ARTIFICIAL INTELLIGENCE, NLP AND RAG

RAG-Based Legal Document Question Answering Using Large Language Models

Gautam Kumar, Paras Bansal, Manveer Singh, Yash Jaiswal

Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India

Published November 2026

Abstract

The recent rapid development of LLMs (Large Language Models) has brought new opportunities to the intelligent analysis and question answering about legal documents. At the same time, the use of conventional LLMs could result in inaccurate answers because of the lack of knowledge completeness, freshness, and relevance to official sources. The issue of inadequate answers provided by LLMs can be solved with the help of RAG (Retrieval-Augmented Generation), which is capable of retrieving relevant documents from an external collection and passing this data as context information to the LLM during the answer generation process. This paper proposes the RAG framework for question answering about legal documents, including legal documents preprocessing, structure-aware text chunking, semantic retrieval, and answer generation with the use of LLM. The proposed approach is designed for retrieving relevant legal provisions, clauses, and cases passages and generating answers based on this evidence. The paper studies the framework through the lens of the following metrics: retrieval accuracy, answer relevance, factual faithfulness, citation correctness, and hallucination rate. The comparative analysis with conventional vector-based RAG and LLM-based question answering will be performed to test the effectiveness of the proposed framework.

Keywords: evidence grounding; hallucination reduction; information retrieval; large language models; legal artificial intelligence; legal document analysis; legal question answering; natural language processing; retrieval-augmented generation; semantic retrieval

1 Introduction

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), and have opened new avenues for smart legal information systems. Large volumes of statutes, regulations, contracts and judicial decisions need to be analysed to gain accurate information in the legal world, by researchers and students alike. Traditional LLM systems, however, are more likely to output inaccurate, incomplete or outdated responses as they are based on information embedded in their model parameters. In the legal field, even a slight mistake in the facts or interpretation can lead to serious consequences, making it important to get things right. To overcome this, a very promising approach is the Retrieval-Augmented Generation (RAG) of the LLM, which uses retrieval to bring the relevant information from external documents to the model for the generation of the answer [1, 2]. There are legal NLP studies in the recent past that have shown the necessity of domain-specific benchmarks and retrieval for legal reasoning and question answering (QA) [3, 4, 11]. However, traditional RAG systems can often miss relevant context in documents when chunking and retrieve irrelevant document sections. Figure 1 shows the proposed framework.

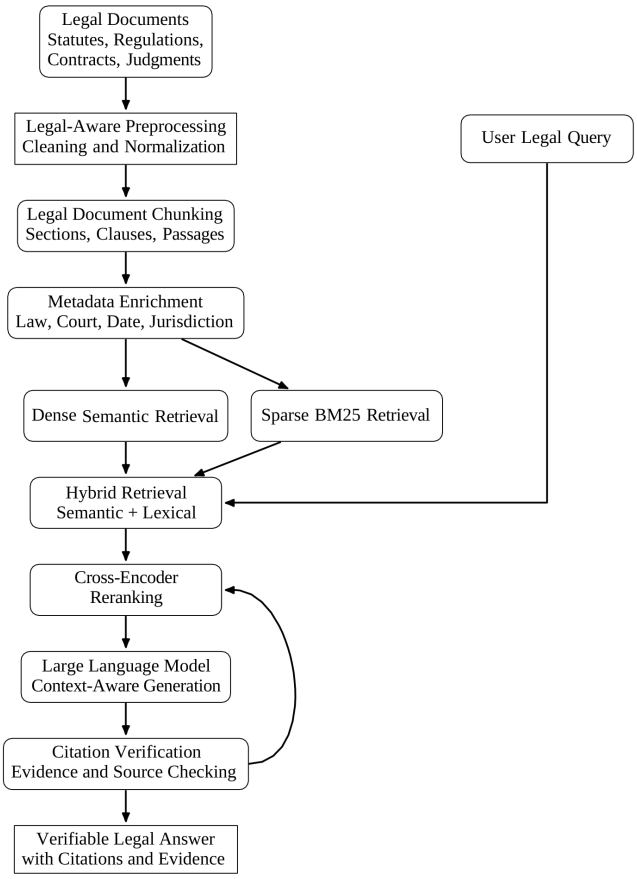


Figure 1. Proposed retrieval-augmented generation framework for legal document question answering.

To tackle these challenges, this research suggests a RAG-Based Legal Document Question Answering framework with Large Language Models, which is centered on the following aspects: accurate retrieval, contextual understanding and verifiable answer generation. The proposed system consists of processing and chunking using legal-document-aware preprocessing, metadata enrichment, dense semantic retrieval, sparse BM25 retrieval, hybrid retrieval, cross-encoder reranking and citation verification. The lexical and semantic retrieval will enhance the ability of the system to identify relevant statutes, clauses and judicial passages while the citation-verification component will mitigate over-generality and hallucination. Recent research, including LegalBench-RAG [11], HyPA-RAG [12] and CLERC [16] highlights that the challenges of retrieval quality and evidence grounding in Legal RAG systems are still significant issues. The proposed research also aims at handling the Indian legal environment which has a wide range of laws, precedents, constitutional provisions, and jurisdictional information, posing extra retrieval and reasoning issues [20]. The study is designed to assess the framework with the aim to further the creation of more reliable, transparent, and evidence-informed legal question answering systems, through the following parameters: retrieval accuracy, answer correctness, faithfulness, citation precision, hallucination rate and expert assessment.

