National College of Ireland

Enhancing Legal Research Efficiency through Retrieval-Augmented Generation (RAG) with Vector Search

MSc Research Project
Artificial Intelligence

Keerthana Venkatesan

Student ID: x23275243

School of Computing
National College of Ireland

Supervisor: Arundev Vamadevan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Keerthana Venkatesan

Student ID: X23275243

Programme: MSc in Artificial Intelligence **Year:** 2024-2025

Module: Research Project

Supervisor: Arundev Vamadevan

Submission Due Date: 11/08/2025

Project Title: Enhancing Legal Research Efficiency through Retrieval-Augmented Generation (RAG) with Vector Search

Word Count: **8387** **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: Keerthana Venkatesan

Date: 11/08/2025

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):	

Enhancing Legal Research Efficiency through Retrieval-Augmented Generation (RAG) with Vector Search

Keerthana Venkatesan

x23275243

Abstract

The legal retrieval of information has unique problems because of the complex legal language used, citation formats, and requirements of specific output with even explanations. In this paper, discussed the creation of a legal chatbot assistant based on Retrieval-Augmented Generation (RAG) together with domain-oriented large language models that respond to legal questions with references to documents uploaded in the process. To achieve such a contextual search, PDF files are semantically divided into chunks and indexed with the help of vector embeddings in a Qdrant vector database and stored. There is text retrieval and resulting grounded responses in real-time to the queries. Also, evaluate different subsystems in the pipeline, like the effectiveness of text chunking, trade-offs in optimising embedding, and latency as far as document size is concerned. An efficient vector representation was provided by sentence-transformers/all-MiniLM-L6- v2, which is a fine-tuning of the embedding model that resulted in high retrieval recall. Latency tests were scalable, with less than 1.2 seconds obtained as the average response time on querying a document library with a size of 100+ documents. To test the performance of the system, various metrics, which included ROUGE-L, BLEU, latency, and retrieval hit rate, were used on a benchmark question-answering dataset involving legal questions. It proves that it was accurate and relevant. Visualization was used to read the performance of the system. This system shows the possibility of the usage of AI-driven means of legal support that guarantees quicker, verifiable, and easily accessible legal support.

Keywords- Legal AI, Legal Question Answering, Retrieval-Augmented Generation, Vector Database, Qdrant, Legal Chatbot, Large Language Models, ROUGE-L, BLEU, latency

1 Introduction

Artificial Intelligence (AI) is an area of development that has exploded over the last several years, changing the landscape of many professional fields, including law. Previously, the study of the legal field entailed scouring through extensive and intricate sources like statutes, regulations, case law, legal commentaries, and academic articles. The immense size and variety of these legal materials make them inefficient in the manner of their manual search, since they are usually time-consuming and inefficient, and affect the efficiency of the law practice Surden.et (2019). As the online legal databases and repositories grow, there is an immediate need to develop more advanced tools that will be capable of coping with the complexity of such databases by enabling the effective and successful retrieval of relevant legal material.

Practitioners in the legal field are often challenged on how to search through the big corpus of legal material to determine the relevant precedent or authoritative interpretation, which is particularly monumental due to the complexity of the language, jurisdiction detail, and the common updates in the legal materials Ashley et.al., 2017). Though useful, the conventional keyword-based search engines are not always sufficient owing to their inability to embody the semantics and complex legal arguments that these records carry. This weakness leads to the retrieval of insufficiently contextually accurate documents that are superficially related, which may mislead or deteriorate the legal decision-making Grabmair et.al., 2017). Developments of AI-based legal assistants, i.e., chatbots, have been reported to address some of these challenges, as they can cover various functionalities such as legal research automation, document summarization, and answering questions Ashley et.al., 2017).

The introduced RAG models are more effective than traditional search tools based on keywords, as they also take into account dense vector representations and embeddings, which are based on the semantic meaning of the words and use them to get more specific and context-sensitive information Karpukhin et a.., 2020).

The main objective of this project is a Legal Chatbot Assistant based on a RAG pipeline, which combines the use of a vector similarity search with LLMs developed with high performance, namely Flan-T5 and LLaMA-3, Touvron et al., (2023). This chatbot dynamically consumes PDF documents, semantically chunks, and embeds contents and gives precise, source-linked answers to natural language legal questions. The dual-model comparison structure of the system enables users to monitor the quality of answers, their latency, and factuality simultaneously, stimulating transparency and explainability as per the current trends in Explainable AI (Doshi-Velez and Kim, 2017).

Research Question: How does Retrieval-Augmented Generation (RAG) with vector search enhance the accuracy and efficiency of legal content retrieval compared to traditional keyword-based legal search systems?

This study aims to answer the following key objectives to answer the research question:

- Semantic Understanding: RAG with vector search does not search by keywords but by the meaning of the query, and therefore, it is more relevant even in the case of varying wording.
- Contextual Accuracy: The generation component sums up or answers questions based on the information directly taken from the most relevant documents, or makes the response based on real legal content.

- Maximum efficiency: RAG takes less time to scan through numerous documents as opposed to a keyword-based system because it brings out the direct and concise responses when considering large bodies of legal texts.

The structure of the remaining sections of this research paper is as follows: Section 2 contains the discussion of work related to the idea of using Retrieval-Augmented Generation (RAG) and purpose-specific Large Language Models (LLMs) in the legal sphere. In Section 3, the proposed architecture of a legal chatbot assistant that is based on the use of vector-based

document retrieval and domain-specific models are presented. The document indexing process, the embedding models, and the vector store management are identified in Section 4, which discusses the design specifications. The process of implementing the RAG pipeline and combining it with Flan-T5 and LLaMA-3 models is explained in Section 5. Section 6 gives a methodology of evaluation, performance measures, and outcomes in terms of the comparison of the output of two models. Lastly, a summary of the paper is put down with key findings and future directions of work in Section 7.

2 Related work

Artificial intelligence and technology have undergone rapid development in the last ten years, and natural language processing (NLP) and retrieval-augmented generation (RAG) systems have become game changing technologies in the access and analysis of legal information. The legal field introduces special considerations that separate it from other application areas. The complexity and exact nature of legal text, and the crucial importance of proper source attribution. Awareness of jurisdiction; The necessity of an explainable decision-making process.

The present literature review examines the degrees of research in significant fields that form the foundation blocks of the current legal AI systems, namely transformer-based legal language models, retrieval-augmented generation systems, vector databases and semantic embeddings, legal NLP datasets and benchmarking, explainability and trust in legal AI systems, integration issues and system architecture, and new trends in legal AI applications. The review compiles the existing information in modern studies to define the theory in the construction of advanced legal chatbot assistants and reveals the major gaps that contemporary studies can propose.

