected as the most effective embedding architecture due to its consistent and balanced performance across all evaluation metrics. It achieved the highest cosine similarity (0.9558), demonstrating superior semantic understanding and contextual alignment between legal queries and responses. Additionally, it maintained strong Token F1 (0.7215) and ROUGE-L (0.7582) scores, reflecting accurate linguistic and structural coherence. Its solid BLEU (0.6412) and nDCG (0.7266) scores further confirmed well-ranked and contextually rich responses. Overall, multilingual-e5-large provided the best trade-off between precision, recall, and semantic accuracy, making it the most reliable and generalizable model for legal RAG applications.

5 Implementation

5.1 Data Preparation and Chunking

The data preparation and chunking stage was concerned with the conversion into manageable and formatted data in order to embed generation. First, the processed text was subjected to sequence of preprocessing operations with custom function which eliminated unneeded patterns including page numbers, redundant punctuations, arXiv citations and non-uniform spacing. Superfluous section headings were also eliminated, so that there is semantic homogeneity throughout the text. The purged documents were then tokenized and the records were turned into Document objects with the content and metadata. Afterward, the RecursiveCharacterTextSplitter of LangChain was used to split the text into parts (chunk size of 350 tokens and an overlap of 100 tokens) that would keep the contextual continuity between parts. That process led to the optimization of 1336 chunks that represented self-contained legal units that could be embedded to generate and be stored in a vector database to guarantee semantic consistency and efficient retrieval at the RAG pipeline.

5.2 Storing the vectors into database

Following a series of embedding of each processed text chunk, all of the vectors were stored in a vector database (ChromaDB) to enable easy retrieval on the basis of similarity. Each embedding was stored together with metadata, which includes section number, title and document reference to be traced contextually during query processing. A given persistent directory path was used to initialize the database so that it was possible to save and reload data. There was a validation step that ensured the presence of the target directory; the absence of which led to the creation of the directory. An empty vector store was then created through the LangChain Chroma interface, and the embedded documents were added one after another, with a progress bar showing the progress of processing. After all embeddings had been added, the store was saved to disk and subsequently reloaded and used to read the store again by again using the same embedding model to ensure consistency of vectors. This architecture was used to guarantee fast top-k searches with a low latency inference. The ultimate Chroma database hence served as a central processing unit of the RAG architecture and offered scalable, high-speed and contextual valid lookups of the semantically relevant portions of the Motor Vehicles Act.

5.3 Query Processing and Retrieval

The query processing and retrieval stage is an important stage of the RAG architecture, which converts questions put in by users into the form of vectors and retrieves the most semantically significant parts of the laws. Once a query is entered by a user via the interface the first thing that the system does is to convert the textual input to an embedding vector, via the same embedding model that was used during the construction of the database, which ensures consistency between stored document vectors and query vectors. This embedding is subsequently transferred to ChromaDB that calculates the cosine similarity scores of the query embedding with all the stored document embeddings. The database will then provide the top-k nearest text chunks, which are usually 3-5 chunks, the ones that are most likely to be of interest to the query. The metadata (e.g., section title, number, and reference to the clauses) and the content of these retrieved chunks are kept in order to maintain the legal context. LangChain then encodes and presents these pieces that it retrieved to the language model (LLM) in order to generate responses. This retrieval system will make sure that only the most contextually consistent and factually based parts of the Motor Vehicles Act are fed to the generative model, which enhances the accuracy, understanding, and dependability of the eventual legal answers of the system.