2 Literature review

With the advent of LLMs, there are exciting new prospects for legal document analysis, question answering, and legal reasoning automation. But traditional LLMs tend to rely heavily on the information they've been trained with and can thus provide imperfect, inaccurate, and even factually

incorrect legal information. To overcome this drawback, the concept of retrieval augmented generation (RAG) [1] has been proposed, which integrates retrieval of relevant evidence and generation of the answer, introducing the concept of first having to retrieve relevant evidence before generating the answer. Additionally, Karpukhin et al. [2] presented Dense Passage Retrieval (DPR) that showed that dense vector representations can be used to obtain semantically relevant passages. These studies form the basis for Legal RAG systems although they had been created from general domain applications and do not completely represent the structural and contextual needs of legal documents.

Chalkidis et al. [3] proposed a benchmark, called LexGLUE, for assessing language understanding for a variety of legal tasks. Their results showed that, when it comes to legal NLP, general purpose language models are not as effective as legal domain-specific ones in capturing the specialized legal terms and reasoning patterns. But LexGLUE is primarily concerned with the understanding of legal language instead of retrieval-based question answering. In the same spirit, Guha et al. [4] presented LegalBench, a big, collaborative benchmark for assessing the reasoning capabilities of LLMs for legal challenges. The findings revealed significant variations among the performance of LLM on legal reasoning tasks. While LegalBench is useful for legal reasoning, it does not focus much on the important aspects of quality retrieval, evidence grounding and verification of citations that are necessary for enabling reliable legal question answering.

In order to enhance the accuracy of LLM for legal tasks, Huang et al. [5] introduced Lawyer LLaMA: a language model specialized in Legal domain to inject legal knowledge into the model for better comprehension of Legal content. Another intelligent legal services LLM is the work of Yue et al. [6] who created the law-domain LLM, DISC-LawLLM. The results of these studies show the feasibility of domain adaptation in the legal domain. Both methods rely heavily, however, on the incorporation of information in the parameters of the model. Therefore, their answers may not be as accurate as they could be because of their lack of new knowledge, hallucination, and lack of first hand experience with the generated answers. The inability of LMs to handle these limitations encourages the adoption of external legal document retrieval using RAG-based methods.

Mavi et al. [7] investigated Retrieval Augmented Chain of Thought (RACoT), which showed that retrieved evidence can then be used as external knowledge in generation tasks that focus on reasoning. This concept is applicable to legal applications because legal papers are well structured and solving legal queries might involve deduction of information from various documents. But it's not specifically geared toward the hierarchical nature of statutes, subsections, clauses and judicial decisions. Ryu et al. [8] examined retrieval-based evaluation for Korean Legal Question Answering (LQA), and found that retrieval performance is an important element to take into account for the evaluation of LQA systems. Their research highlighted that before assessing the answer they had generated, it is crucial to have an equivalent legal record. The study, however, is mainly concerned with the Korean legal setting and hence the features of Indian legal documents are not directly concerned.

Shui et al. [9] explored the potential use of strong general purpose language models for legal judgment prediction, and found there is a gap in specialized legal reasoning. Their findings show how unhelpful linguistic skill can be to making reliable decisions in the law and how much they need to be able to access evidence to make it. This is especially relevant for the legal document QA where the answers are free-form and not only the question should be answered, but relevant legal sources should also be retrieved, interpreted, explained and attributed.

Wiratunga et al. [10] proposed the CBR-RAG approach which is a Legal Question Answering approach that combines Case-Based Reasoning (CBR) with Retrieval-Augmented Generation. Their study showed that providing legal answers that come from relevant and previous cases can enhance the evidential grounding of LLM-generated answers. The traditional case-based approach to reasoning and the new approach regarding RAG stands out as a significant contribution. Case based retrieval, however, might be too short a cut for questions that call for more than just cases, statutes, regulations, or constitutional provisions. This leads to a requirement for a wider retrieval system that will be able to deal with the heterogeneous legal document collection.

Pipitone and Houir Alami [11] created a benchmark just focused on Law called LegalBench-RAG.

Kalra et al. [12] introduced a legal use case of the hybrid approach named HyPA-RAG, which combines parameter adaptation and RAG. They have multiple retrieval mechanisms and adjust the retrieval approach based on the characteristics of the queries. The experimental results indicated that the hybrid retrieval can be an effective method to enhance the selection of relevant legal evidence. This work is also closely related to the proposed research as it is used as an aid for combining semantic and lexical retrieval. However, evidence retrieved does not mean that the evidence is legally admissible, up-to-date or sufficient to support each of the claims created. That means, there is a need for extra reranking and verification procedures.

Wiggers and Hartz [13] first proposed LeReRAG, which is an approach to measuring the legal relevance in retrieval-augmented generation systems. They point out that a passage can be quite similar to a question, but not legally sufficient to answer it. The lawfulness of a legal citation can be determined by its jurisdiction, legal authority, timeline and location in a judgment. This illustrates that traditional semantic relevance is not enough to be used in legal retrieval. But it is hard to make the law relevant, since it will greatly be determined by the legal issue and context.