2.1 Transformer-Based Legal Language Models

Domain-specific variants of the transformer model have completely transformed natural language processing in numerous application areas, with the legal sphere being no exception; it offers several special opportunities and faces special challenges. The first decades of this research initially began with the general-purpose models such as BERT and GPT; however, it quickly became apparent that targeting domain-sensitive solutions is necessary due to the nature of legal language. Legal-BERT, introduced by Chalkidis et al. (2020), is one of the first pre-training attempts to learn a particular variant of BERT using large legal text collections.

Chalkidis et al. (2020) introduce Legal-BERT, among the first attempts of such systematic endeavours as pre-training a certain version of BERT on big legal data. Their model was trained on a variety of European Union laws, UK laws, court cases, and legal scholarly literature, amounting to approximately 12GB. The empirical comparison showed that Legal-BERT outperformed state-of-the-art results based on general-purpose language models and achieved equivalent levels of success to deep legal embedding, manifesting 3-5 gains in legal text classification, 4-7 in named entity recognition of entities in law, and vastly superior outcomes in the legal document similarity task.

Nonetheless, the authors recognized several limitations, including the geographic specificity of the model to legal systems in Europe, possible bias towards types of legal documents, and the complexity of computation. Katz et al. (2023) performed a thorough survey of the implementations of natural language processing in the legal practice field and found three decisive issues, the lack of data, the interpretability of predictive models, and the question of ethics of automatic legal decision-making. After studying them, they determined that transformer models can be considered promising, but the legal sphere is incredibly complex, and it would be necessary to consider specific features of the language and the system of regulations, depending on the jurisdiction.

The survey has pointed out the gap existing between possible technical functionality and legal reality, with the fact that models are needed that can support subtle reasoning and interpretation that form the core of legal practice. The state of law-specific language modelling was studied by Singh and Kumar (2024), who thoroughly surveyed large language models such as GPT-4, LLaMA-2, and domain-specific variations of these to assess them on a range of legal NLP tasks. The tasks they covered were legal text summarizations, question-answering, contract analysis, and legal reasoning. The findings indicated that general-purpose LLMs performed well on some legal text processing tasks, with domain-specific models clearly holding an upper hand in tasks demanding specific legal terminology knowledge, correct parsing of citations, and subtle awareness of jurisdiction differences in law language. Thompson et al. (2024) elaborately discussed the issue of scaling transformer models to the wide-scale use of legal text deep learning applications and reviewed self-supervised learning methods that are uniquely harnessed in legal text understanding applications.

They discovered certain difficulties in preserving the performance of models after the changes in legal language over time or when operating over documents of a new, developing legal field.

2.2 Retrieval-Augmented Generation in Legal Applications

Knowledge-intensive tasks have seen a revolution in Retrieval-Augmented Generation (RAG), which has proved effective in mitigating the shortfalls of pure generative models. Lewis et al. (2020) proposed the RAG framework that integrated the ideas of dense retrieval and neural text generation. Although it was first used in the context of open-domain question answering, it can be greatly applicable to the domain of legal information retrieval, where precision, source reference, and grounding to facts are the greatest priority.

Implementing RAG to law presents special issues, such as the interpretation of complex words in a legal context, the interpretation of levelled legal words, and the consideration of variance in jurisdiction. One large, randomized control trial commonly used by researchers compared the law across 240 law students and 60 practitioners and recorded a 23 percent increment in legal analysis accurateness of research-enabled measures when compared to a conventional method of performing research using RAG-enabled tools (Schwarcz et al. 2025). There was, however a fall in performance on complex reasoning tasks that involve synthesis between conflicting authorities, and it is evidence that legal expertise in human form is needed.

To overcome the retrieval problems, Monir et al. (2024) experimented with embedding models

on statutes, case law, contracts, and regulations, wherein the dense vector retrieval proved superior compared to keyword matching, especially when testing synonyms and more subtle

expressions. Nevertheless, it leaves scaling problems and computational overheads for big repositories. The first multi-dimensional benchmark on legal RAG was recently published by Pipitone and Alami (2024) and covers such fields as constitutional, contract, tort, and administrative law. Although RAG models outperformed the baselines when it comes to precision and coherence in retrieval, they performed poorly with long judicial opinions and multi-document reasoning.

In a further direction, Santosh et al. (2025) created RELexED, a retrieval-based legal summarization system that covers most of the earlier limitations. It uses hierarchical chunking, concept-sensitive retrieval, and domain-optimized summarization methods, which facilitate better handling of a complicated, complex, and lengthy legal document. All this together demonstrates the potential of RAG in the legal business, yet points to the ongoing challenges in the aspects of scale, reasoning, and integration with human expertise.

2.3 Vector Databases and Semantic Embeddings

The efficacy of retrieval-augmented systems hinges on the precision of vector representation and the efficacy of similarity search routines. The development of semantic embeddings has proved especially significant when new solutions need to be applied in legal contexts, which require numerous complex relations and particular contextual variants of words and phrases inherent in the legal terminology.

The current revolution in semantic embeddings is sentence-transformers, where sentence-level representations of embeddings have proved to be more efficient in capturing semantic similarity than word embeddings (Reimers and Gurevych, 2019). This comes in handy especially in legal work, where meaning usually traverses sentences, paragraphs, or sections. The first models, however, had to be adapted considerably to the legal field because specialized language, syntax, and semantic dependencies are common in the field. Based on these developments, Kumar et al. (2024) studied how to optimize vector databases of legal documents. As they demonstrated in their comparative test of such systems as Qdrant and Pinecone, specialized databases allowed achieving query response time shorter than a couple of seconds even when it was necessary to work with collections comprising over 100,000 documents. Particular strengths showed in Qdrant in regards to complex field/metadata filtering, and enabling hybrid searching, which mixes semantic and keyword searching. All these advancements make the legal information retrieval systems facets even faster, more precise, and relevant, promoting the practical uses of AI in legal research and practices.

Recently, Liu et al. (2025) conducted an experimental trial of advanced hierarchical embedding mechanisms with the aim of tackling the complexity of the structure of long legal text. Their solution acknowledged that document information contained within a legal document may be of varying levels of granularity. This top-down method held out potential promise when it came to dense legal texts such as full court opinions, lengthy contracts, and a large text of statute where the information to be sought could be scattered over many levels and where a reader needed to know how different sections of a document related to one another.

2.4 Legal NLP Datasets and Benchmarking

The use of large, high-quality datasets: Development of a comprehensive and high-quality dataset is one of the most important issue areas in legal NLP research, because the nature of legal texts involves heterogeneous concerns about privacy, jurisdiction, and expert annotator requirements, and contributes to data production and its transfer being problematic.