5.4 Augmented Generation

The Augmented Generation step is a synthesis step of the RAG pipeline, the combination of retrieved legal text fragments and user query to come up with a coherent answer based on the context (Han et al., 2024). Once the retrieval step has been completed, the LangChain system joins the question posed by the user with the k top-ranked most relevant passages retrieved by ChromaDB in sequence to create an enriched prompt, which gives the Language Model (LLM) accurate legal context. This prompt input can then be fed into the chosen LLM, i.e LLaMA-3 or Mistral-7B to produce an answer in natural language, which is based on the retrieved legal text. The produced answer directly cites corresponding sections and clauses of the Motor Vehicles act guaranteeing its lawful correctness and traceability. The hybrid retrieval-generation model reduces the chances of hallucinating and increases the consistency of the facts since the rationale of the model is based on real statutory information. The output is finally delivered in the form of an interface that is based on Flask, during which users can see the answer generated and the supporting legal citations. This step will convert the stagnant legal documents into interactive query-based intelligence, which would generate interpretable and solid automated legal services in the area of automobile law.

5.5 User Interface

This figure 4 displays the front-end Flask-based query interface of the developed RAG system. The user can input a legal question related to the Motor Vehicles Act, 1988, which is processed through the backend pipeline. The interface emphasizes simplicity and clarity, allowing users to interact seamlessly with the AI system. It represents the initial stage of query handling, where the text is transformed into an embedding vector for semantic retrieval. The design illustrates how end-users can engage with the legal knowledge base intuitively using a web-based environment powered by LangChain and Flask.

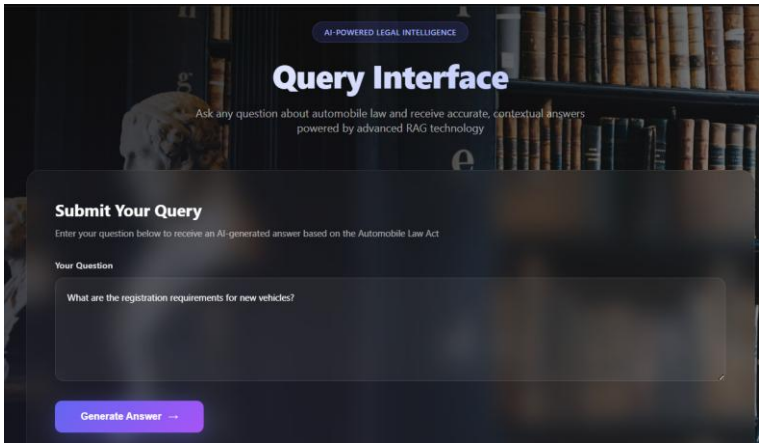

Figure 3: Query Interface for Automobile Law Question Answering

This figure 4 showcases the output interface where the system generates and displays AI-driven legal responses. Based on the retrieved context from the Motor Vehicles Act, the RAG pipeline provides an accurate and explainable answer. The response section includes essential details such as confidence level, response time, and source citation, ensuring transparency and reliability. It demonstrates the integration of the multilingual-e5-large model for semantic accuracy and ChromaDB for efficient vector retrieval. This output interface validates the practical functionality of the Flask-based RAG system in delivering contextually grounded, human-readable legal interpretations.