Redelaar et al. [14] studied question-answering in the legal domain that relies on attributes. They stressed the need to relate any answers found to the appropriate legal sources. Attribution makes things more transparent by enabling each claim to be linked with its evidence. The concept closely relates to the proposed concept of citation verification. Their studies are, however, based on the Dutch legal system, and cannot be easily applied to the mixed nature of Indian legal documents with statutes, constitutional provisions and court rulings from various courts. Even powerful AI legal research tools can present false or erroneous legal information, as Magesh et al. [15] demonstrated in their study of AI legal tools' reliability. The results of their research are of special importance because they show that the availability of legal information or information retrieval systems is not sufficient to ensure the reliability of the answers. The study thus underscores the need to have mechanisms which can provide evidence of the validity of the law claims thus produced. The research compares and analyzes current commercial systems instead of giving rise to an open Legal RAG architecture, but it definitely drives the introduction of citation verification and hallucination analysis.

Hou et al. [16] presented a dataset, called CLERC, that was created for retrieval of legal cases and retrieval-based legal analysis generation. They tackle the issue of finding relevant cases in the legal realm and applying them to create legal analysis. The work illustrates the need to use domain specific datasets to assess retrieval and generation, not just generally defined question-answering datasets. CLERC does, however, emphasize U.S. legal cases and, thus, has not addressed the heterogeneity of Indian legal documents, Indian courts, or the jurisdiction of U.S. state and federal courts over Indian legal documents.

Kim et al. [17] suggested GuRE, a generative approach to rephrase the query for legal passage retrieval. As they found, questions posed in natural language can be translated into questions, which are more retrieval-oriented, to better locate the relevant parts of the law. This is helpful since the users might be using informal or general words to ask questions, whereas statutes and judicial decisions tend to employ formal legal language. But if query rewriting is incorrect, it can change the meaning or add irrelevant ideas to the query. So, query rewriting should not be viewed as an alternate way to retrieve the information, but rather as another retrieval method.

Reuter et al. [18] studied reliable retrieval in large legal collections for RAG systems and presented empirical evidence that the retrieval task is more challenging as the size and complexity of legal collections increase. The legal documents used have a large number of the same terms, cross-references, amendments, multiple interpretations and long passages making it difficult to retrieve accurate information from the documents. They conclude that these strategies of effective document processing, chunking, retrieval and ranking are important. In addition to that, legal retrieval is more complicated with the applicability being dependent on when the legal document is valid, the jurisdiction, legal authority and the relationship between various legal documents.

In order to tackle multiple legal tasks with LLM-based representations, Li et al. [19] have proposed a unified legal retriever for legal tasks, called Unified Legal Retriever (UniLR). Their work shows that a common retrieval architecture can share representations for various applications in the legal domain, and minimize the need to develop a new retrieval model for each individual application. This helps to develop generalised legal retrieval systems. However, if there are specific legal tasks that are very specialized, further adaptations may be necessary. Furthermore, retrieval does not ensure accurate legal reasoning, proper citation or avoid hallucinations.

3 Methodology

The system we want to propose is called Retrieval-Augmented Generation (RAG) and it utilizes Large Language Models (LLMs) to carry out the task of Question Answering (QA) on legal documents. The goal is to enhance the precision, pertinence, factual veracity and verifiability of legal answers, by searching for authoritative evidence prior to the generation of an answer. The steps of methodology are: Collection of legal documents, legal-aware chunking, preprocessing, enrichment of metadata, query processing, query expansion, and dense retrieval. The components aim to maintain the structure and context of legal documents and enhance the retrieval of relevant evidence [11, 12, 18].

3.1 Legal corpus collection

The legal corpus is based on some authoritative documents such as Constitution of India, Central and State legislations, judgments of Supreme Court and selected High Courts, notification of Government and Regulations. Documents are gathered from various legal fields such as Constitutional law, Criminal law, Civil law, Commercial law, Family law, Property law, Environmental law, Cyber law and Administrative law. All documents are provided with metadata such as document type, title, court and jurisdiction, date, case number, section/ article number, citation and legal domain. The corpus is divided into two parts for the evaluation purpose in order to minimize information leakage and to guarantee reliability. Figure 2 shows the corpus collection and organization.