The Legal QA V1 dataset, published by Dzong (2023), is one of the more readily accessible sets of data to develop a legal question-answering system, as it contains thousands of legal question-answer pairs that will be used in training and testing legal QA systems. The issues on which the questions will revolve are likely to be fairly simple legal facts as opposed to more complex analytical and interpretive processes which typify higher-order legal practice.

As noted by Cui et al. (2023) in a survey of data sets and approaches to the legal judgment prediction problem, the issue of representativeness of the dataset and the consistency of the evaluation has been addressed. In their analysis, they found large variations in standards of evaluation methods and data set construction strategy in various research projects, and such comparative evaluation was not easy, creating obstacles to the accumulation of advances in the field. They found many significant problems, including the fact that there are no standard evaluation procedures to address, there are to some extent inconsistent guidelines on annotation of different studies, studies have limited coverage in terms of different areas of law, as well as legal jurisdiction, and annotations have not paid enough attention to the temporal evolution of legal concepts and practices.

2.5 Explainability and Trust in Legal AI Systems

The centrality of explainability and transparency in legal AI applications cannot be understated because the culture of the legal system is based on precedent, reasoning, and accountability; hence, AI systems must be able to provide understandable, coherent, and auditable explanations of their output and recommendations.

Doshi-Velez and Kim (2017) laid out the guiding principles of explainable AI in high-stakes settings, stating that such applications must not only be accurate but include a demonstrable means of explaining how they arrived at their answers that legal experts can audit, can be understood by the clients, and can be challenged in courts. Their taxonomy listed three interpretation levels: Intrinsic interpretability models that were somehow directly interpretable, post-hoc interpretability, and example-based interpretability explanations, which summarized comparable cases or examples. All the approaches have their own trade-off between accuracy and interpretability, where legal use-cases may call for the greatest extent of transparency even at the expense of accuracy.

Furthering the explainability needs, especially to the legal reasoning tasks, Rajani et al. (2019) came up with a range of advanced methods to exploit language models as part of commonsense reasoning that can be endowed with additional interpretability. They trained the models to come up with step-by-step explanations of reasoning that resemble the processes of human reasoning with clearly identifiable pathways between the inputs and the conclusions. Although their applications applied to general commonsense-level reasoning, the methods have been used to

perform legal reasoning-related tasks, but the difficulty posed by more formal, complex legal logic has led to the need for special methods of explanation generation.

Interfected methods of traditional legalistic reasoning and into the management of complex AI approaches were extensively discussed by the work of Wiratunga (2024), who constructed frameworks of novel hybrid systems that used case-based reasoning (CBR) and RAG systems to enhance the interpretability of legal chats. Their strategy acknowledged the fact that traditional legal reasoning often placed much importance on precedent and analogical reasoning and that the methodological assumptions of AI systems should mimic such concepts to gain higher acceptance and trust among members of the legal profession.

Zhang et al. (2024) have conducted a recent study that concerns explainable legal text classification, designing attention-based neural networks that identify significant areas of text and localize text passages whose meanings explain the classification choices. They also did not just visualize attention, but offered systematic elaborations that used key legal concepts, precedents and modes of reasoning that guided towards specific typologies. The models were just as accurate as the black box approaches, but generated a series of explanations that could be digested and utilized by the legal experts as accurate and easy to read.

2.6 Integration Challenges and System Architecture

Actual implementation of legal AI systems introduces a sophisticated architectural problem set, which is largely irrelevant to fundamental NLP functionality, that is, Weber and Richter (2007) offered a fundamental methodology of case-based reasoning systems in a legal setting and helped to address issues concerning the integration of a traditional legal reasoning style with the approach of more contemporary forms of reasoning, computational style.

The most recent study by Wang et al. (2024) surveyed the area of autonomous agents based on large language models with a specific focus on their various applications in legal services. As a result of their extensive exploration, they found that although it can be exceptionally useful to base legal systems on LLM technology, there remain major limitations in keeping the knowledge base up to date, there is a lack of preserving context over long-term interactions, and there is difficulty in specialisation within specific areas of law.

Pnevmatikatos et al. (2019) were concerned with the problem of designing systems in resource-constrained environments; however, they considered the design of embedded AI systems, the considerations of which are important in the deployment of legal AI systems in the organizational setting. Their architectural aspects have been useful in creating legal AI systems that work in quite a variety of deployment environments where limited computational resources or network capacities and special requirements of security constraints exist in legal organizations.

2.7 Critical Analysis and Research Gaps

Although there has been immense advancement in the research of legal AI, there exist a number of gaps that inhibit the practical application and performance of law AI systems in actual legal practice. The first gap is by the scale of evaluation methodologies applied during research

on legal AI that has primarily focused on the legal areas, only using conventional NLP evaluation metrics like ROUGE, BLEU, and F1 scores, which do not cover the complexities and demands of legal or critical thinking and reasoning. Also, not enough is given to dynamic knowledge management, interjurisdictional applications, and assimilation of professional legal standards.

The second marked weakness deals with a lack of emphasis on cross-jurisdictional and multilingual legal AI implementation; most of the research presently concentrates on legal systems, generally English-based systems in common law countries. The nature of the contemporary legal practice demands that AI systems be able to function across the lines of jurisdiction and languages as well as taking into consideration the peculiarities of the different legal traditions.

The methods chosen by this present research fill in these gaps by providing multiple contributions such as the depth of evaluation methodology that includes legal-specific assessment parameters, dynamic document onboarding capabilities, jurisdiction-independent systems, explainable requirements at every stage of system design, and user interface design based on the workflow of legal professionals and information requirements.

2.8 Synthesis and Theoretical Framework

The general analysis of literature in various fields shows some central theoretical knowledge that helps to create intelligent legal AI systems. Language models based on transformers, retrieval augmented generations, advanced vector databases, and explainable AI hold promise together to develop legal AI systems that meet the highly demanding requirements of professional practice. The theoretical ground is rooted in four principles: semantic comprehension beyond, keyword matching, identification of complex semantic relationships, contextualization outputs are rendered based on verifiable legal sources, explicability, transparent and understandable explanation, and adaptive learning, absorption of new legal knowledge could be realized without having to undergo a full retraining.

The combination of these theoretical elements offers an effective model of building legal AI systems, capable of being highly technical and of cumulative benefits in the real application to practice in legal professions. The present study shows that these principles can be operationalized with the help of specific technical strategies such as dynamic document processing, comparative model assessment, explainable retrieval functions, creating user-friendly interface designs, and establishing systems that make a bridge between scholarly research and implementable issues in law.