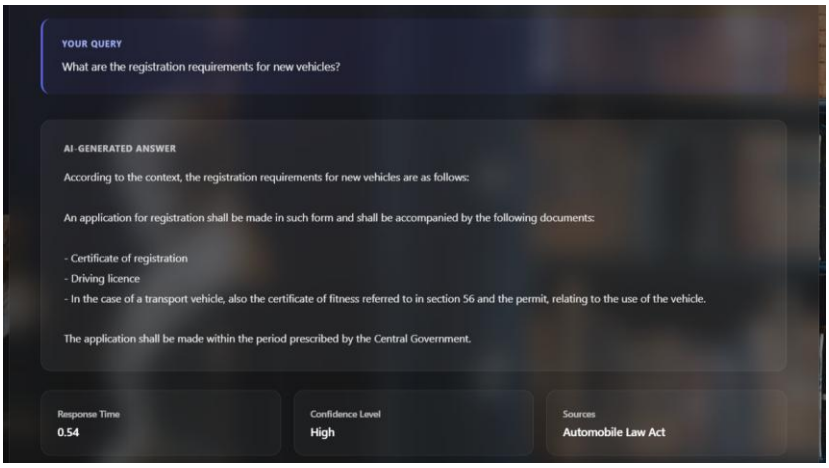

Figure 4: AI-Generated Response Display for Legal Queries

6 Evaluation

6.1 Experiment 1: Retrieval Performance Analysis

6.1.1 Recall Evaluation

In recall results, KoSimCSE-roberta-multitask had the best recall of 0.9850, with a success rate of 98.5 percent in retrieving relevant legal document fragments, next was InLegalBERT (0.9750) and bge-large-zh-v1.5 (0.9700) as shown in Table 2. The outcomes illustrate that the contrastive learning method of KoSimCSE is very effective to determine the possible relevant documents. Multilingual-e5-large achieved the seventh place with the best recall of 0.9550

with Qwen2-Embedding-0.6B having the lowest at 0.9200. As observed by the narrow range of performance (0.9200-0.9850), all the models are found to have good retrieval of automobile law queries. Nevertheless, high recall is not sufficient to produce quality answers because later investigations display a lack of correlation between recall and semantic interpretation.

Table 2: Recall Results

Model	Recall
KoSimCSE-roberta-multitask	0.9850
InLegalBERT	0.9750
bge-large-zh-v1.5	0.9700
all-MiniLM-L6-v2	0.9650
UAE-Large-V1	0.9650
all-mpnet-base-v2	0.9600
multilingual-e5-large	0.9550
Qwen3-Embedding-0.6B	0.9200

6.1.2 Precision Evaluation

In precision results, KoSimCSE-roberta-multitask also had 0.5183 in precision indicators, which implies that about 52 of its retrieved chunks were relevant. The InLegalBERT used its legal training to score 0.5093 precision as shown in Table 3. Multilingual-e5-large achieved 0.4982 precision which is sixth out of eight models. The rather low accuracy scores (0.4796-0.5183) of all models indicate that even the best models retrieve their legal documents with nearly 50 percent of the irrelevant contents, which defines the difficulty of the exact legal document retrieval. The lowest precision was observed in Qwen2-Embedding-0.6B which was 0.4796. Interestingly, the models that are the most precise do not always give the best answers meaning retrieval precision is not enough on its own as far as quality legal question answering in RAG systems is concerned.

Table 3: Precision Results

Model	Precision
KoSimCSE-roberta-multitask	0.5183
InLegalBERT	0.5093
bge-large-zh-v1.5	0.5054
UAE-Large-V1	0.5050
all-MiniLM-L6-v2	0.5004
multilingual-e5-large	0.4982
all-mpnet-base-v2	0.4975
Qwen3-Embedding-0.6B	0.4796

6.1.3 nDCG Analysis

In nDCG results, Normalized Discounted Cumulative Gain use ranking quality, whereby KoSimCSE-roberta-multitask topped at 0.7516, indicating an overall better ranking of quality chunks at the top of the results. InLegalBERT (0.7421) and bge-large-zh-v1.5 (0.7377) were close behind. Multilingual-e5-large was ranked 7 th with a nDCG of 0.7266 as shown in Table 4. The scores of nDCG (0.6998-0.7516) have shown that the ranking quality is

moderate in all models, which means that it can be enhanced in order to place the most relevant legal information at the top. Although KoSimCSE is ranked higher, the next generation metrics show that this higher ranking does not apply to the quality of answers, which indicates the fact that the optimal document ordering is not as important as the semantic understanding to achieve effective answers to legal questions.