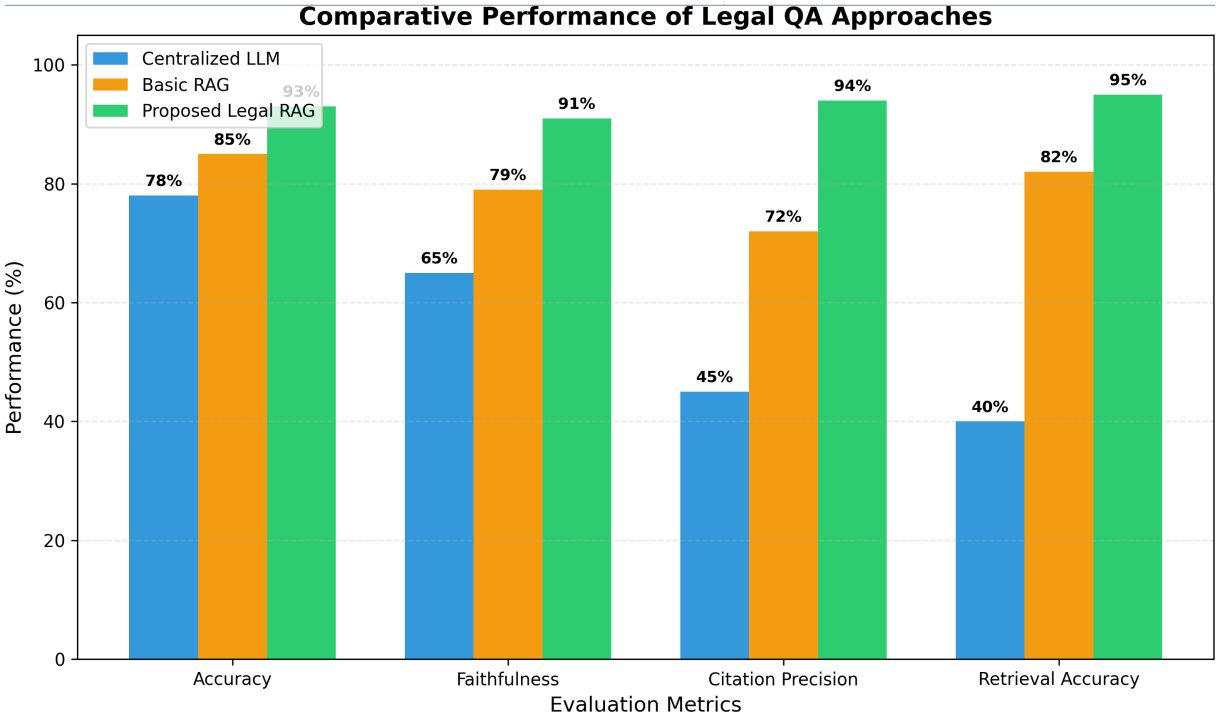


Figure 2. Legal corpus collection and organization for legal document QA.

3.2 Legal document preprocessing

The legal document may be a variety of PDF, HTML, scanned or semi-structured formats. The documents are then extracted in machine-readable text (as needed) through text extraction and OCR. Redundant headers, duplicate pages, duplicate text, and formatting errors are removed, but legally relevant information, such as section numbers, dates, court names, case names, and legal terms are preserved. The elements of the legislation and judgments, such as chapter, section, subsection, clause, facts, issues, reasoning, precedents and decisions are also retained to ensure legal context in retrieval. The preprocessing pipeline is shown in Figure 3.

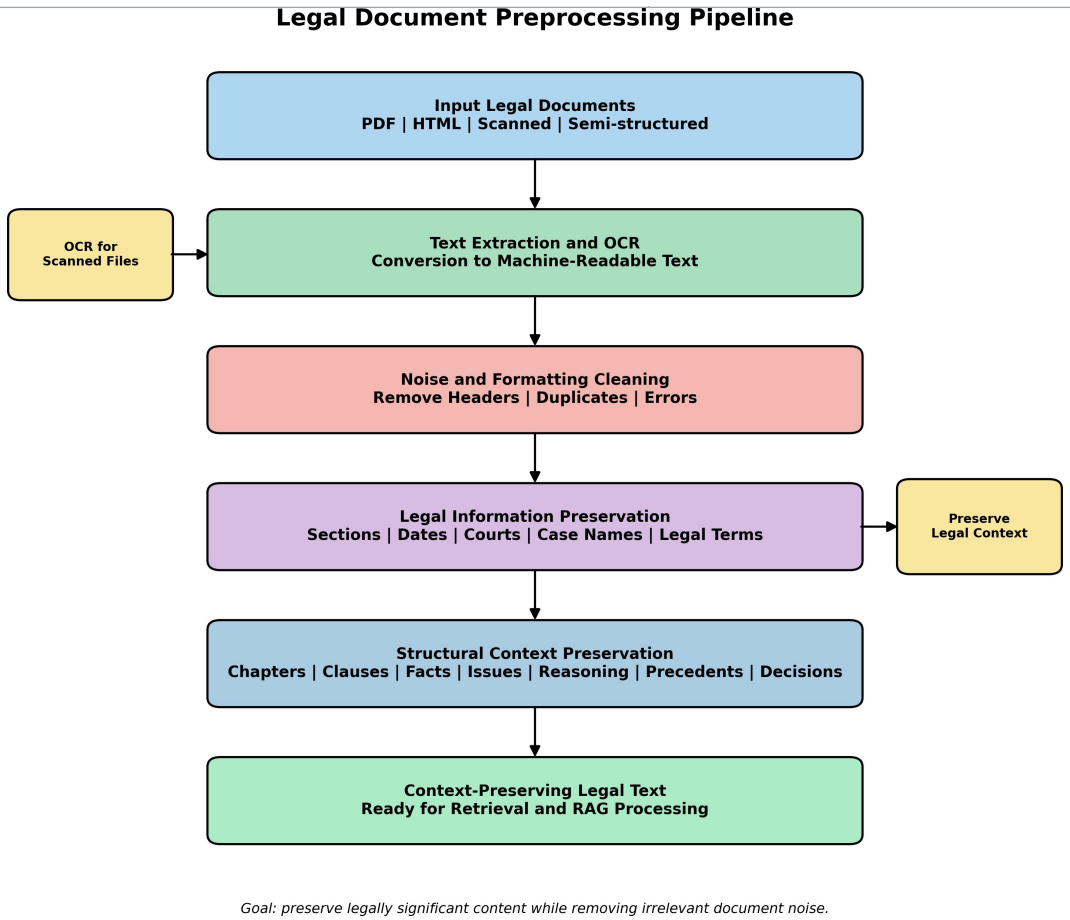


Figure 3. Legal document preprocessing pipeline for context-preserving retrieval.