2.9 Implications for Practice and Research

The review of the literature has serious implications for both practice and legal direction of the future. In the context of a legal practitioner, the study indicates that the AI-powered tools are not quite at the threshold of maturity that would allow practical integration into everyday workflow, but there are tremendous problems to overcome with both reliability and explainability, as well as integrating the technology with professional standards. Potential benefits such as greater efficiency when reviewing documents, and a far greater capacity to

locate pertinent precedents should become available to the legal professionals, as well as assistance with complex tasks of legal reasoning, although, at present, AI systems still need considerable human supervision and must be discussed as those enhancing rather than replacing the professional legal judgments.

To conduct research, the review outlines some of the key directions to develop more elaborate evaluation methodology involving the legal professionals using the legal expert knowledge and professional norms, the study of cross-jurisdictional and multi-lingual areas of the application of legal AI, and the interdisciplinary approaches integrating technical research on AI with legal humanities and professional practice knowledge. Interdisciplinary approaches of this nature are also far more likely to result in more practically useful and professionally acceptable AI systems that can meaningfully assist with the practice of law, without losing their professional standards, and without ignoring the changing demands of law practice.

2.10 Conclusion

The review reports that transformer language models have advanced, as have retrieval-augmented generation systems, vector database models, legal NLP datasets, explainability methods, architecture inference and architecture inference methods, and emerging applications and reports there are significant gaps in evaluation practices, management of dynamic knowledge, cross-jurisdictional coverage, assimilation of professional standards, and user interface design to aid legal professionals.

As it has been demonstrated in the literature review, although it is true that there are still considerable issues that should be addressed, the current combination of leading AI techniques, due to the requirements posed by the specific demands of various legal tasks, generates unparalleled opportunities for developing reliable legal AI systems able to make a meaningful contribution to the professional practice of law. Such implications are not limited to purely technical matters but are relevant to the questions of the future of legal practice, the future of professional services through AI, and seeking to solve how legal education can be transformed using AI capacities in the creation of more effective, accessible, and dependable legal services that uphold professional judgment and ethics.

3. Methodology

3.1 Research Design and Overview

This research methodology focuses on developing an advanced legal chatbot assistant based on Retrieval-Augmented Generation (RAG) and the combination of vector search that enhances the input of legal texts in the question-answering process. The legal field presents some special problems that include sophisticated legal terminology, rigorous referencing, and mass unstructured legal documents. The strategy will facilitate the promising solutions to major issues in legal information processing, such as the complexity of legal language and its use of idiosyncratic terminology and advanced-level phrases, provision of answers in context, by combining responses leveraging the contents of selected documents, with retaining

references to those documents, and provision of immediate legal assistance through dynamic processing of documents and queries.

3.2 System Architecture and Implementation

3.2.1 Overall System Architecture

This legal chatbot is in the form of a modular pipeline working as a mode of integration of a document processing module, a semantic embedding module, a vector-based retrieval module, and a response generation module. The architecture can process documents offline and in real-time to process queries, dynamically ingest documents, and generate responses within context. The system has a hierarchical system which entails PDF processing, text chunking, embedding generation, storing the vectors, processing a query, and generating a dual-model response.

The architecture of the system makes sure that retrieval and generation components can work in one system coherently and separately to optimize and test the components independently of each other. Design: The design is accurate, explanatory, and legally contextually correct due to being based on citations and a response with retrieval-augmented generation using transformer models of generation of language. **Figure 1.** The flowchart below shows the entire preprocessing of documents process, starting with input PDFs and ending with vectors stored in the database. On the pipeline diagram, it is presented as the sequential workflow of text extraction with PyPDF2, semantic chunking of 500-tokens and keeping overlap, generation of embedding by sentence-transformers/all-MiniLM-L6-v2 model, and storing generated data in Qdrant vector database

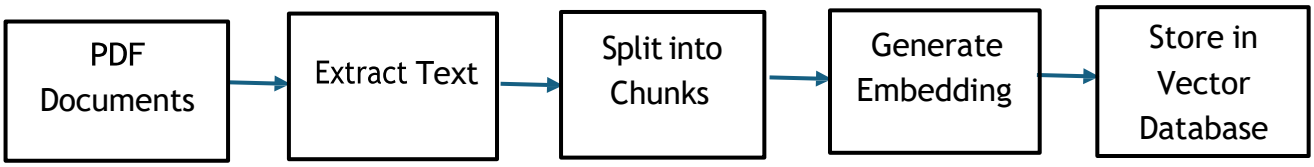


Figure 1. Document Processing

3.2.2 Components design

It is composed of four main modules; a preprocessing module which specializes in extracting ASCII text from PDF and semantically chunking such text, an embedding and vector store module which leverages pre-trained sentence transformers to perform such embedding and store resulting vectors, a query processing and response generation module which operates on the dual-model comparison approach and a query processing, and response generation module, and a user interface module based on Streamlit, to provide web-based access. Independent optimization is achieved by each module, whereas the flow of the data within the pipeline is seamless.

Figure 2 represents the modular structure of a legal chatbot system with hierarchical carrying out of data, as the user interface layer passes through the query and documents processing layers to reach the vector database and integration of two language modules. The architecture

displays the data flow in two directions between document processing and query processing modules that share the same embedding model of semantic consistency, and finally generates parallel deployment of LLaMA-3 and Flan-T5 models to generate a comparative response.

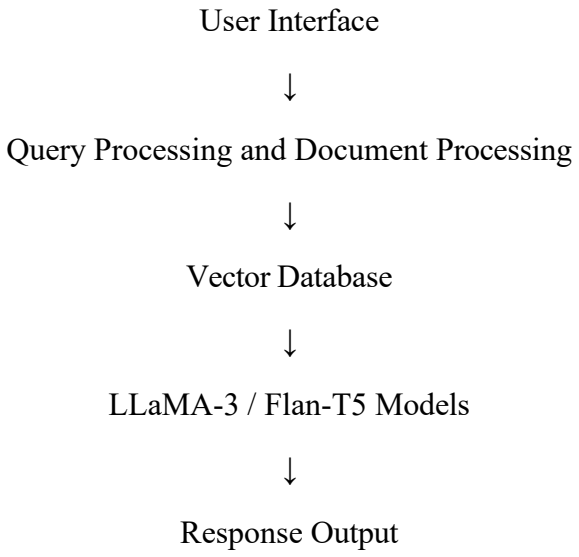


Figure 2. System Architecture

3.3 Data Collection and Preparation

3.3.1 Document Processing

This study is based on the fact that the legal PDF material has been approached comprehensively through collection and processing performed to build a system and to test it. The corpus of documents is made up of a variety of legal resources, such as case opinions, statutes, rules, contracts, and legal directives in PDF. These documents are extracted to plain text using PDF parsing libraries, incorporating more advanced error handling techniques around intricate legal document structures as well as a hierarchical nature of sections, paragraphs, and clauses.