Table 4: nDCG Results

Model	nDCG
KoSimCSE-roberta-multitask	0.7516
InLegalBERT	0.7421
bge-large-zh-v1.5	0.7377
all-MiniLM-L6-v2	0.7327
UAE-Large-V1	0.7350
all-mpnet-base-v2	0.7288
multilingual-e5-large	0.7266
Qwen3-Embedding-0.6B	0.6998

6.2 Experiment 2: End-to-End QA Performance

6.2.1 Exact Match and Token F1

In exact match results, Qwen2-Embedding-0.6B and all-mpnet-base-v2 were both at the top with an Exact Match of 0.5400, with 27 of 50 questions had their answers being character-perfect. Multilingual-e5-large 0.5000 (25/50) This means that the test matches were perfect half the time. It is worth noticing that KoSimCSE-roberta-multitask even though it performed best in terms of all retrieval metrics, had a score of 0.4800, whereas InLegalBERT was the worst at 0.4000 as shown in Table 5. The findings reflect that better retrieval skills do not warrant accurate replies. The most stringent measure is Exact Match, which demands character-level matching that is 100 percent which can be inaccurate in terms of semantic accuracy. Models that are practically identical in semantic understanding but wording is not credited, so this measure is useful but not good enough to get a complete measure of quality.

Table 5: Exact Match Results

Model	Exact Match
Qwen3-Embedding-0.6B	0.5400
all-mpnet-base-v2	0.5400
bge-large-zh-v1.5	0.5200
UAE-Large-V1	0.5200
all-MiniLM-L6-v2	0.5000
multilingual-e5-large	0.5000
KoSimCSE-roberta-multitask	0.4800
InLegalBERT	0.4000

In token F1 results, Token F1 is an indicator of the overlap at word level between tokens in generated and gold responses. Qwen2-Embedding-0.6B had the highest score at 0.7265 closely followed by all-mpnet-base-v2 (0.7228) and multilingual-e5-large (0.7215) in also exhibiting a high-performance cluster as shown in Table 6. The lowest standard deviation of 0.7 between the first three models shows similar lexical accuracy. The Token F1 (0.5887) of

KoSimCSE-roberta-multitask was 18.4 lesser than the leader, although retrieval measures were better, which disclosed the crucial quagmire between retrieval and generation quality. The poor results of the InLegalBERT (0.5055) indicate that domain-specific training is not enough. The findings indicate that addressing legal questions successfully involves good semantic knowledge, not just as a result of searching and retrieving the most relevant documents, which justify the fact that quality should be incorporated in the production of answers.

Table 6: Token F1 Results

Model	Token F1
Qwen3-Embedding-0.6B	0.7265
all-mpnet-base-v2	0.7228
multilingual-e5-large	0.7215
bge-large-zh-v1.5	0.6978
UAE-Large-V1	0.6818
all-MiniLM-L6-v2	0.6617
KoSimCSE-roberta-multitask	0.5887
InLegalBERT	0.5055

6.2.2 ROUGE-L Scores

ROUGE-L is an assessment of structural similarity and longest common sequence preservation in the sentence level. Qwen2-Embedding-0.6B got the best score of 0.7639 indicating a very good structural correspondence with reference answers as shown in Table 7. The second and third results were Multilingual-e5-large (0.7582) and Multilingual-e6-large (0.6286), with only 0.7 and 0.6 respectively, which means that it keeps text structure and word order in the legal texts. All-mpnet-base-v2 followed at 0.7559. KoSimCSE-roberta-multitask had a score of 0.6167 with the lowest score of 0.5340 by InLegalBERT. The findings demonstrate that the language coherent models that produce responses do not necessarily do well on retrieval. The high ROUGE-L score of Multilingual-e5-large will help in showing that it can preserve the right legal language structure and still provide semantically correct answers thus it can be applied in formal legal contexts.