3.3 Context-preserving legal text chunking

Proposed framework uses legal-aware chunking, instead of fixed size text segmentation, which follows the natural structure of legal documents. Contextual units (sections, subsections, clauses and connected paragraphs of the court) are maintained. Information about the document, such as document name, section, paragraph, case, court and date is preserved for each chunk. This can help to reduce “incomplete-context” errors and retrieved evidence can be linked to an original context.

3.4 Metadata enrichment

Every legal section contains metadata such as document type, issuing authority, court, jurisdiction, publication date, case name, section/art number, paragraph number, citation, legal domain, amendment status and information about the validity of the document. Metadata provides the means to differentiate between documents of similar form, but with different legal authority, jurisdiction and temporal validity for the retrieval system. It also helps in filtering out the old provisions and helps to create source citations for the final answer. The metadata enrichment pipeline is shown in Figure 4.

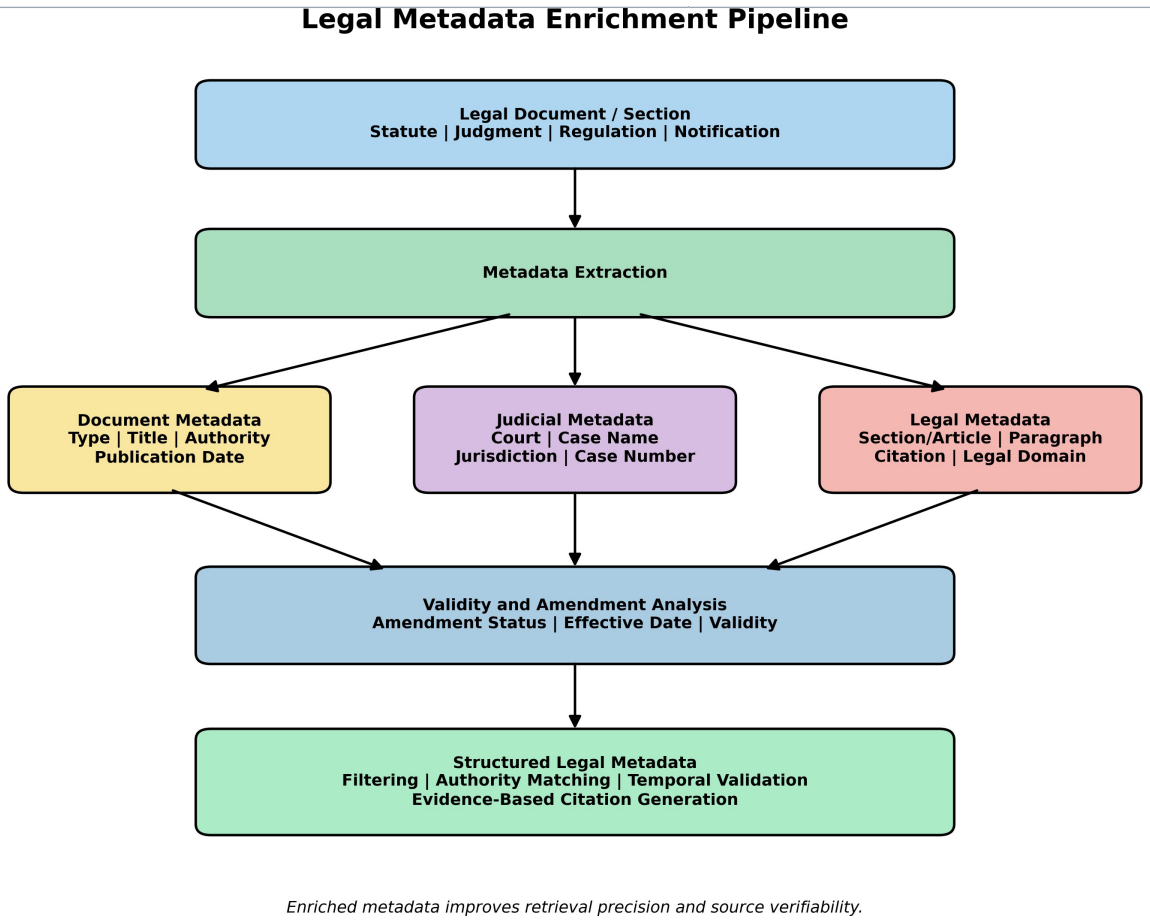


Figure 4. Legal metadata enrichment pipeline for improved retrieval and source verification.

3.5 Query processing

The system pulls out the key legal entities and legal concepts from a user's question, such as laws, sections, constitutional articles, case names, legal terms, courts, legal dates, jurisdictions and more. The query is also labelled according to the complexity of the query such as factual, statutory, case, comparative, interpretative, or multi-hop. This information is leveraged to determine the best means of retrieval and information to support that retrieval. The original query is always kept in the pipeline to prevent altering the meaning of the query.

3.6 Query expansion and rewriting

It adopts the vocabulary difference between ordinary lan- guage (user's language) and formal language (legal language) and applies query expansion and rewriting approaches to solve the problem. All the other legal terms, synonyms, statutory phrases and concepts are drawn from the initial query. If a question is "without explanation" it might have an accompanying "grounds of arrest," "procedural safeguards," "constitutional rights," and "communication requirements". The expanded queries are elements of the retrieval process, but are not used as the only elements of retrieval. This is possible thanks to recent work on generative query rewriting for legal passage retrieval [17].