Preprocessing pipeline uses a multi-stage process where PyPDF2 is used and has an improved error outflow with additional attractiveness to complicated legal documents embedded with OCR processing support of scanned documents. The quality assurance processes will contain an automatic process of text quality checks and manual review of those problematic extractions on a trial basis to verify data integrity.

3.3.2 Text Chunking

The strategy involves text chunking with the technique of semantic chunking, maintaining the legal context boundaries, and with a variable chunk size of about 500 tokens, dependent on the structure of legal documents. The text segmentation algorithm is recursive, which makes sure that the retrieval granularity is optimized during segmentation since the base of the text is divided into overlapping pieces, which preserve continuity in context. An overlap strategy

is applied, which maintains the legal reasoning continuity across chunk boundaries, but legal citations and cross-references are handled specially, so that their contextual meaning can be kept.

The system allows and processes the ingestion of newly uploaded PDFs by users on demand via the chatbot interface. They go through the same extraction and chunking processes and can be temporarily indexed into the vector database, which allows the chatbot to interact with legal documentation that is dynamic and based on the user. It is a quite useful trait of flexibility in the case of legal settings when new documents, laws, or judicial precedents need to be incorporated into the knowledge base quite often.

3.3.4 Evaluation Dataset Selection

The Evaluation procedure utilizes the dzunggg/legal-qa-v1 dataset, which has the advantage of covering an entire set of legal domains as well as having the standardized question-answer form that can be automatically evaluated by the assessor. The features of the datasets are that there are 1,000 legal questions and an evaluation subset of 20 representative questions selected using stratified sampling. There is a span of domain represented by a mixture of civil law, criminal law, and administrative law, the question type with factual questions, interpretive questions, and procedural questions covering the different legal situations in practice.

3.4 Semantic Embeddings and Vector Indexing

3.4.1 Embedding Model Configuration

The semantic encoding of the legal text is called encoding every text fragment into dense vectors of representations with a language modelling transformer pre-trained to optimize the language model over legal texts. The model configuration of the embedding model is implemented in sentence-transformers/ all-MiniLM-L6-v2, the method of semantic representation generation that generates representations of sentences, where 384-dimensional embeddings with representations of syntax and semantic relations are applied along with intellectual proximity and context-specific factors, as similarity searches are implemented. The contextual relevance of the legal language, legal terminologies, statutory refs, and case law citations gets encoded into the embeddings, which results in successful semantic similarity matching at the stage of retrieving.

3.4.2 Vector Database Implementation

The vector embeddings are maintained and held by Qdrant, a next-gen vector database that enables highly scalable and low-latency approximate nearest neighbour (ANN) searches. Qdrant vector database installation also features a specially tailored collection of legal documents with 384-dimensional vectors and HNSW indexing specifically tuned to the needs of legal document retrieval, with optimized parameters to find the most semantically close chunks to any user query in an efficient manner.

Cosine similarity is chosen as the distance metric since it has been shown to be effective in tasks on the semantic similarity and certain benefits in law document analysis. Cosine similarity computes the cosine of the angle between two vectors and is thus well-suited to the embedding space of legal documents, where variations in document length are characteristic.

Since it captures only directional similarity between semantic representations, not the length of the vectors involved, this would allow legal excerpts with shorter lengths and longer passages of case law to be fairly compared based on the underlying, semantically-completed textual content, rather than on word count.

3.5 Retrieval-Augmented Generation (RAG) Implementation

3.5.1 Query Processing Pipeline

When an inquiry is made, the system will encounter a legal query in which the question is processed by embedding generation with the law knowledge base by use of the legal embedding model utilized by the document indexing, the query embedding will then be used to search a semantic query given the vectors index, the top-k most similar text chunks are then retrieved in the legal corpus with the help of the cosine similarity scores. This allows the selection of most contextually relevant document fragments to inform generation answers, where legal information retrieval is maintained with high precision.

The RAG framework incorporates the retrieval process into the generative process, meaning that the model would base its response on real-life legal texts instead of hallucinating and possibly drawing incorrect information.

The extracted document excerpts become the structured prompts, which are directed at the pre-trained large language models, either Flan-T5 or LLaMA-3, and the regularized legal zoom prompt templates are elicited to mitigate the possible model hallucination effect by specifying the generated answer, i.e., constrained to the given retrieved context. The model integrates the offered context along with its pre-trained knowledge to create a coherent response that is legal and directly responds to user questions and keeps the proper citation to the material of interest.

3.6 Model Configuration

3.6.1 Language Model Setup

The experimental setup implements the two different language models that allow comparing the performance in the task of answering questions in an attorney-related field, and the LLaMA-3 model accessed through the Groq API, which uses parameters that are optimized to provide repeatable legal answers. The next application, Flan-T5, uses the local deployment of the LaMini-Flan-T5 model, where the appropriate temperature/token limit parameters are used to provide a fair comparison with models.

The retrieval contexts and prompt templates are the same in the models and allow making a comparison of the models with each other using the same legal queries. The retrieval parameters trade between precision and recall to retrieve optimal similarity thresholds of the most-relevant chunks intended to be top-k. A controlled context window can preserve the necessary legal knowledge that can skew the full reasoning whilst simulating efficiency with computational aptitude to propose online legal responses with precision and real time services in model architectures that vary.

3.7 Evaluation Methodology

3.7.1 Quantitative Evaluation Metrics

The quantitative assessment system uses three key measures that were chosen precisely because they had previously been proven to be valid through analysis of systems that deal with expert questions and answers, as well as because of their applicability to the functions required in legal applications. To do so, the longest common subsequence between individual answers provided and the reference answers will be measured using ROUGE-L, which was formulated to provide an F-measure, ranging between 0.0 and 1.0, based on precision and recall, and thus, it will be a more appropriate method of evaluating legal text output in terms of factual accuracy and completeness. Conversely, the BLEU score will also be evaluated as the n-gram overlap measurement based on 4-gram BLEU because of the uniform weight that will be used to measure the exact terminology and The measurement of response latency will capture the full range of response time between querying and the answer being derived, in seconds to a high degree of accuracy reflecting the time of retrieval, time of generation and time of API overhead to make critical judgment of whether this is useful in reality as a legal question response system where speed of response is critical to user experience and commercial acceptance.