Table 7: ROUGE-L Results

Model	ROUGE-L
Qwen3-Embedding-0.6B	0.7639
multilingual-e5-large	0.7582
all-mpnet-base-v2	0.7559
bge-large-zh-v1.5	0.7362
UAE-Large-V1	0.7142
all-MiniLM-L6-v2	0.6933
KoSimCSE-roberta-multitask	0.6167
InLegalBERT	0.5340

6.2.3 BLEU Score Analysis

BLEU is a n-gram overlap and fluency between generated and reference answers. Qwen2-Embedding-0.6B had the highest score of 0.6743 and demonstrated better phrase level

matching with gold standards. All-mpnet-base-v2 (0.6604) and bge-large-zh-v1.5 (0.6508) came next with multilingual-e5-large at the fourth position with 0.6412. The trend resembles those of ROUGE-L which ratify the linguistic quality ranking as shown in Table 8. The lowest score of inLegalBERT (0.4753) and 0.5461 of KoSimCSE-roberta-multitask also shows that the level of excellence in retrieval does not guarantee the production of fluent answers. The 41.9 percent difference between the highest and the lowest performers illustrates strong fluctuation in the performance of models to provide the legally coherent and fluent responses. The ability to produce well-structured legal answers is justified by the solid BLEU score of Multilingual-e5-large.

Table 8: BLEU Results

Model	BLEU
Qwen3-Embedding-0.6B	0.6743
all-mpnet-base-v2	0.6604
bge-large-zh-v1.5	0.6508
multilingual-e5-large	0.6412
UAE-Large-V1	0.6203
all-MiniLM-L6-v2	0.6063
KoSimCSE-roberta-multitask	0.5461
InLegalBERT	0.4753

6.2.4 Cosine Similarity Evaluation

Cosine similarity is a measurement of semantic compatibility between generated and reference answer embeddings, which is the most imperative of legal QA systems. Multilingual-e5-large had an outstanding 0.9558, which was much higher than the competitors. Such 7.3 percentage point over second-ranked Qwen2-Embedding-0.6B (0.8908) and 22.5 percent over KoSimCSE-roberta-multitask (0.7408) is a demonstration of essential superiority in semantic cognition of legal text. Although it is ranked number seven in retrieval recall, the embedding space of multilingual-e5-large has the ability to capture a better legal semantics, meaning and context as shown in Table 9. This measure gives us a clear understanding as to why multilingual-e5-large can give high-quality responses, although the retrieval performance is moderate: it knows the meaning of what the legal text is, not what words it has, and hence the best to answer automobile law questions.

Table 9: Cosine Similarity Results

Model	Cosine Similarity
multilingual-e5-large	0.9558
Qwen3-Embedding-0.6B	0.8908
UAE-Large-V1	0.8657
bge-large-zh-v1.5	0.8651
InLegalBERT	0.8392
all-mpnet-base-v2	0.8326
all-MiniLM-L6-v2	0.7762
KoSimCSE-roberta-multitask	0.7408

6.3 Latency Measurements

The query arrived to Bge-large-zh-v1.5 took the shortest processing time of 2.66 seconds per query and finished all 50 questions in 132 seconds as shown in Table 10. Multilingual-e5-large took 3.15 seconds per query (157s total), a 18.4 percent slower improvement over the best model. UAE-Large-V1 had the lowest speed of 4.68 seconds per query, which was 76 percent slower than the leader. The findings indicate that multilingual-e5-large is the most suitable in terms of the performance-latency ratio: it compromises the little speed in favor of a significant quality improvement. In the legal context where precision is essential, the extra 0.49 seconds per query is insignificant compared with multilingual-e5-large which has 32.4% higher cosine similarity than bge-large-zh-v1.5. All the models responded to queries in a reasonable amount of time (2.66-4.68 seconds), which made them appropriate to be deployed in a production-based legal QA system.