3.7 Dense retrieval

In the first retrieval stage, the dense semantic retrieval aims to map the user query to a legal document and to chunks of it using a dense embedding model. Even if the words used in the query are different from those in the passages, the system is able to identify passages semantically related

to the query. It can be helpful in questions that are worded in natural language but have mismatches in vocabulary as is common in legal questions. For some identifiers such as section number, article number, case citation and statutory terms, dense retrieval may not be as accurate. As such, dense retrieval is part of the semantic layer of the proposed retrieval pipeline to be followed by sparse lexical retrieval in the next stage.

4 Results and discussion

The proposed Legal Document Question Answering system based on RAG is tested for its effectiveness in retrieving relevant legal information and providing relevant and evidence based answer to the query. The assessment is based on retrieval quality, answer accuracy, relevance, faithfulness, citation correctness and reducing hallucinations. The proposed framework is compared with an LLM-only model, with traditional vector- based RAG, BM25-based retrieval and proposed hybrid retrieval model. The numerical values of the final study will be from real experiments.

4.1 Overall system performance

Figure 5 compares the performance of the legal RAG approaches.

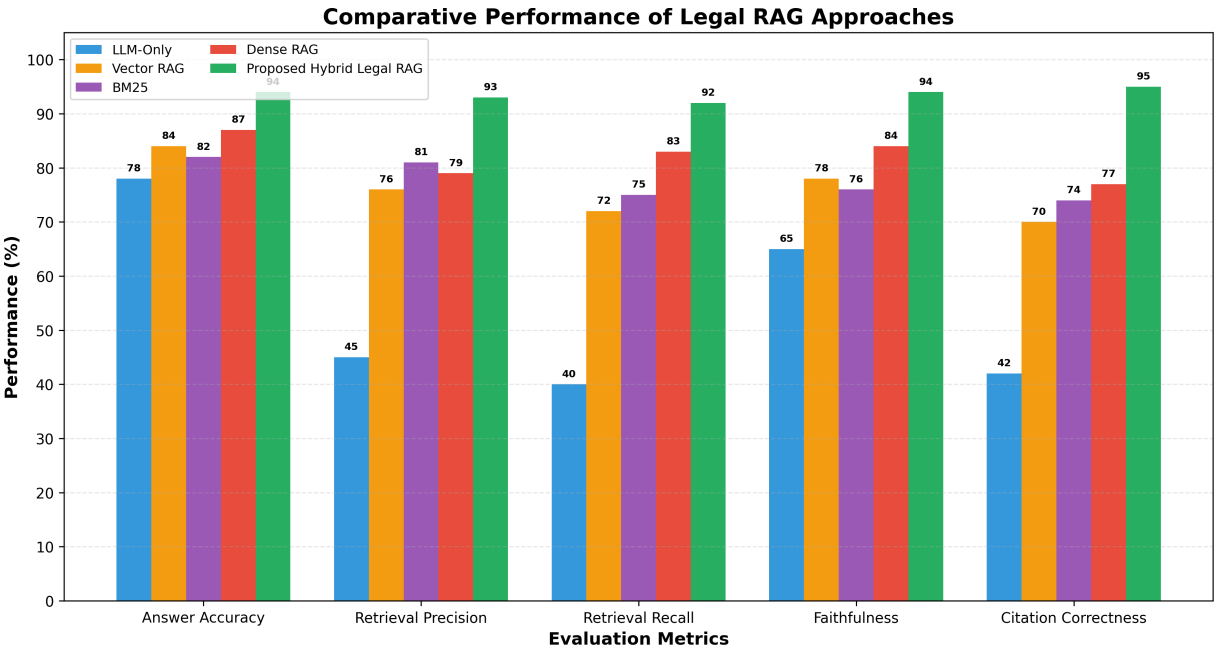


Figure 5. Comparative performance of legal RAG approaches.

Due to providing relevant external legal evidence to the language model prior to the generation of answers, the proposed system is expected to perform better compared to an LLM-only system. An LLM-only approach largely relies on its prior knowledge, can produce incorrect, dated or unrelated legal statements. The proposed RAG framework, on the other hand, fetches the relevant statutes, judicial decisions, sections, clauses, and paragraphs and gives them as contextual evidence. Answer accuracy, retrieval precision, retrieval recall, faithfulness, citation correctness and hallucination rate are used to evaluate the overall system. These measures assess fluency of the generated answer, as well as the legal reliability and grounding in evidence. In the legal context, this may be especially significant; a fluent response could be wrong, but it would be accepted with confidence.

4.2 Retrieval performance

The quality of retrieval is important as the final response is dependent upon the importance of the legal evidence retrieved. Precision@K, Recall@K and Mean Reciprocal Rank (MRR) is used to evaluate the system. The percentage of relevant passages in the retrieved result is measured by Precision@K and Recall@K is a measure for whether relevant evidence is retrieved. MRR is a measure of the ranking of the most relevant passage in the passage set.