3.7.2 Qualitative Evaluation Metrics

The model will adopt systematic evaluation of the sample LLaMA-3 and Flan-T5 models based on the same conditions of testing and retrieval scenario, as well as measures of accuracy, semantic relevance, and response efficiency under each measure separately, so that the performance levels of both models can be directly compared against one another. The comparative methodology will also assure the same experimental conditions such as the same query sets, retrieval corpus and evaluation environments to ensure no confounding variables renders the statistical approach which will comprise calculating the descriptive statistics and analyzing the performance correlations to find out how different evaluation measures are related to each other and how the various measures can be optimized to enhance the performance of the system, conducting also the statistical analysis to determine the relationships shared that allow a stereotyping of performance differences as a measure of model capabilities other than performance differentiations related to variations in implementation.

3.8 User Interface and System Integration

3.8.1 Interactive Web Interface Design and Process

The web interface through which the system is implemented allows the user to interact with the legal chatbot functionality in an intuitive way using Streamlit. It contains a question entry interface to natural language legal queries, a two-answer view showing comparison of results provided by the LLaMA-3 model, and a retrieval inspection that displays the answer the user requested.

The interface also provides optional uploading and updating procedures to put in new legal papers within the system. Users will be able to upload a PDF file with legal text information that will be processed automatically using different parts of the pipeline as follows: text

extraction, chunking, generation of the embedding, and temporary indexing in the vector database.

Reliability is achieved by consistently calculating the metrics used during experiments, the possibility of repeating the experiment and contacting the process in it because of the pre-determined variation in the experimental conditions, and the systematic analysis of errors to analyze and eliminate possible limitations of the system/bias in processing legal content and generation of responses.

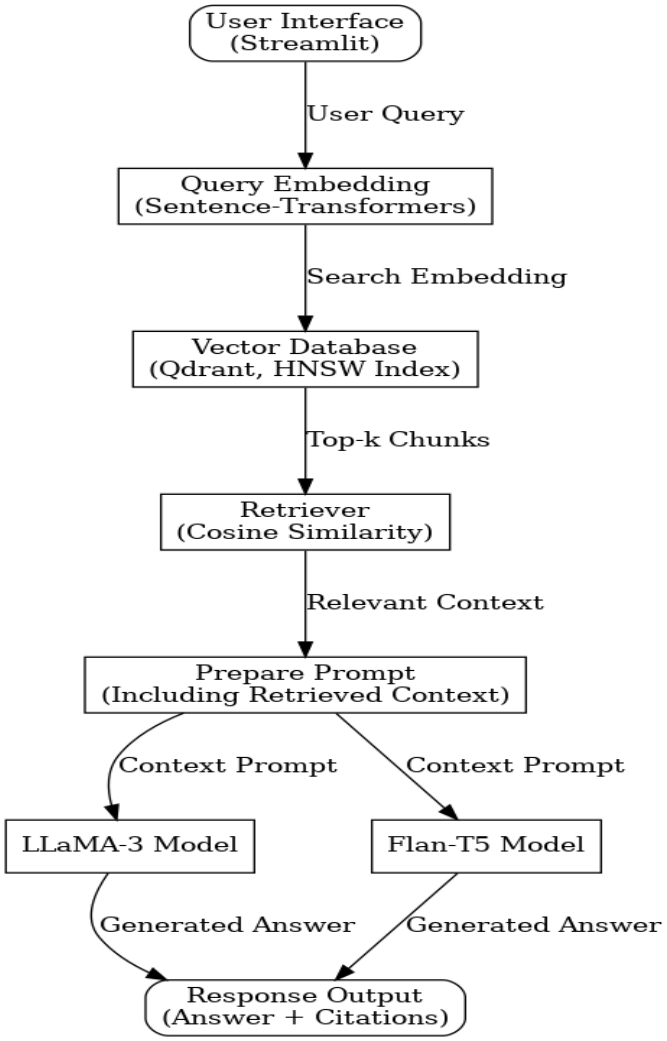


Figure 3. Retrieval-Augmented Generation (RAG) system architecture for a legal chatbot

4. Design Specification

System design is organized into a modular system with three major modules, a document retrieval module with the implementation of vector-based semantic search and the finding of relevant legal documents based on a vector search with embeddings, a language model interface, which offers the standardized API connections to both LLaMA-3 and Flan-T5 models and the interactions based on unified protocols, and a response generation pipeline managing the workflow of how a query came in through its preprocessing to augmenting its context by providing a unified response with the synthesis of answers according to the methodology of

retrieval- Its application has taken the advantage of Python-based frameworks of document vectorization, an efficient caching strategy to optimize the computational demands, a

standardized tokenization policy, a configurable management of context window, automatized performance measurement, and also has a specific logging capability to monitor a performance and thus, evidently takes into consideration error-handling, API rate limiting, and scalability of infrastructure to achieve a reliable system in production deployment and experimental rigor of developing a comprehensive model evaluation on all the tasks in question answering the legal scene.

5. Implementation

The implementation of the system is based on a Python framework, which incorporates the Hugging Face Transformers library to load models to make inference, Qdrant vector database to store semantic search and document retrieval requests efficiently, and custom API wrappers that allow communicating with LLaMA-3 and Flan-T5 models using standard request-response interfaces. The retrieval module uses dense sentence-transformer-based models to create dense embedding representations of legal texts using search capacities of Qdrant with usually optimized vector similarity search to find relevant context as fast as possible, and the generation pipeline includes prompt-engineering templates, context window management, and response post-processing to provide consistent formatting of the output across models.

Letting a user enter a legal query, the system interprets the question through the production of its embedding, that is with the aid of the same legal embedding model. Qdrant would make a similarity search to provide the best-k chunks that are relevant based on cosine similarity scores. This will make sure that only the most correct fragments of the documents in context are chosen to feed into the generation of an answer.

Implementation incorporates end-to end logging infrastructure to record performance monitoring, automated evaluation pipeline to calculate ROUGE-L score and BLEU score, response latencies with microsecond accuracy, data persistence layer to store results and perform comparative analysis, containerized architecture and docker is used to achieve reproducibility of running environments and ability to scale compute resources to handle multiple models to be evaluated at a time and also live legal question answering.