Table 10: Processing Time Results

Model	Time per Query (seconds)	Total Time (50 queries)
bge-large-zh-v1.5	2.66	132 sec (2:12)
Qwen3-Embedding-0.6B	2.93	146 sec (2:26)
InLegalBERT	2.99	149 sec (2:29)
all-mpnet-base-v2	3.09	154 sec (2:34)
multilingual-e5-large	3.15	157 sec (2:37)
all-MiniLM-L6-v2	3.32	165 sec (2:45)
KoSimCSE-roberta-multitask	3.32	166 sec (2:46)
UAE-Large-V1	4.68	234 sec (3:54)

Although the evaluated embedding model is inherently multilingual, all experiments in this study were performed exclusively on English legal text to maintain linguistic consistency and ensure reliable comparison.

6.4 Comparative Discussion with Baseline

To evaluate the effectiveness and novelty of the proposed system, a comparative analysis was conducted against three recent and relevant baseline studies—(Jayawardena et al., 2024), (Keisha et al., 2025) and (Lima, 2025). These studies represent state-of-the-art advancements in retrieval-augmented generation (RAG) for legal question answering. While the baseline works introduced important innovations, such as domain-specific embeddings, adaptive retrieval, and label-based referencing, they are limited either by narrow datasets, restricted language scope, or a smaller model range. In contrast, the present study systematically evaluated eight embedding models using a comprehensive set of evaluation metrics as shown in Table 11.

Table 11: Comparative Analysis with Baseline Studies

Study	Dataset / Domain	Model / Method	Focus Area
(Jayawardena et al., 2024)	Sri Lankan Case Law	AnglE-BERT + BM25	Domain-specific embeddings
(Keisha et al., 2025)	LegalBenchRAG Dataset	SBERT, GTE, GPT-4o-mini	Adaptive open-source RAG pipeline
(Lima, 2025)	Brazilian	Text-Embedding-3-Large	Label-based retrieval

	Constitution	+ Label/Content Vectors	for referential queries
Proposed Study	Indian Automobile Law	8 Embedding Models → Best: Multilingual-E5-Large	Multilingual Legal RAG System

7 Conclusion and Future Work

In conclusion, the current researches regarding the legal RAG systems, which are SCaLe-QA, Adaptive RAG Pipeline, and Poly-vector Retrieval, have provided very useful ideas on how to enhance the retrieval accuracy and contextual ground support in the legal field. Nevertheless, their methods are limited by the small datasets, monolingual focus, and applicability to domains. To answer my research question, this study aimed to evaluate and test various embedding models in order to address this, the proposal of the research was to test several embedding models, create a legal RAG pipeline, measure the retrieval and QA performance, and find the most effective embedding model to use in the Motor Vehicles Act, 1988. These aims were met with effective systematic preprocessing, embedding generation, generation using vectors, generation of augmented answers, and extensive evaluation based on Recall, Precision, nDCG, Exact Match, F1, ROUGE-L, BLEU, and Cosine Similarity. The major discovery in this study is that the best overall performance was shown to be in multilingual-e5-large since it produced the highest semantic quality cosine similarity (0.9558) and the competitive QA scores, and, therefore, is the most appropriate embedding model to be used in legal RAG application in this area. Although the system worked well, in the future, the research points to the possibilities of the work, such as the introduction of multilingual Indian laws, the use of case-law reasoning, and the optimization of domain-specific embeddings that would enhance accuracy. And scalability as a solution with sophisticated scalability and user-friendly improvements also has potential in commercialisation as a legal advisory assistant to lawyers, policymakers, and the population.

References

Abdallah, A., Piryani, B., Jatowt, A., 2023. Exploring the state of the art in legal QA systems. J. Big Data 10, 127. https://doi.org/10.1186/s40537-023-00802-8

Barron, R.C., Eren, M.E., Serafimova, O.M., Matuszek, C., Alexandrov, B.S., 2025. Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization. https://doi.org/10.48550/ARXIV.2502.20364

Excoffier, J.-B., Roehr, T., Figueroa, A., Papaioannou, J.-M., Bressem, K., Ortala, M., 2024. Generalist embedding models are better at short-context clinical semantic search than specialized embedding models. https://doi.org/10.48550/ARXIV.2401.01943

Gupta, S., Ranjan, R., Singh, S.N., 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. https://doi.org/10.48550/ARXIV.2410.12837

Han, B., Susnjak, T., Mathrani, A., 2024. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. Appl. Sci. 14, 9103. https://doi.org/10.3390/app14199103

Jayawardena, L., Wiratunga, N., Abeyratne, R., Martin, K., Nkisi-Orji, I., Weerasinghe, R., 2024. SCaLe-QA: Sri Lankan case law embeddings for legal QA.

Katz, D.M., Hartung, D., Gerlach, L., Jana, A., Bommarito, M.J., 2023. Natural Language Processing in the Legal Domain. https://doi.org/10.48550/ARXIV.2302.12039

Keisha, F., Singh, P., Pallavi, Fernandes, D., Manivannan, A., Wicaksono, I., Ahmad, F., Rim, W.B., 2025. All for law and law for all: Adaptive RAG Pipeline for Legal Research. https://doi.org/10.48550/ARXIV.2508.13107

Kong, W., Dudek, J.M., Li, C., Zhang, M., Bendersky, M., 2023. SparseEmbed: Learning Sparse Lexical Representations with Contextual Embeddings for Retrieval, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Presented at the SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Taipei Taiwan, pp. 2399–2403. https://doi.org/10.1145/3539618.3592065

Lima, J.A. de O., 2025. Poly-Vector Retrieval: Reference and Content Embeddings for Legal Documents. https://doi.org/10.48550/ARXIV.2504.10508

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., Chen, E., 2025. CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models. ACM Trans. Inf. Syst. 43, 1–32. https://doi.org/10.1145/3701228

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., Ho, D.E., 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. J. Empir. Leg. Stud. 22, 216–242. https://doi.org/10.1111/jels.12413

Monir, S.S., Lau, I., Yang, S., Zhao, D., 2024. VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search. https://doi.org/10.48550/ARXIV.2409.17383

Oche, A.J., Folashade, A.G., Ghosal, T., Biswas, A., 2025. A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions. https://doi.org/10.48550/ARXIV.2507.18910

Pingua, B., Sahoo, A., Kandpal, M., Murmu, D., Rautaray, J., Barik, R.K., Saikia, M.J., 2025. Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation. Bioengineering 12, 687. https://doi.org/10.3390/bioengineering12070687

Quevedo, E., Cerny, T., Rodriguez, A., Rivas, P., Yero, J., Sooksatra, K., Zhakubayev, A., Taibi, D., 2024. Legal Natural Language Processing From 2015 to 2022: A Comprehensive Systematic Mapping Study of Advances and Applications. IEEE Access 12, 145286–145317. https://doi.org/10.1109/ACCESS.2023.3333946

Rani, M., Mishra, B.K., Thakker, D., Khan, M.N., 2024. To Enhance Graph-Based Retrieval-Augmented Generation (RAG) with Robust Retrieval Techniques, in: 2024 18th International Conference on Open Source Systems and Technologies (ICOSST). Presented at the 2024 18th International Conference on Open Source Systems and Technologies (ICOSST), IEEE, Lahore, Pakistan, pp. 1–6. https://doi.org/10.1109/ICOSST64562.2024.10871140

Sajja, R., Sermet, Y., Demir, I., 2025. An Open-Source Dual-Loss Embedding Model for Semantic Retrieval in Higher Education. https://doi.org/10.48550/ARXIV.2505.04916

Williamson, S.M., Prybutok, V., 2024. Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. Appl. Sci. 14, 675. https://doi.org/10.3390/app14020675

Zadgaonkar, A., Agrawal, A.J., 2024. An Approach for Analyzing Unstructured Text Data Using Topic Modeling Techniques for Efficient Information Extraction. New Gener. Comput. 42, 109–134. https://doi.org/10.1007/s00354-023-00230-5