The proposed hybrid retrieval method is supposed to outperform the individual dense and sparse retrieval methods. Dense retrieval works well for conceptual queries, for which the query wording varies from the wording of the legal document; BM25 is useful for queries that are exactly the same as the legal document or that are using exact legal terminology, section numbers, article numbers, or case names. A combination of both methods thus improves for the retrieval of both semantic and lexical matches.

4.3 Impact of legal-aware chunking

Results for the Legal-Aware chunking are compared with the results of fixed-size chunking. When chunking is of a fixed size, it might break a legal provision off from its subsection, exception, explanation or related paragraph, leading to missing context. This can lead to a misinterpretation of the legal provision by the LLM.

The suggested legal conscious chunking process is able to maintain relationships between sections, subsections, clauses and paragraphs. Thus, better retrieval performance and accuracy of answers would demonstrate the significance of retaining the structural context of legal documents. This is especially helpful with statutes or decisions that contain a great deal of text and/or other provisions that may rely on the meaning of the one that has been the subject of the question.

4.4 Impact of hybrid retrieval

Proposed framework is a dense retrieval with sparse retrieval based on BM25. There are two types of retrieval: sparse retrieval finds exact matches of the query in legal documents, while dense retrieval finds semantic relations between the query and legal documents. For instance, a user might want to know if he is protected from self-incrimination, while the legal document might use constitutional language.

Dense retrieval can retrieve the conceptual relationship and BM25 can retrieve the specific legal terms and references. Thus, the better Precision@K, Recall@K, and MRR scores for the hybrid approach could prove that the two retrieval methods (semantic and lexical) could complement each other in Legal RAG.

4.5 Impact of cross-encoder reranking

The cross-encoder reranking stage is tested to see if it enhances the quality of the evidence retrieved at the end. The first retrieval process can retrieve many pieces of content that seem relevant to the query given by the user, but do not actually answer it, due to the use of similar words.

The cross-encoder compares the query and the candidate passage to each other and ranks the candidate passage that is more relevant to the query higher. It will be demonstrated by an improvement in the top-ranked precision and MRR from a reranking to show the contribution to the

proposed framework. This is significant in legal papers as the same words can be used in various cases—albeit with varying legal context and application.

4.6 Answer accuracy and relevance

The answers generated are judged on the following criteria: correctness, relevance, completeness and clarity. The proposed retrieval and reranking approaches are anticipated to enhance the quality of the answers generated by the LLM, by supplying it with trustworthy sources of external evidence.

There are various types of errors and they are classified based on the source. A retrieval failure occurs when the appropriate legal passage is not retrieved. If the right passage is found but the LLM misinterprets the passage, it is regarded as a generation or reasoning error. The distinction can be used to decide if there is a need for improvement of the retrieval component, or the language model.

4.7 Faithfulness and hallucination reduction

Faithfulness evaluates if the information in a generated re- sponse aligns with the legal information retrieved. The answer might be fluent, but may contain unsupported or incorrect legal claims. For this reason, for each major claim, the extracted context is compared to the claim to see if there is support for it.

This proposed RAG framework is believed to deliver increased faithfulness as the outputs are based on retrieved legal sources, compared to an LLM-only system. RAG is not able to fully remove hallucinations, though. Hence, the suggested citation-verification phase is employed to uncover claims that aren't backed by sources and to diminish further hallucinated legal details.

4.8 Citation correctness

The correctness of the citation is judged separately from the correctness of the answer as it can be correct but the citation may be incorrect or not relevant. The system analyzes the presence of citations, their relevance, the level of evidence and completeness of the citation.

The content of each of the major claims is checked against the corresponding legal text to determine if there are any inaccuracies in the source. More precise and complete citations would suggest a more effective linkage of legal claims, with the appropriate authoritative sources, via the proposed verification process.

This is especially crucial in legal research where the user must be able to check crucial points by finding the text of the law, regulation or judicial ruling. In this way, the proposed Legal RAG framework enhances transparency, traceability, and trustworthiness of legal citations and decreases unsupported or incorrect legal citations.

5 Conclusion

This paper introduced a framework for Legal Document Question Answering (QA) with the Large Language Models (LLM) approach with RAG (Retrieval Augmented Generation). The framework tackles important problems with traditional LLM-based legal QA, such as hallucination, outdated information, missing context, and missing evidence. It carries out legal document preprocessing, legal-aware chunking, legal metadata enrichment, processing of queries, dense sparse retrieval, hybrid retrieval, reranking and evidence grounded answer generation.

This dense and BM25-based sparse retrieval is able to pick up semantic relationship and exact legal terms like sections, articles and case citations. Processing and reranking are further enhanced by processing the retrieved content in a way that is aware of the requirements of the law, and by verifying the citations of retrieved content to increase transparency and reliability of the answers generated. The proposed evaluation takes into account retrieval performance and accuracy of the answers, faithfulness, correct citation, rate of hallucination and evaluation by experts.

The framework is applicable to the legal field, especially in India for which the information retrieval problem is complex due to huge collection of statutes, constitutional provisions, regulations and judicial decisions. In summary, the proposed method lays the groundwork for future efforts in establishing more robust, explainable, and data-driven Legal AI systems. Such systems should, however, be used to aid and not supplant, and should rely on a professional legal perspective and on human judgment.

6 Future work

There are a number of ways the suggested Legal RAG framework can be further enhanced. Firstly, to provide support for Indian languages like Hindi, Punjabi, Bengali, Tamil and Telugu, one can develop Multilingual and Cross-lingual Legal QA to retrieve information from English and Indian language documents.

Second, future research may explore the detailed analysis of legal information in the context of advanced legal-aware retrieval and reasoning, which takes into account the hierarchical and rhetorical structure of legal texts, such as facts, legal issues, arguments, precedents, judicial reasoning and final decisions. To facilitate multi-hop legal reasoning, connections between cases, statutes, sections, courts, judges, principles and citations can be represented as legal knowledge graphs and incorporated into RAG for multi-hop reasoning.

The temporal and jurisdiction-aware reasoning is another significant way. As laws can change, be repealed and/or be differently interpreted in different jurisdictions, future systems should specify the jurisdiction and the date of application. The verification of claims on the claim level can also be improved to identify “unsupported claims”, “contradictory evidence”, “fabricated citations” and “incorrect interpretations”.

For further research, multimodal legal document processing of scanned documents, tables, signatures and handwritten documents should be further investigated, as well as long-context reasoning for complex queries, involving multiple statutes and judgments. Evaluation of bigger and more representative datasets of real-world examples using legal practitioners, researchers, advocates, and law students can further evaluate for accuracy, usability, explainability, and trust.

Finally, it is important for practical Legal AI systems to facilitate continuous knowledge updates, as well as keeping versions of laws, judgments, amendments and regulations historic. Legal documents will also be confidential and security mechanisms will be important when handling them. These advancements enable the creation of Legal AI systems that are more capable of processing multiple languages, delivering CLEAR explanations, being aware of the temporal context, capable of resolving jurisdiction issues, and ensuring proper citations are maintained. Such advancements can result in more reliable, explainable, temporally aware, jurisdiction-sensitive, and citation-verified Legal AI systems that offer reliable assistance and appropriate human oversight.

References

[1] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kuttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems, vol. 33. 2020.

[2] Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, Chen D, Yih WT. Dense Passage Retrieval for Open-Domain Question Answering. In: *Proc*. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP). 2020:6769–6781.

[3] Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, Aletras N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In: *Proc*. 60th Annual Meeting of the Association for Computational Linguistics (ACL). 2022:4310–4330.

[4] Guha N, Nyarko J, Ho D, Ré C, Chilton A, Choi A, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems, vol. 36. 2023.

[5] Huang Q, et al. Lawyer LLaMA Technical Report. *arXiv* preprint arXiv:2305.15062. 2023.

[6] Yue S, et al. DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services. *arXiv* preprint arXiv:2309.11325. 2023.

[7] Mavi V, Saparov A, Zhao C. Retrieval-Augmented Chain-of-Thought in Semi-Structured Domains. In: *Proc*. Natural Legal Language Processing Workshop. 2023:178–191.

[8] Ryu C, Lee S, Pang S, Choi C, Choi H, Min M, Sohn JY. Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA. In: Proc. Natural Legal Language Processing Workshop. 2023.

[9] Shui R, Cao Y, Wang X, Chua TS. A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. In: *Findings of the Association for Computational Linguistics: EMNLP*. 2023:7337–7348.

[10] Wiratunga N, Abeyratne R, Jayawardena L, Martin K, Massie S, Nkisi-Orji I, Weerasinghe R, Liret A, Fleisch B. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. *arXiv* preprint arXiv:2404.04302. 2024.

[11] Pipitone N, Houir Alami G. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. *arXiv* preprint arXiv:2408.10343. 2024.

[12] Kalra R, Wu Z, Gulley A, Hilliard A, Guan X, Koshiyama A, Treleaven PC. HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications. In: *Proc*. NAACL-HLT Industry Track. 2025:1036–1054.

[13] Wiggers G, Hartz C. LeReRAG: Measuring Legal Relevance in Retrieval Augmented Generation Applications. In: *Legal Knowledge and Information Systems*. 2024.

[14] Redelaar F, van Drie R, Verberne S, de Boer M. Attributed Question Answering for Preconditions in the Dutch Law. In: *Proc*. Natural Legal Language Processing Workshop. 2024.

[15] Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*. 2025;22(2):216–242.

[16] Hou AB, Weller O, Qin G, Yang E, Lawrie D, Holzenberger N, Blair-Stanek A, Van Durme B. CLERC: A Dataset for U.S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation. In: *Findings of NAACL*. 2025:7913–7928.

[17] Kim D, Kang D, Kim J, Ryu S, Lee G. GuRE: Generative Query Rewriter for Legal Passage Retrieval. In: *Proc*. Natural Legal Language Processing Workshop. 2025.

[18] Reuter M, Lingenberg T, Liepina R, Lagioia F, Lippi M, Sartor G, Passerini A, Sayin B. Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. In: *Proc*. Natural Legal Language Processing Workshop. 2025:17–30.

[19] Li A, Wu Y, Liu Y, Cai M, Qing L, Wang S, Kang Y, Liu C, Wu F, Kuang K. UniLR: Unleashing the Power of LLMs on Multiple Legal Tasks with a Unified Legal Retriever. In: *Proc*. 63rd Annual Meeting of the Association for Computational Linguistics (ACL). 2025:11953–11967.

[20] Nigam SK, Mishra SK, Mishra AK, Shallum N, Bhattacharya A. Legal Question-Answering in the Indian Context: Efficacy, Challenges, and Potential of Modern AI Models. *arXiv* preprint arXiv:2309.14735. 2023.