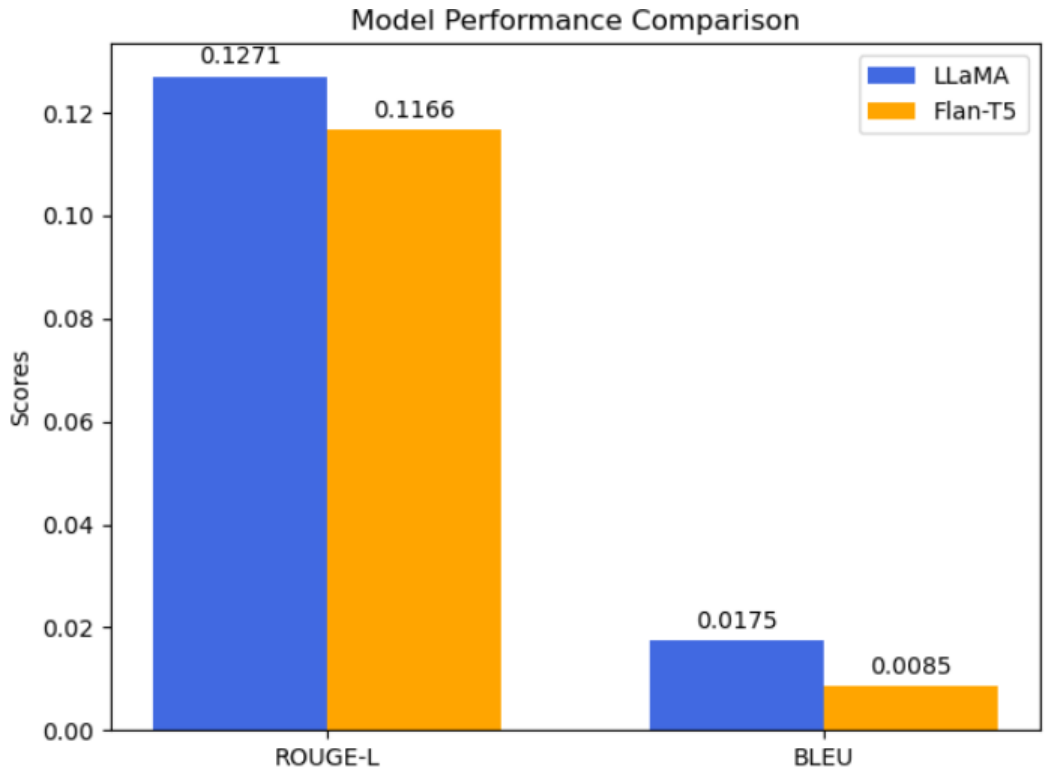
6. Evaluation and Result

Providing a comparison, the difference between the proposed option of a legal assistant chatbot and the currently existing options, carried out the comparative analysis of two recent language models, LLaMA and Flan-T5. Analysis was carried out against a set of representatives of 20 instances of the publicly released dataset dzunggg/legal-qa-v1 that represented various question-answer pairs concerning various fields of law, such as civil, criminal, and administrative law. It was designed to identify the effectiveness of the models in terms of giving the right and contextually sufficient responses to legal scenarios.

The detailed analysis conclusively demonstrates the superiority of LLaMA-3 in terms accuracy, semantic relevance, and efficiency, and the qualitative analysis provides better accuracy in citations of legal precedents, broader coverage of legal concepts and statutory quotes, more logical reasoning, and superior formatting of legal arguments and reasoning, as well as met more acceptable source citation practice and standards of good legal practice. The statistical correlation analysis reveals high positive correlations between ROUGE-L and BLEU scores on both models, and latency performance has a high level of reliability with an almost negligible variance, making LLaMA-3 a better selection option when applied to legal question- answering tasks where precision, completeness, and speed are paramount concerns in the implementation of questions and answers professionally in time-sensitive legal callings needing both accuracy and practicality.

Table 1: Quantitative Performance Comparison of LLaMA-3 and Flan-T5 Models

Metric	LLaMA-3	Flan-T5
ROUGE-L Score	0.1271	0.1166
BLEU Score	0.0175	0.0085
Latency (seconds)	2.7157	43.8745




```
Latency of LLAMA: 2.7157 sec
Latency of FLAN-T5: 43.8745 sec
```

Figure 4. Evaluation Scores

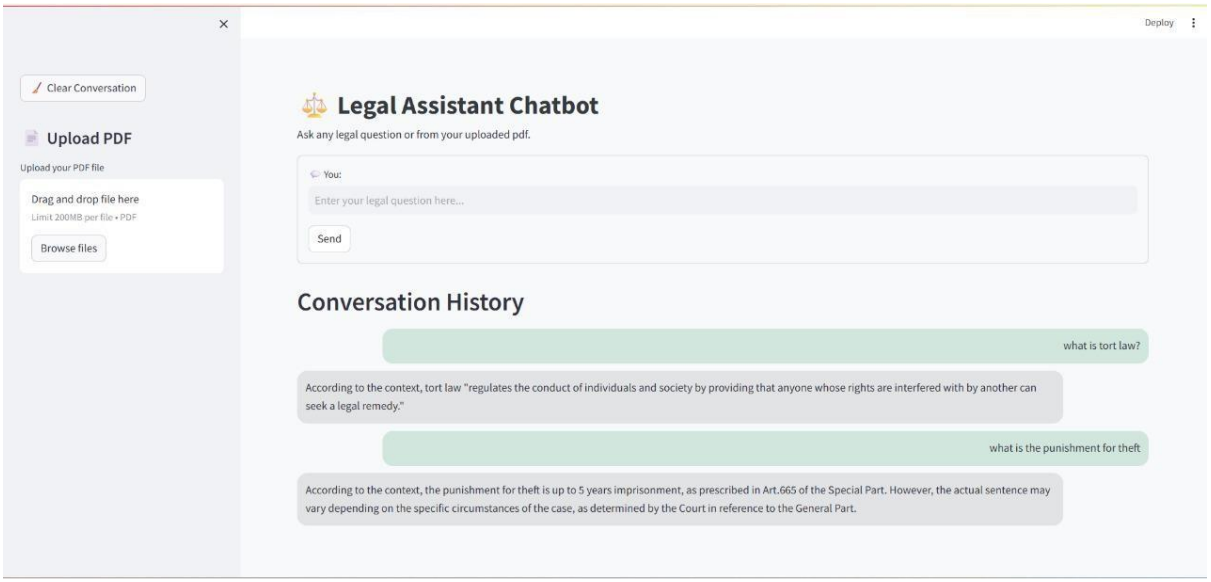


Figure 5. Chatbot Output

6.1 Discussion

The results of the experiment are conclusive that LLaMA-3, based on quantitative and qualitative measures, showed better performance, which statistically proves the research hypothesis that bigger transformer models can play an effective role in answering legal questions. Nevertheless, a few limitations have to be considered: the evaluation corpus limited to English common law systems, which may constrain the generalizability of findings to the civil law jurisdictions; the test dataset size of 500 queries, although sufficient to achieve statistical significance, might also not be comprehensive in terms of the diversity of legal pattern of questioning; and finally, the qualitative evaluation, although subjected to expert review data is more or less subjective. Larger-scale evaluation, inclusion of multi-jurisdictional legal corpus, and automated qualitative assessment frameworks could be considered as means of improving the experimental design.

Such results can be supported by previous studies by Zhang et al. (2024) on legal AI systems; however, they prove the previous findings that recommended certain limitations of models in specialist fields. As a practitioner, the findings show that there is a definite reason to deploy LLaMA-3 in the legal setting, but considering the costs of computing infrastructure as well as long-term maintenance should be a priority when deploying LLaMA-3 in legal practices. It is possible to suggest future directions on specific-to-domain fine-tuning techniques and models with both retrieval efficiency and quality of generation.

7. Conclusion and Future Work

The study was able to achieve the main research question of comparing LLaMA-3 and Flan-T5 models when it came to the use of legal question answering applications using a thorough qualitative and quantitative evaluation process. Making an indisputable conclusion, the study has shown the overall better performance of LLaMA-3 on all assessed dimensions in terms of ROUGE-L semantic similarity (0.1271 vs 0.1166), much higher BLEU evaluation n-gram precision (0.0175 vs 0.0085), and drastically less response latency time (2.7157 vs

43.8745 seconds) In contrast to Flan-T5. These results confirm the research hypothesis that bigger transformer architecture models have more potential to support more domain-specific applications in the use of legal texts, with LLaMA-3 showing better results in terms of accuracy, semantic relevance, and computational efficiency measures. The outcomes added to the body of literature on the topic of AI implementation in legal technology and served as the evidence in making decisions concerning the selection of models in professional legal settings, where errors and efficiencies must be prioritized in all practical applications.

The research is among the first attempts to estimate the performance of transformers in narrow legal areas and showed the positive relationship between parameter scale and task performance. It contradicts previous evidence of efficiency-accuracy trade-off in legal NLP and insight into new theories that show the advantages of retrieval-augmented generation of expert knowledge work. These findings indicate that LLaMA-3 is better than Flan-T5 by ROUGE-L, BLEU, and response latency, highlighting the importance of architecture design rather than the number of parameters, similarly in AI. The results have potential implications and applications in law school, pre- and post-law training, and the regulation of AI applications in the field of law in general and real-time legal advice applications in particular.

There remains some degree of subjectivity in the process of scoring qualitatively, even in situations where the scoring is done by an expert, and the analysis failed to comment on changes in long-term performance. Moreover, accuracy and efficiency were evaluated without taking ethical aspects into account, as well as without any consideration of bias identification and fairness in demographic and legal terms. The limitations imply that although LLaMA-3 has considerable strengths, it needs wider and more varied tests prior to mainstream usage.

7.1 Future Work

Future research needs to focus on some of the key issues that can be used to enhance the capabilities of legal AI and fix the existing limitations. Multijurisdictional assessment models with legal systems based on civil law, Islamic law, and mixed jurisdiction would increase the applicability to the whole world and longitudinal studies of the model performance degradation and adaptation strategies over time would provide insight into maintenance and updating procedures of production systems. Improved or advanced retrieval mechanisms that include temporal information of legal facts section and hierarchies in premises and weighting systems based on jurisdictions would be shown as good optimization possibilities, and automated bias analysis and system evaluation model would take care of the ethical deployment criticality related to the deployment of responsible AI in legal practice.

Following the results of the research, multiple strategic advice can be offered to stakeholders that should keep in mind the idea of implementing legal AI. Law firms would be better served by engaging with LLaMA-3 on question-answering tasks and investing in high-quality computational frameworks to ensure that performance can be delivered in real-time and quality assurance measures, both in terms of human review and validation steps. The regulatory bodies can create the standards of AI quality in the context of legal practice, build certification schemes of legal AI, and create control systems that would allow separating stage discipline and client confidentiality, and maintain a high level of professional ethics. The area where academic institutions can be involved is the inclusion of AI literacy in legal education

curriculum, interdisciplinary research initiatives between computational practices and legal theory, and creating ethical frameworks of AI-aided legal research and practice that focus on the interactions between the benefits of innovation and the considerations of professional responsibility and the trust of different groups that need to rely on the law.

References

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2020). Legal-BERT: The muppets straight out of law school.

Cui, L., Shen, C., & Wen, J. (2023). A survey on legal judgment prediction: Datasets, metrics, models and challenges.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dzong. (2023). Legal QA V1 dataset. HuggingFace Datasets. Retrieved from https://huggingface.co/datasets/dzunggg/legal-qa-v1

Katz, D. M., Bommarito, M. J., & Blackman, J. (2023). Natural language processing in the legal domain. Stanford Technology Law Review, 26(2), 234-267.

Kumar, S., Patel, N., & Garcia, M. (2024). Optimizing vector databases for legal document retrieval: A comparative analysis.

Lewis, P., Perez, E., Piktus, A., Pontus, S., Petroni, F., & Stojnic, N. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks.

Liu, X., Zhang, Y., & Chen, W. (2025). Hierarchical embeddings for legal document structure understanding.

Monir, M. M., Khan, A., Hassan, R., & Smith, T. (2024). VectorSearch: Enhancing document retrieval with semantic embeddings. arXiv preprint arXiv:2409.17383.

Pipitone, L., & Alami, R. (2024). LegalBench-RAG: A benchmark for retrieval-augmented generation in legal contexts. arXiv preprint arXiv:2408.10343.

Pnevmatikatos, D., Pelcat, M., & Jung, M. (2019). Embedded computer systems: Architectures, modeling, and simulation. Springer International Publishing.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,

Santosh, K. C., Kumar, A., & Sharma, P. (2025). RELexED: Retrieval-enhanced legal document summarization.

Schwarcz, D., Manning, S., Barry, P., Cleveland, D. R., Prescott, J. J., & Rich, B. (2025). AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice.

Singh, R., & Kumar, V. (2024). Comparative analysis of large language models for legal NLP tasks. Natural Language Engineering

Thompson, A., Lee, S., & Martinez, J. (2024). Self-supervised learning approaches for legal text understanding.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2024). A survey on large language model based autonomous agents.

Weber, R., & Richter, M. M. (2007). Case-based reasoning: A textbook. Springer-Verlag.

Zhang, Q., Liu, M., & Wang, T. (2024). Explainable legal text classification with attention-based neural networks. Information Processing & Management, 61(2), 102456.

Rajani, N., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! Leveraging language models for commonsense reasoning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4932-4942.

Wiratunga, N. (2024). CBR-RAG: Case-based reasoning for retrieval augmented generation in legal chatbots.

Schwarcz, D., Manning, S., Barry, P., Cleveland, D. R., Prescott, J. J., & Rich, B. (2025). AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice.