IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 February 2025, accepted 6 March 2025, date of publication 11 March 2025, date of current version 20 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3550145

SURVEY

Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey

MAHD HINDI, LINDA MOHAMMED, OMMAMA MAAZ, AND ABDULMALIK ALWARAFY, (Member, IEEE)

Department of Computer and Network Engineering, College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding author: Abdulmalik Alwarafy (aalwarafy@uaeu.ac.ae)

This work was supported by United Arab Emirates University (UAEU) under Grant 12T047.

ABSTRACT Retrieval-Augmented Generation (RAG) is a promising solution that can enhance the capabilities of large language model (LLM) applications in critical domains, including legal technology, by retrieving knowledge from external databases. Implementing RAG pipelines requires careful attention to the techniques and methods implemented in the different stages of the RAG process. However, robust RAG can enhance LLM generation with faithfulness and few hallucinations in responses. In this paper, we discuss the application of RAG in the legal domain. First, we present an overview of the main RAG methods, stages, techniques, and applications in the legal domain. We then briefly discuss the different information retrieval models, processes, and applied methods in current legal RAG solutions. Then, we explain the different quantitative and qualitative evaluation metrics. We also describe several emerging datasets and benchmarks. We then discuss and assess the ethical and privacy considerations for legal RAG and summarize various challenges, and propose a challenge scale based on RAG failure points and control over external knowledge. Finally, we provide insights into promising future research to leverage RAG efficiently and effectively in the legal field.

INDEX TERMS Information retrieval, large language model (LLM), legal technology, prompt engineering, retrieval-augmented generation (RAG).

I. INTRODUCTION

Legal technology, which is also referred to as legal tech, emerged around 2010 as a technological solution or tool used in the legal domain to support legal services implemented in different sectors to ordinary users and legal professionals, including lawyers and other legal practitioners [1]. Technological innovations have influenced the evolution of legal tech solutions, enabling them to provide high-quality services in terms of efficiency, transparency, cost, and time [2]. This began with digitalizing legal content, followed by automating routine legal tasks, and now moving toward advanced Artificial Intellegence (AI) integration [3].

Natural Language Processing (NLP) has enabled many applications in the legal sector, covering a wide range of areas, e.g., legal research, electronic discovery, contract review, document automation, and legal advice [4]. Innovations in Large Language Model (LLM) techniques with the emergence of transformer models have realized superior performance, parallelizability, and faster training times for sequence transduction tasks compared with traditional recurrent or convolutional neural networks [5], [6], [7], [8], [9], thereby opening the door to more powerful LLM-driven applications in various domains, including the legal domain. However, in legal applications, LLMs have exhibited high hallucination rates [10]. To address this issue, LLMs have been enriched with prompt engineering [11], fine-tuning processes [12], [13], or retrieval-augmented generation (RAG) [14], to obtain better results in terms of precision and

The associate editor coordinating the review of this manuscript and approving it for publication was Hai Dong.

TABLE 1. List of acronyms provided in the paper.

AI	Artificial Intelligence
API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy
Eval	Evaluation
FAISS	Facebook AI Similarity Search
FP	Failure Point
FT	Fine-tuning
IEM	Investigation Enhancement Model
IR	Information Retrieval
KG	Knowledge Graph
LEAs	Law Enforcement Agencies
Legal RAG	Retrieval-Augmented Generation in Legal Domain
LJP	Legal Judgment Prediction
LLM	Large Language Model
M&A	Mergers and Acquisitions
MAN	Manual
MAP	Mean Average Precision
MAUD	Mergers and Acquisitions Understanding Dataset
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MMR	Maximum Marginal Relevance
MRR	Mean Reciprocal Rank
NDA	Non-Disclosure Agreement
NLP	Natural Language Processing
QA	Question-Answer
RAG	Retrieval-Augmented Generation
RAGAs	Retrieval-Augmented Generation Assessment [26]
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RT	Re-Training

hallucination. RAG was proposed to enhance generators to achieve less hallucination and offer more interpretability and control [14].

Different studies have proved that RAG outperformed fine-tuning processes for existing knowledge encountered during training and entirely new knowledge [15], [16]. Thus, since RAG was first introduced in 2020 by Lewis et al. [14], different RAG systems have been rapidly developed for various domains, including the legal domain. The most powerful feature of RAG is its ability to adapt recent or specific external knowledge rapidly and dynamically retrieve relevant information from external sources during the generation process [17]. Numerous legal applications have demonstrated that a combination of RAG and fine-tuning methods perform well [18], [19], [20], [21].

In 2024, more than 20 legal RAG pipelines were implemented using various embedding, retrieval, enhancement, and generation methods. These pipelines have been frequently integrated with other approaches, e.g., prompt engineering, which is essential for all RAG pipelines, and with knowledge graphs (KG) [22], and fine-tuning (FT) [18], [19], [20], [21], [23], [24] or embedded within multi-agent frameworks [25]. In addition, legal RAG pipelines span various applications across the legal domain, ranging from specialized systems focused on specific legal fields to more comprehensive legal platforms.

A. CONTRIBUTION AND ORGANIZATION

Given the absence of a comprehensive literature survey on RAG systems within the legal domain, this paper attempts to bridge this gap by providing a detailed overview of all currently available RAG methods relevant to the legal domain. The primary contributions of this paper are summarized as follows.

- We highlight various techniques utilized in RAG methods specific to the legal field.
- We examine how these methods contribute to improvements in accuracy and interpretability.
- Our analysis of RAG methods and techniques in the legal field offers insights into various applications, methodologies, evaluations, datasets, and benchmarks.
- We extensively outline and describe some open challenges for RAG application in the legal domain and provide deep insights into promising future research directions in RAG applications.

The findings of this work are expected to guide legal tech researchers who aim to use cutting-edge technology to optimize LLM- driven legal applications and practices for various tasks. In addition, this study will serve as a contemporary reference for RAG methods in the legal field.

The remainder of this paper is organized as follows: Section II provides an overview of RAG methods, RAG main stages, and techniques as well as a classification of legal RAG methods, applications, and datasets. Section III explores advanced methods that improve retrieval accuracy in legal RAG systems, addressing the unique needs of legal information retrieval tasks. Section IV explains the quantitative and qualitative metrics used to analyze retrievers and generators in RAG systems, and Section V describes relevant emerging datasets and benchmarks. Section VI, evaluates ethical and privacy considerations in legal RAG systems. Section VII focuses on the main challenges of legal RAG systems. Section VIII provides insights into promising research directions. Finally, the paper is concluded in Section IX.

II. RAG IN THE LEGAL DOMAIN

A. OVERVIEW OF RAG

RAG comprises three key processes, i.e., the information retrieval (IR), augmentation, and generation processes. Many processes and techniques are involved before and after the IR process is performed to enhance the process and its outcomes.

The IR process is a critical element in the RAG framework. The generator will likely produce poor outcomes if the retrieved data are inaccurate or inconsistent with the query. A powerful IR method can outperform the performance of a combined IR LLM [23]. thus, the IR process is the backbone of the entire RAG pipeline [27]. IR techniques have been improved over decades from initial traditional sparse IR techniques [28], e.g., BM25 [29] and TF-IDF [30], to more advanced dense Transformer-based embedding models, e.g., DPR [31]. Transformer-based retrieval methods outperform nonneural methods, specifically for legal document retrieval tasks [32]. In addition, KG embeddings have been integrated with IR embeddings to optimize the IR process [33]. Advanced RAG systems involve pre- and post-retrieval

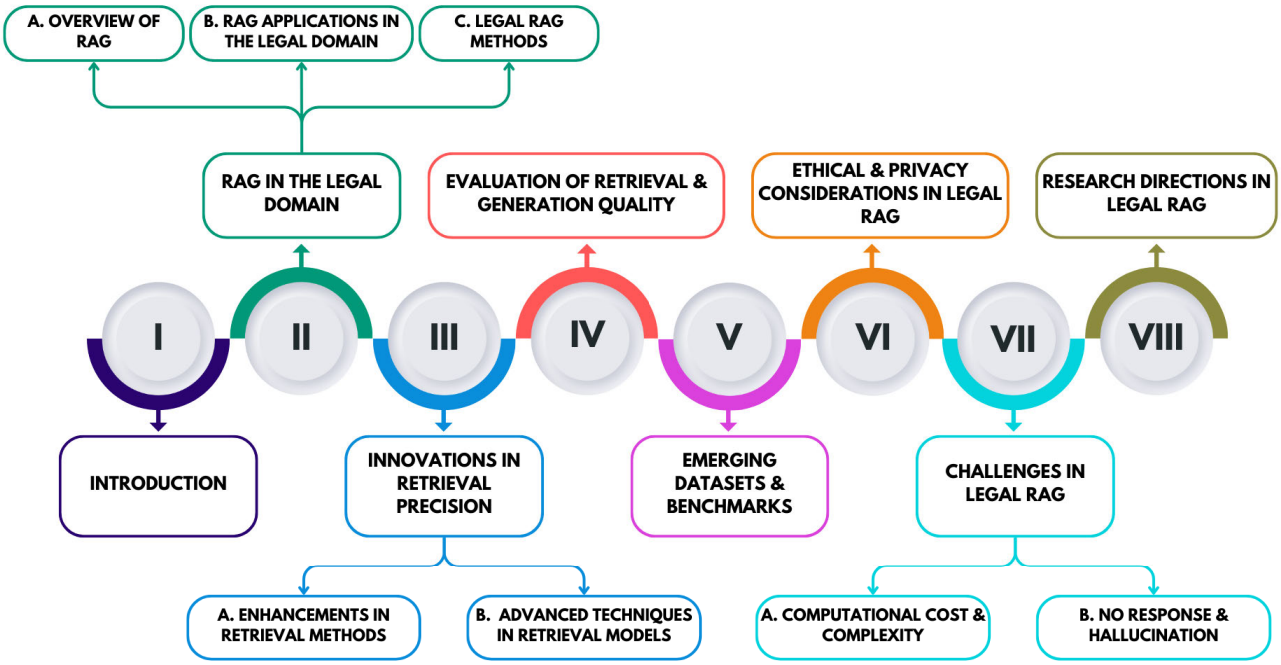

FIGURE 1. Paper organization.

enhancements to enrich the IR process and produce accurate and precise results [34]. The IR process can be optimized based on the complexity of the problem and the required reasoning steps. An effective method to optimize IR is to apply different methods for IR rather than relying on a single IR process. Selecting an appropriate IR process depends on the complexity of the task and required reasoning steps [34]. In addition, many methods and techniques are used to optimize IR, e.g., the embedding model and query enhancements [35]. The chunking method, which plays a pivotal role in IR, is influenced by the nature of the dataset to be retrieved, how the texts are organized in the dataset, what will be retrieved from the datasets, and how much semantic similarity is important in the IR process. The precision of the IR process is strongly dependent on the selection of the chunking strategy [34], [36], [37], [38].

The augmentation process, which integrates retrieved information and query fragments in the LLM, can be performed in three ways in the input, output, and intermediate layers of the generator [36]. KG embeddings can enrich the prompt for more accurate response generation by integrating the retrieved triplets with the retrieved chunks and the original user's query, as proposed in [22] and illustrated in Fig. 2. Prompt engineering with one or few shots has demonstrated more accurate responses compared with zero-shot prompting [23].

In the generation process, the LLM can be retrained or fine-tuned on legal data using a parameter-accessible LLM or the RAG pipeline may utilize a parameter-inaccessible "frozen" LLM [34], [36].

B. RAG APPLICATIONS IN THE LEGAL DOMAIN

Despite being introduced in 2020 [14], RAG was not applied to the legal domain until 2023. The first research paper was by Shui et al. [23], who employed RAG to predict legal judgments. While they did not discuss the abstract term RAG explicitly, they did introduce the RAG system and referred to the process as "LLMs coordinate with IR (LLM + IR)."

We conducted an extensive literature review by examining relevant papers from four major academic databases: SCOPUS, IEEExplore, Web of Science, and Google Scholar, in addition to preprints posted on arXiv. The query we used to search for related papers included three main terms. The first term included all keywords related to legal, including 'legal', 'legal case', 'judiciary,' 'judicial,' and 'law'. The second term included all keywords related to the RAG, including "retrieval", "augmented" and "generation". The third term included all keywords related to LLM models, including "LLM", "transformer model" and "generative AI". As RAG was first introduced in 2020, the search was restricted to articles published between 2020 and 2024. The gathered papers were then scanned based on titles, abstracts, and keywords to determine relevant articles for further analysis. As this field is newly emerging, the final selection includes only 22 papers.

As shown in Fig. 3, RAG research experienced a notable surge in 2024, spanning various applications within the legal domain. The RAG methods proposed in the legal field address various areas, e.g., privacy law, legislative texts, public law, criminal law, statutory law, and immigration law. These applications are employed in various systems, including

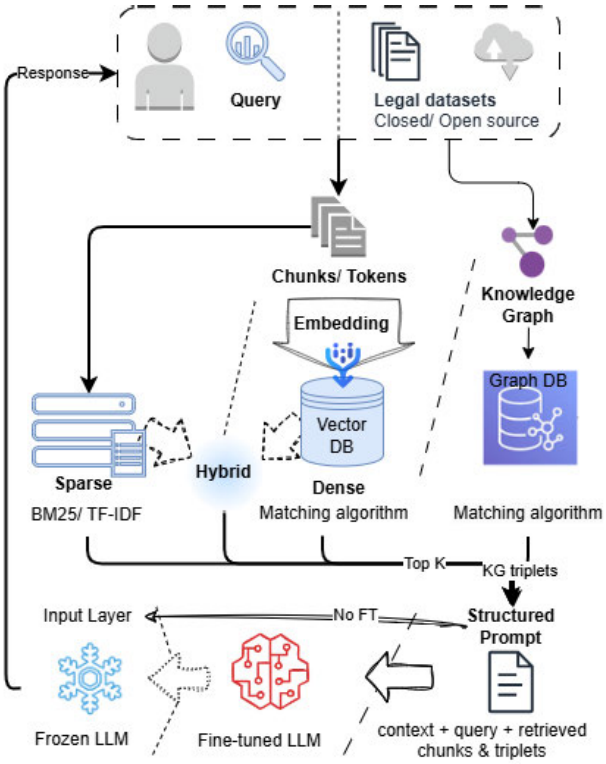

FIGURE 2. RAG pipeline enriched with knowledge graph and LLM fine tuning (inspired and elicited from legal RAG systems).

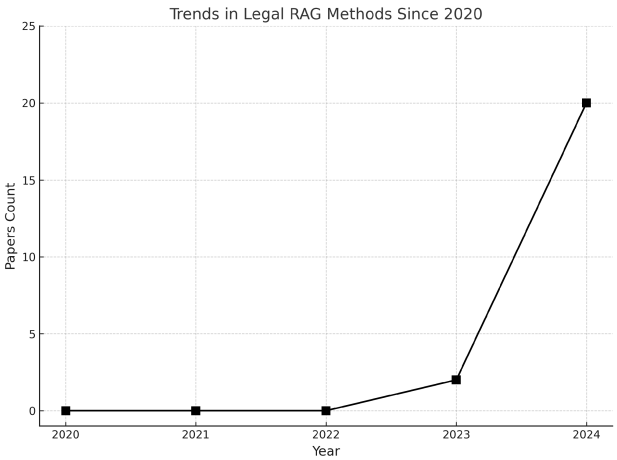

FIGURE 3. Trends in legal RAG methods since 2020.

question answering, recommendations, legal advice, case reasoning, legal chatbots, digital assistants, and predicting legal judgments. In addition, three datasets have been introduced recently to provide benchmarking solutions when developing and evaluating RAG models. Detailed information about the legal RAG methods and relevant applications is given in Table 2.

The common objectives of legal RAG systems are to enhance efficiency, accuracy, accessibility, and contextual understanding and to improve legal and regulatory services targeting different fields in the legal domain. In addition, RAG methods in the legal domain attempt to address the limitations of previous LLM-driven systems in terms of hallucination [18], [25] and outdated knowledge [49]. They also improve IR [20], [44], support legal accessibility and usability [24], and automate complex legal processes [45], [52].

1) CASE STUDY: SEMANTIC INTERLINKING OF IMMIGRATION DATA USING LLMS AND RAG FOR KNOWLEDGE GRAPH CONSTRUCTION

The study on semantic interlinking of immigration data using LLMs and RAG presents a transformative approach to managing complex legal data, specifically in the context of the U.S. Adjustment of Status (AOS) process [22]. The framework addresses inefficiencies in processing heterogeneous, paper-based immigration records by converting them into structured, interconnected knowledge graphs (KGs). By employing advanced key-value mapping strategies and integrating Retrieval-Augmented Generation (RAG) techniques with LLMs, the system enables accurate extraction of entities and relationships from legal documents. The constructed KGs, ingested into Neo4j, provide a detailed representation of the AOS process, allowing legal professionals to retrieve context-aware and semantically enriched insights with simple English queries. This innovation enhances decision-making and data management while maintaining data privacy through the potential use of local LLMs. The results demonstrate how this methodology simplifies the complexities of legal processes, offering a scalable and adaptable solution for immigration and other legal domains.

C. LEGAL RAG METHODS

Legal RAG systems are built on a variety of architectural designs, each tailored to handle the complexity of legal text retrieval, augmentation, and generation. While some frameworks employ a standard retrieval-then-generation pipeline, others integrate more advanced mechanisms such as iterative retrieval, knowledge graph embeddings, and adaptive augmentation. These enhancements refine the retrieval process, improve factual grounding, and optimize generative fluency.

A comprehensive RAG pipeline in the legal domain is illustrated in Fig. 2. Here, the RAG system is enriched with a fine-tuned generator and the embedding of a KG. The user query and legal documents are preprocessed, chunked, embedded, and stored in a vector database, and the similarity between the query and the chunks is calculated. Here, relevant chunks are retrieved, integrated, and passed to the fine-tuned or pre-trained (i.e., frozen) LLM, which can be enriched with the KG triplets using a structured prompt, and then the response is generated accordingly. Fig. 2 summarizes all currently surveyed RAG systems in the legal domain.

It is essential to incorporate the appropriate strategies and optimization at each pipeline stage to build an effective legal RAG system. The following sections explore key

TABLE 2. Legal RAG methods and applications.

Application Type	Ref	Year	Method Name	Field of Legal Domain	Purpose
Case Reasoning	[39]	2024	CASEGPT	Criminal, civil, and admin cases	Enhance case-based reasoning
	[40]	2024	MVRAG	Case law	Case law retrieval and legal research applications
	[23]	2023	LJP	Criminal law	Evaluate the competency of LLMs in LJP
Chatbot	[41]	2024	Chat-EUR-Lex	Legislative documents	Improve the accessibility of EU laws
	[42]	2024	TaxTajweez	Tax law	Provide income tax advisory services
Dataset	[43]	2024	CLERC	Case Law	Assist in the development and evaluation of RAG models
	[44]	2024	LegalBench-RAG	Agreements, privacy	Evaluate the retrieval of RAG systems in the legal domain
Dataset & QA	[21]	2024	LLeQA	Statutory law	Generate long-form answers to statutory law questions
Digital Assistant	[45]	2024	PRO	Patent law	Optimize patent response generation
	[46]	2024	/	Privacy and data protection	Support governance applications
Legal Advisory	[47]	2024	/	Banking regulations	Enhance legal advice in the banking sector
	[22]	2024	/	Immigration law	Digitize immigration records into KG
	[25]	2024	Chatlaw	Broad	Mitigate the hallucination problem inherent in LLMs
QA	[48]	2023	Eval-RAG	Statutory law	Enhance the evaluation of texts generated by LLMs
	[18]	2024	DRAG-BILQA	Border inspection	Provide legal responses to border inspection-related queries
	[20]	2024	CBR-RAG	Broad	Improve contextual understanding and provide evidence-based QA
	[49]	2024	HyPA-RAG	AI-specific law	Addressing the limitations of LLMs in complex contexts
	[50]	2024	/	Privacy law	Assist users in identifying potential privacy risks in policies
	[24]	2024	SAVIA	Public law and legislative	Facilitate access to regional legislation
Recommendation	[51]	2024	/	Public procurement	Improve the efficiency of procurement processes
	[19]	2024	IEM	Forensic investigations	Assist LEAs in crime resolution and suspect identification
	[52]	2024	LexDrafter	Legislative documents	Assist in drafting Definitions articles for legislative documents

aspects, including retrieval sources, models, augmentation mechanisms, generation methods, and training approaches.

1) RETRIEVAL SOURCE

The effectiveness of the RAG process is strongly dependent on the external knowledge, which is also referred to as nonparametric memory [14], and the processes to be performed on the external dataset prior to retrieving the required information [44]. The strength of RAG lies in its ability to utilize the retrieval of the most relevant text chunks/segments pertaining to the query or the user's question from the knowledge source, thereby enhancing the LLM's ability to produce the most accurate and appropriate response. Conversely, it will be difficult for the RAG system to produce a response if the information required for the relevant query is not present in the dataset. In some cases, the LLM will provide a hallucinated response [53] or no response [54]. Thus, the external knowledge should be complete, including all information that is relevant to the application.

The dataset retrieval process could be closed-sourced from a specific dataset, which is more suitable for domains that do not involve rapid changes in knowledge, e.g., legal domain in most cases. Alternately, the dataset retrieval process could be open-sourced, where data can be retrieved directly from the Internet and other sources, which is more suitable for applications and domains that involve rapidly changing knowledge [36]. Thus, it is expected that most legal RAG systems employ the closed-source retrieval mechanism. Table 3 summarizes the retrieval dataset types and their corresponding datasets. Additional details about emerging datasets in the legal domain are discussed in Section V.

2) EMBEDDING MODELS

Embedding models convert legal text into high-dimensional vector representations, allowing retrieval models to measure similarity between queries and legal documents [51]. These models capture the meaning of text beyond simple keyword matching, making them essential for dense retrieval.

The most common embedding models used for dense retrievers are transformer-based embeddings, and legal RAG systems primarily employ BERT and BERT-based models along with Open AI's ADA-002 model. In addition, customized models have used non-English embeddings, e.g., bge-large-zh-v1.5, text2vec, and bge-m3 for Chinese dataset embedding [40], and embed-multilingual-v3.0 for Italian dataset embedding [51]. Some RAG pipelines enrich the model with KG embeddings, which enhance the retriever and make the IR process more interpretable [22], [25], [49]. Detailed information on transformer-based embedding models is provided in Table 5.

3) RETRIEVAL METHODS

Transformer-based dense retrieval methods are mainly used in legal RAG applications, and cosine similarity is the most commonly used search algorithm for the retriever. In addition, sparse retrieval has been tested experimentally [43] and outperformed some dense retrievers. Sparse retrieval is employed in three pipelines using BM25 as a ranking model [23], [50], [52]. By combining sparse and dense retrievers, a hybrid approach has been employed in a legal RAG pipeline [49]. As discussed in Section II-A, selecting an appropriate IR method is dependent on various factors related to the nature of the source of the IR and the task related to the IR. For example, in complex legal scenarios [18], dense

TABLE 3. Overview of RAG systems and their associated datasets.

Language	Dataset	Dataset Retrieval	Method
Chinese	BorderLegal-QA, JEC-QA	Closed source	DRAG-BILQA [18]
	Comprehensive legal data	Open & Closed source	Chatlaw [25]
	LeCaRDv2	Closed source	MVRAG [40]
	CAIL (Chinese AI and Law)	Closed source	LJP [23]
English	EUR-Lex legal acts from Energy domain	Closed source	LexDrafter [52]
	Adjustment of Status (AOS)	Closed source	[22]
	50,000 court case summaries	Closed source	CASEGPT [39]
	CLERC	Closed source	CLERC [43]
	NYC Local Law 144 (LL144) of 2021	Closed source	HyPA-RAG [49]
	LegalBench-RAG	Closed source	LegalBench-RAG [44]
	Specific for Patent Response System	Closed source	PRO [45]
	Income Tax Manual	Closed source	TaxTajweez [42]
	EUR-Lex	Closed source	[46]
	Open Australian Legal QA (ALQA)	Closed source	CBR-RAG [20]
English (translated)	Korean Legal QA	Closed source	Eval-RAG [48]
English and Italian	EU legal acts	Closed source	Chat-EUR-Lex [41]
French	LLeQA	Closed source	LLeQA [21]
Italian	Laws enacted in the Emilia-Romagna region	Closed source	SAVIA [24]
	National Authority for Anti-Corruption (ANAC)	open-source	[51]
Not mentioned	Crime-related data	Closed source	IEM [19]
	Domain-specific dataset	Not specified	[47]
Vietnamese	Vietnamese Legal Document	Closed source	[50]

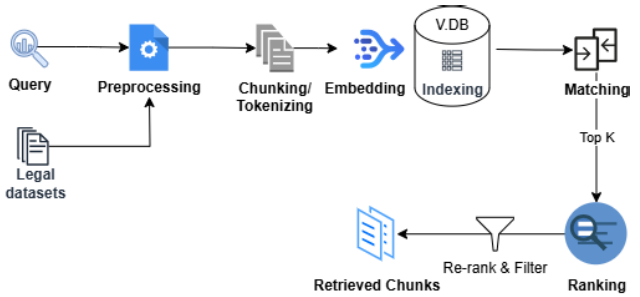

FIGURE 4. Dense retrieval approach used in most legal RAG methods.

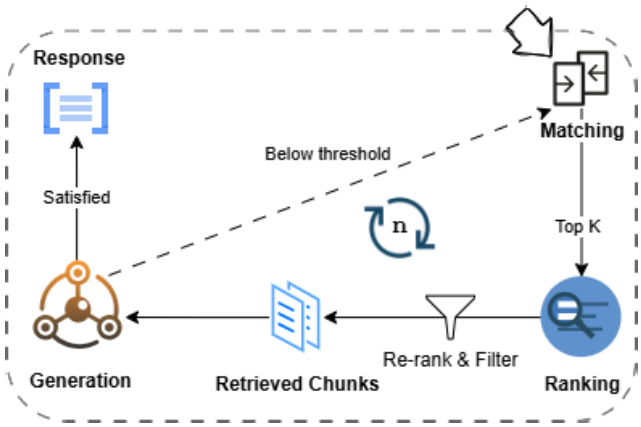

FIGURE 5. Iterative retrieval process with n retrievals.

retrievers are more effective than sparse retrievers in terms of capturing semantic similarity [55]. Additional information about advanced methods that improve the retrieval process in legal RAG systems is given in Section III. Fig. 4 illustrates an advanced dense retrieval approach, and the best-performing transformer-based models in the retrieval process in surveyed legal RAG methods are summarized in Table 4.

4) RETRIEVAL PROCESS

One-time process retrieval is the most common IR process in legal RAG pipelines. The iterative retrieval process used in previous studies [20] and [25] is illustrated in Fig. 5, where the retrieval is repeated n times until a predefined threshold is met. The adaptive retrieval process, (Section III), has also been used in [18] and [49] (Fig. 6), where the RAG system can determine whether to initiate the retriever based on a predefined threshold.

5) AUGMENTATION MECHANISM

The technique of combining the retrieved chunks/segments along with the original query in a structured prompt and passing it to the input layer of the generator (the LLM) was used in all the surveyed methods. Meanwhile, some methods applied additional prompt engineering techniques with one-shot or few-shot prompting to enhance the performance of the LLM to generate more accurate responses [21], [23].

6) GENERATION

GPT-4 is an LLM that is primarily used as a generator in legal RAG pipelines, followed by llama-based LLMs, both GPT-4 and llama-based LLMs of which are the best-performing generators used in different proposed pipelines. Generally,

TABLE 4. Best-performing transformer-based models in the retrieval stage in legal RAG systems.

Method	Model	Transformer Type	Fine-Tuned
CASEGPT [39]	legal-BERT	Domain-specific BERT-based	Yes
CBR-RAG [20]	AnglEBERT	Contrastive learning-enhanced Transformer	No
CLERC [43]	ft-LegalBERT DPR	Legal domain-specific retrieval model	Yes
Eval-RAG [48], Chat-EUR-Lex [41], [51], [47], [42]	text-embedding-ada-002	Dense embedding model for retrieval	No
[21]	CamemBERT	Dense vector retrieval model for French law	Yes
LegalBench-RAG [44], [47], [51]	text-embedding-3-large	Dense embedding model for retrieval	No
SAVIA [24]	Sentence-BERT	Semantic search embedding model	No
HyPA-RAG [49]	distilBERT	Dense embedding model for retrieval	Yes
[22]	Cypher Query Generation based on GPT-3.5	Decoder-Only Transformer	No
[40]	bge-large-en-v1.5	Dense embedding model for retrieval	No

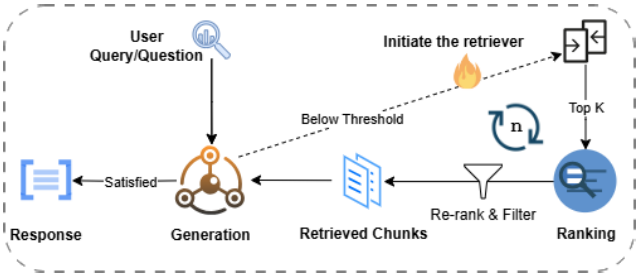

FIGURE 6. Adaptive retrieval process, wherein RAG has the ability to activate/deactivate the retriever.

Open AI LLMs are mainly used because they can generate accurate responses, and Open AI models have outperformed the Cohere models [51]. Seven legal RAG methods surveyed in this study employed FT processes in parameter-accessible LLMs to enhance the performance of the LLMs. It is important to mention here that combining FT processes with RAG can fail if the retrieval is not performing accurately [43]. Table 5 provides additional information about best-performing LLMs.

7) TRAINING APPROACHES FOR LEGAL RAG MODELS

Training legal RAG models involves optimizing retrieval accuracy, augmentation strategies, and generative fluency through fine-tuning and parameter-efficient adaptations [21], [49]. Some systems train the retrieval and generation modules separately, while others fine-tune them jointly to improve overall performance [56].

Legal factor recognition training is a key component in certain RAG frameworks such as DRAG-BILQA. In this approach, the model first estimates its confidence in generating an accurate response. If the confidence score meets a predefined threshold, the response is outputted directly; otherwise, additional retrieval is triggered to refine the input context, thereby enhancing answer accuracy [18].

Dense retrieval models (e.g., DPR, ColBERT) are typically trained using contrastive learning to improve query-document matching. DRAG-BILQA employs a dual-encoder recall model for initial retrieval and a cross-encoder re-ranking model to refine the set of retrieved legal passages. The recall model is trained with in-batch negative sampling so that positive legal document embeddings are closer to their corresponding queries than negative samples. Common hyperparameter settings include batch sizes of 32-64, learning rates between 1e-5 and 3e-5, and training epochs ranging from 2 to 10 (see also Section II for further details on contrastive learning) [18], [21].

Generation model fine-tuning adapts large language models (LLMs) for domain-specific legal text generation. Parameter-efficient tuning methods such as QLoRA and prompt tuning are widely adopted to minimize computational costs. For example, DRAG-BILQA fine-tunes ChatGLM2-6B using QLoRA with a LoRA rank of 8, a dropout rate of 0.1, a maximum sequence length of 512, and a batch size of 16 at a learning rate of 1e-5. Additionally, prompt tuning is performed with a batch size of 8 over 4 steps to optimize prompt-specific learning. In contrast, full fine-tuning is applied in specialized cases—such as adapting CamemBERT for case law reasoning using a batch size of 32, a learning rate of 2e-5, weight decay of 0.01, and 20 training epochs with the AdamW optimizer [18], [21].

While Section II outlines the core components of legal RAG systems—including retrieval, augmentation, and generation—the overall performance depends critically on both the architectural design and the precision of the retrieval mechanisms. Advanced methods that integrate multiple datasets and optimization techniques are further discussed in Section III.

III. INNOVATIONS IN RETRIEVAL PRECISION

In this section, we discuss advanced methods used to improve retrieval precision and relevance in legal RAG systems. These innovations address the challenges inherent in legal IR, including the complexity of legal language, domain specificity, and need for responses that are contextually accurate.

A. ENHANCEMENTS IN RETRIEVAL METHODS

1) PRERETRIEVAL OPTIMIZATION

In RAG systems, it is essential to efficiently retrieve relevant documents from the data source. Thus, the pre

TABLE 5. Summary of legal RAG methods and techniques.

Method	Retriever	Retrieval Type	Matching Algorithm	Retrieval Process	Best Generator	RT/ FT LLM
DRAG-BILQA [18]	DPR	Dense	Dot product	Adaptive	ChatGLM2-6B	FT
Chatlaw [25]	KG	KG	Not mentioned	Iterative	Not mentioned	FT
MVRAG [40]	bge-m3	Dense	Cosine similarity	Once	Not mentioned	No
LJP [23]	BM25	Sparse	BM25	Once	GPT-4	No
LexDrafter [52]	TF-IDF	Sparse	TF-IDF	Once	Vicuna-7b-v1.5	No
[22]	KG	KG	Not mentioned	Once	GPT-3.5, GPT-4	No
CASEGPT [39]	BERT	Dense	Cosine similarity	Once	GPT-3	No
CLERC [43]	ft-LegalBERT DPR	Dense	Cosine similarity	Once	(Results vary)	No
HyPA-RAG [49]	BM25 + distilBERT + Semantic KG	Hybrid + KG	Not mentioned	Adaptive	GPT-4o	No
LegalBench-RAG [44]	text-embedding-3-large	Dense	Cosine similarity	Once	Not mentioned	No
PRO [45]	PRLLM derived from LLaMA2	Dense + KG	Cosine similarity	Once	PRLLM-70B	RT + FT
TaxTajweez [42]	ADA-2	Dense	Cosine similarity	Once	GPT-3.5-turbo	No
[46]	text-embedding-3-large	Dense	Cosine similarity	Once	GPT-4	No
CBR-RAG [20]	Hybrid AnglEBERT	Dense	Cosine similarity	Iterative	Mistral-7B	No
Eval-RAG [48]	ADA-2	Dense	Not mentioned	Once	GPT-4	No
Chat-EUR-Lex [41]	ADA-2	Dense	KNN	Once	GPT-4	No
LLeQA [21]	CamemBERT	Dense	Cosine similarity	Once	WizardLM	FT
SAVIA [24]	Sentence-BERT	Dense	FAISS	Once	Mixtral-8x7B-Instruct	FT
[51]	OpenAI's 3-large	Dense	Cosine similarity	Once	OpenAI (Not specified)	No
IEM [19]	Not mentioned	Dense	Euclidean norm	Once	/	FT
[47]	ADA-2	Dense	Dot product	Once	GPT-4	No
[50]	BM25	Sparse	BM25	Once	llama2-7B-50b-chat	FT

retrieval stage employs several techniques to enhance the accuracy of retrieved information; these techniques include data granularity adjustments, indexing enhancement, and query formulation along with and the selection of a suitable embedding model. Collectively, these techniques facilitate highly precise and structured retrieval of information, which is crucial for legal RAG systems due to the inherent complexity and specificity of legal data [36].

In addition, KG integration transforms conventional data handling approaches and plays a crucial role in improving retrieval precision by structuring legal data into interconnected entities and relationships. For example, combining KGs with LLMs allows retrieval systems to generate context-aware responses using KG triplets as additional context [22], [45], [49]. The PPNet framework encodes legal relationships from judicial sources into a KG, which improves the accuracy of responses [45]. Furthermore, hybrid systems, e.g., HyPA– RAG, utilize KG triplets alongside dense and sparse methods for adaptive query tuning [49].

Query rewriting enhances the retrieval process by reformulating user inputs to better align with indexed data while incorporating related concepts that users may not explicitly mention but are contextually relevant. Some frameworks, e.g., the PA-RAG framework, adapt queries dynamically by selecting the number of rewrites and the top related retrieved chunks (K) values based on query complexity [49]. In addition, the multiview RAG (MVRAG) framework introduces intention-aware query rewriting and leverages multiple domain viewpoints to refine queries in knowledge-dense contexts [40].

In terms of chunking strategies, dividing long legal documents into smaller, manageable chunks improves retrieval precision. Chunk-based embedding strategies ensure that contextual details are preserved within smaller text fragments, which reduces the noise associated with embedding the entire document [47]. HyPA–RAG employs multiple techniques to evaluate the model, including sentence-level, semantic, and pattern-based chunking to balance token constraints and context. It has been demonstrated that pattern-based chunking using corpus-specific delimiters achieves the best retrieval precision, with top scores in context recall and faithfulness. Sentence-level chunking excels in context precision and F1 scores; thus, it is suitable for precise retrieval tasks. In addition, unless heavily tuned, semantic chunking underperforms compared with simpler methods [49].

An efficient indexing mechanism is essential for rapid similarity searches in high-dimensional spaces. To balance search speed and accuracy, CASEGPT employs the Hierarchical Navigable Small World algorithm, which is a state-of-the-art indexing technique. In addition, the system implements an incremental indexing mechanism to support real-time updates, which facilitates the seamless integration of new cases without requiring complete reindexing [39].

2) POST-RETRIEVAL STRATEGIES

The post-retrieval procedure ensures that the retrieved information is presented appropriately and efficiently. Legal documents frequently contain nuanced details that are critical for accurate interpretations; thus, post-retrieval innovations focus on organizing and refining the retrieved content. In the reranking process, rearranging document chunks is essential to reduce the total document pool. This serves as a filter and enhancer in IR, and it provides a more accurate input for processing language models [36]. Legal RAG systems employ various reranking techniques to enhance retrieval precision. For example, PRLLM reranks the retrieved paragraphs according to their significance by prioritizing the most critical paragraph of the examiner, which is followed by other comparable passages [45]. CaseGPT implements a multifactor approach that integrates domain-specific factors, including case recency, citation frequency, and jurisdictional relevance. In addition, it balances relevance and diversity in the retrieved cases using a diversity-aware retrieval process based on the maximum marginal relevance method [39]. Similarly, in ChatLaw, reranking is performed using a two-step evaluation process. Here, an LLM first assesses each document's relevance to the query. Then, a critic model performs critical evaluations by refining the results iteratively by reviewing and selecting the best content. This iterative process ensures that the final response is self-assessed, optimized, and highly relevant [25]. Another approach is used in MVRAG, where documents are reranked based on a recalculated relevance score that integrates multiperspective alignment [40].

B. ADVANCED TECHNIQUES IN RETRIEVAL MODELS

As legal tasks become increasingly complex, advanced retrieval techniques are required to improve the precision and relevance of the results. These techniques are designed to handle the challenges of legal IR tasks by adapting to the nuances of legal queries.

1) HYBRID RETRIEVAL APPROACHES

Sparse and dense embedding approaches capture distinct relevance features. Sparse embeddings are particularly useful for tasks that are dependent on keyword matching and require high precision, focusing on specific words or phrases. However, dense embeddings excel in tasks that demand semantic similarity and contextual understanding because they generate continuous vector representations that capture nuanced meanings beyond surface-level word matching. By adopting hybrid retrieval models, systems can leverage the advantages of both approaches, where the precision of the sparse embeddings are combined with the contextual depth of the dense embeddings to improve the overall retrieval accuracy and relevance [36]. For example, HyPA–RAG adopts a hybrid search engine that combines (dense and sparse) and KG retrieval methods to improve retrieval accuracy [49].

2) ADAPTIVE RETRIEVAL MECHANISMS

Adaptive retrieval mechanisms optimize the retrieval process by adjusting to the complexity of the given query. For example, HyPA–RAG employs a domain-specific query complexity classifier that categorizes queries based on their complexities, which helps the system select the most appropriate retrieval strategy, e.g., the number of subqueries to generate and the top k number of documents to be retrieved. This approach ensures that the retrieval process is efficient while maintaining relevance to the legal context [49]. The dynamic RAG framework for border inspection legal question answering (LQA) controls the retrieval process dynamically based on confidence scores. In the framework, after generating an initial response, the system calculates the confidence score of relevant legal factors. If the confidence score is low, the system triggers an additional retrieval process to enhance the response [18]. The adaptive retrieval process is illustrated in Fig. 6.

3) FINE-TUNING RETRIEVAL MODELS

Fine-tuning retrieval models is essential for aligning embeddings with legal-domain-specific data, particularly when the context diverges considerably from the For example, HyPA–RAG fine-tunes its distilBERT model on legal corpora and CASEGPT adopts a fine-tuned version of Legal-BERT to achieve enhanced retrieval performance [39], [49]. In addition, CamemBERT is fully fine-tuned on the long-form LQA (LLeQA) dataset to improve its ability to handle complex legal queries [21]. Table 4 shows the best Transformers-based retrievers, of which four models are enhanced by fine-tuning processes.

4) ADVANCED SAMPLING STRATEGIES

In contrastive learning, different techniques, e.g., negative sampling and hard negative sampling, play important roles in training the retrieval model to better distinguish between relevant and irrelevant documents. Negative sampling exposes the model to relevant and irrelevant pairs of documents. In contrast, hard negative sampling challenges the model by selecting tough negative examples, thereby improving the model's ability to classify documents accurately. These strategies enable the model to learn in a low-dimensional, high-quality embedding space where relevant question–provision pairs are placed closer together than irrelevant ones [21].

IV. EVALUATION OF RETRIEVAL AND GENERATION QUALITY

Despite the growth in RAG-related research, we only found a few articles outlining state-of-the-art techniques in the legal domain. The evaluation of RAG systems in most state-of-the-art systems was divided into two parts, i.e., retrieval evaluation and response evaluation. The metrics used to evaluate the effectiveness of the retrieval process include precision, recall, mean reciprocal rank (MRR), and mean

average precision (MAP). Precision is the fraction of relevant instances among the retrieved instances, and recall is the fraction of relevant instances retrieved from the total number of relevant cases. MRR is the average of the reciprocal ranks of the first correct response to a set of queries, and MAP is the mean of the average precision scores for each query [57]. These four metrics can be used to evaluate how effectively a retriever identifies and ranks relevant documents in response to the user's query [58].

For response evaluation, the primary goal is ensuring that the response is relevant to the user query and avoid hallucinations. Generation metrics, e.g., METEOR [18], [21], the bilingual evaluation understudy (BLEU) [52], and the recall-oriented understudy for gisting evaluation (ROUGE) [43], are used to determine the response quality for RAG systems in the legal domain. METEOR combines precision, recall, and sentence fluency to accurately calculate the similarity between automatically generated and reference responses to evaluate the effectiveness of text generation tasks. BLEU measures the overlap between a generated response and a set of reference responses by focusing on the precision of n-grams. Finally, ROUGE counts the number of overlapping units, e.g., n-grams, word sequences, and word pairs, between the generated and reference responses considering recall and precision. Utilizing these metrics to evaluate retrieval and generation tasks helps build a robust, efficient, and user-centric RAG system. However, there are no ground truth answers to queries; thus, the focus of the evaluation has shifted to quantitative aspects, wherein the retriever and generator are evaluated separately [59]. In other words, the nature of RAG systems makes them generate unstructured text, which means that qualitative and quantitative metrics are required to assess their performance accurately. Therefore, we adopted a similar approach to Table 5 of [59], indicating the evaluation metrics used for RAG-based systems in the medical domain and ethical principles considered in surveyed studies. By adopting these metrics and considerations, we created Table 7, which shows the evaluation metrics employed in surveyed RAG-based studies in the legal domain. Specifically, we reviewed 22 papers to check for references to the five evaluation metrics to assess their usage in RAG-based legal applications. The five- evaluation metrics were correctness, completeness, faithfulness, fluency, and relevance (context relevance and answer relevance). Correctness means that the response generated by the RAG system must perfectly align with the expected response or be a relevant statement that conveys the same information [60].Completeness refers to RAG-generated responses that are comprehensive and cover all aspects of the anticipated response. Faithfulness indicates that the response must be based on the provided context. RAG systems are frequently utilized in contexts where the factual accuracy of the generated text with respect to the grounded sources is highly significant, e.g., law [26]. Fluency is the ability of a RAG system to generate readable and clear text. Finally, relevance comprises two parts, i.e., the context relevance, which checks if the context of the retrieved information is relevant to the query, and the response relevance, which indicates how relevant the generated response for the given query.

Table 6 shows the quantitative metrics used for each evaluation aspect. These metrics, obtained from surveyed studies, are traditional indicators and do not yet represent a standardized framework for quantifying the quality aspects of RAG systems [34]. Metrics definitions are summarized in Table 9.

Note that there is no standardized evaluation method for RAG systems, and various frameworks utilize different metrics to evaluate RAG systems. One of such dedicated frameworks is RAG assessment (RAGAs) [26]. RAGAs has been employed in previous studies [42], [49] to assess the performance of a RAG pipeline considering four factors, i.e., faithfulness, relevance to the query, relevance to the context, and recall of the context. The RAGAs framework was designed to serve as a universal standard to assess RAG pipelines without requiring access to ground truths. The system uses OpenAI's GPT-4 to determine a score ranging from 0 to 1 for each of the four metrics. The RAGAs score is calculated by determining the average of the assigned scores.

V. EMERGING DATASETS AND BENCHMARKS

The advancement of Retrieval-Augmented Generation (RAG) systems in legal technology heavily relies on high-quality datasets that enable effective retrieval, reasoning, and interpretability. Several benchmark datasets have been introduced to improve legal question answering (LQA), legal information retrieval (IR), and case law analysis. This section reviews key datasets and benchmarks that support the development of RAG systems in the legal domain.

Legal Question Answering (LQA) datasets play a crucial role in training and evaluating models that generate precise legal responses. For example, the BorderLegal-QA Dataset [18], is specialized for legal queries related to border inspections, and it contains 1,329 pairs of questions and answers covering 51 types of questions. The goal is to offer expertly curated question–answer pairs that are applicable to realistic border inspection situations. In addition, the JEC-QA dataset is a collection of multiple-choice questions from the National Unified Legal Professional Qualification Examination. This dataset contains a total of 26,365 questions, and it acts as a standard to assess legal QA systems. The CJRC dataset was created from real-world accounts in Chinese court records, and it contains approximately 10,000 documents and nearly 50,000 question– answer pairs covering a wide range of reasoning scenarios. The CAIL2020 and CAIL2021 datasets [18] improve the reasoning skills required to answer legal questions. The CAIL2020 dataset contains 10,000 legal documents, and the CAIL2021 dataset presents multisegment questions with approximately 7,000 question–answer pairs. The Open Australian Legal Question-Answering Dataset [20] contains more than 2,100 question–answer–snippet triplets generated by GPT-4 using the Open

TABLE 6. Quantity metrics for each quality aspect.

Metric	Correctness	Context Relevance	Faithfulness	Answer Relevance	Fluency
Accuracy	✓	✓	✓	✓	✓
Recall	✓	✓	✗	✗	✗
Precision	✓	✓	✗	✗	✗
MRR	✓	✓	✗	✗	✗
BLEU	✓	✓	✓	✓	✓
ROUGE/ROUGE-L	✓	✓	✓	✓	✓

TABLE 7. Metrics to Evaluate RAG-Based Systems in the Legal Domain. We assess whether the ethical principles of Privacy, Safety, Robustness, Bias, and Trust are Considered.

Method	Correctness	Completeness	Faithfulness	Fluency	Relevance	Privacy	Safety	Robust	Bias	Trust	Eval
DRAG-BILQA [18]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	Auto
[46]	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	Man
CASEGPT [39]	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	Both
CBR-RAG [20]	✓	✗	✓	✗	✓	✗	✓	✓	✗	✗	Auto
Chatlaw [25]	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	Auto
CLERC [43]	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	Both
HyPA-RAG [49]	✓	✓	✓	✗	✓	✗	✗	✓	✗	✓	Both
Chat-EUR- Lex [41]	✓	✗	✓	✗	✓	✓	✗	✓	✗	✓	Both
LLeQA [21]	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	Auto
IEM [19]	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	Both
LegalBench-RAG [44]	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	Both
LexDrafter [52]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	Auto
PRO [45]	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	Man
[47]	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	Both
[51]	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	Man
Eval-RAG [48]	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗	Both
SAVIA [24]	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	Both
[22]	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	Man
TaxTajweez [42]	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	Both
MVRAG [40]	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	Both
[50]	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	Both
LJP [23]	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto

Australian Legal Corpus. This dataset allows LLMs to enhance their skills when answering legal questions. The LLeQA dataset [21] was created to help develop models that can provide in-depth responses to legal questions in French. This dataset comprises 1,868 legal questions explained by experts with detailed answers based on applicable legal provisions sourced from a collection of 27,942 statutory articles. The LLeQA dataset improves on previous work by adding new kinds of annotations, e.g., a comprehensive taxonomy of questions, jurisdiction information, and specific references at the paragraph level, which makes it a versatile resource for progressing research in LQA and other related legal activities.

Legal IR datasets are critical for evaluating the retrieval precision of legal RAG systems. For example, the Chatlaw Legal Dataset [25] contains about 4 million data samples in 10 main categories and 44 minor categories. This dataset includes different legal areas, e.g., case classification, statute prediction, and legal document drafting, as well as specialized tasks, e.g., public opinion analysis and named entity recognition. This variety guarantees the thorough inclusion of legal processing assignments. The Case Law Evaluation and Retrieval Corpus [43] is the main dataset created from digitized case law retrieved from the Caselaw Access Project by Harvard Law School. This platform contains more than 1.84 million federal case documents and was created for

IR and RAG tasks.In addition, the Chat-Eur-Lex dataset was created specifically for the Chat-EUR-Lex project [41] to improve the accessibility of European legal information using chat-based LLMs and RAG. The EUR-Lex repository contains approximately 37,000 legal acts in English and Italian, which are divided into approximately 371,000 texts or ''chunks'' to improve search results. Note that this dataset does not include documents without XML or HTML data and corrections, which guarantees both quality and significance. The main goal is to help create a conversational interface that offers simplified explanations of complicated legal documents and allows for customized interactions for users requiring legal information. The specialized LeCaRDv2 dataset [40] is uniquely curated for legal case retrieval and is known for its thorough selection of legal cases and careful methodology. It functions as a standard to assess legal retrieval systems, and it covers various legal topics and situations. This dataset contains in-depth case descriptions and is organized to help test retrieval models, especially in complex and uncommon legal cases, ultimately improving the functioning and comprehension of legal IR systems.

LegalBench-RAG [44] is a comprehensive benchmark that was constructed using the four primary datasets. The ContractNLI dataset focuses on NDA-related documents and contains 946 entries. The Contract Understanding Atticus Dataset includes private contracts and has a total of 4,042 entries. The Mergers and Acquisitions Understanding Dataset (MAUD) comprises M&A documents from public companies, with a total of 1,676 entries. Finally, the Privacy QA dataset comprises the privacy policies of consumer applications with a total of 194 entries. In total, these datasets contribute to a robust corpus of legal documents, amounting to approximately 80 million characters across 714 documents, and they form the basis for the 6,889 question–answer pairs that constitute the LegalBench-RAG benchmark.

While the datasets mentioned above all contribute to legal RAG applications, they differ in terms of structure, purpose, and impact on RAG performance. The following comparisons highlight key distinctions:

Legal Question-Answering vs. Case Law Retrieval: Datasets like JEC-QA, LLeQA, and BorderLegal-QA focus on question-answering tasks, making them valuable for improving the precision of RAG systems in legal inquiries. In contrast, datasets such as CJRC, Case Law Evaluation and Retrieval Corpus, and LeCaRDv2 focus on case law retrieval, enhancing RAG's ability to fetch relevant case precedents.

Structured vs. Unstructured Legal Texts: The Chatlaw Legal Dataset and LegalBench-RAG incorporate structured annotations, making them useful for legal document classification and knowledge extraction. On the other hand, CAIL2020, CAIL2021, and Chat-Eur-Lex deal with unstructured legal documents, requiring RAG models to improve document chunking and summarization techniques.

Monolingual vs. Multilingual Data: Datasets such as Chat-Eur-Lex and LLeQA introduce multilingual legal data (English, Italian, and French), helping RAG systems adapt to cross-jurisdictional applications, whereas datasets like JEC-QA and BorderLegal-QA are domain-specific and monolingual.

Regulatory vs. Contractual Focus: The Open Australian Legal QA Dataset and Privacy QA dataset specialize in regulatory compliance, helping RAG models interpret legal policies and statutes. Meanwhile, datasets like ContractNLI and the Mergers and Acquisitions Understanding Dataset emphasize contract analysis, which is useful for legal contract review automation.

These differences determine how effectively a RAG system performs specific legal tasks. The choice of dataset affects model interpretability, retrieval accuracy, and domain adaptation, ultimately shaping the development of more robust legal AI applications. Although existing datasets offer a useful basis for RAG-based legal AI, a number of limitations still exist. Most datasets are restricted to English and Chinese, leaving gaps in legal systems that use languages such as Arabic and French. Additionally, certain legal domains, such as international law and regulatory compliance, remain underrepresented, limiting the applicability of current models. Furthermore, existing benchmarks primarily emphasize QA accuracy, while often overlooking crucial aspects such as interpretability and explainability. Addressing these gaps requires expanding datasets to cover diverse legal systems, refining benchmarks to evaluate interpretability, and incorporating human-in-the-loop evaluation methods. Moreover, integrating multiple datasets to develop hybrid models could further enhance precision and contextual understanding in legal AI applications. Table 8 summarizes the details of the compared datasets.

VI. ETHICAL AND PRIVACY CONSIDERATIONS IN LEGAL RAG

When utilizing RAG-based LLMs in the legal field, addressing ethical concerns, including bias, privacy, hallucination, and safety, is crucial. These issues can be resolved by implementing strong data privacy measures, advocating for transparency and accountability, addressing bias, emphasizing human supervision, and promoting human–machine collaboration. However, the analysis performed in this study shows that only a few papers have addressed these concerns, which indicates that there is considerable room for improvement. Table 7 displays whether ethical values, e.g., privacy, safety, robustness, bias, and trust, were considered in the 22 articles reviewed in this study. All definitions of ethical principles are listed in Table 9.

VII. CHALLENGES IN LEGAL RAG

A. COMPUTATIONAL COST AND COMPLEXITY

Many legal RAG methods encounter challenges associated with the computational cost related to the use of parameter-inaccessible LLMs as generators and embeddings utilizing APIs [40], which can make it inefficient to rely on a powerful LLM, e.g., GPT-4. However, the

TABLE 8. Comparison of legal datasets for RAG systems.

Dataset	Task Type	Domain/Application	Size	Structure
BorderLegal-QA	QA	Border legal queries	1,329 QA pairs	QA pairs
JEC-QA	QA	Legal professional exams	26,365 questions	Multiple-choice questions
CJRC	QA	Chinese court records	50,000 QA pairs	Legal case documents
CAIL2020	QA	Legal reasoning	10,000 documents	Legal text with QA pairs
CAIL2021	QA	Multi-segment QA pairs	7,000 QA pairs	Advanced legal reasoning
Open Australian Legal QA	QA	QA-snippet triplets	2,100 triplets	QA with legal snippets
LLeQA	QA	French legal QA	1,868 QA pairs	Expert-annotated legal QA
CLERC	IR	Case law retrieval	1.84M documents	Digitized case law
LeCaRDv2	IR	Legal case retrieval	Legal case retrieval	Curated case law selection
Chatlaw Legal Dataset	RAG	Legal document processing	4M data samples	Various legal tasks
Chat-Eur-Lex	RAG	European law processing	37,000 legal acts	Chunked legal texts
LegalBench-RAG	RAG	Legal document retrieval	80M characters	Various legal tasks

TABLE 9. Metrics and ethical principles definitions.

Metric/Ethical Principle	Definition
Accuracy	The proportion of correctly classified instances out of the total number of instances [61].
Correctness	The response being either an exact match to the expected answer or a relevant statement that accurately conveys the same information.
Completeness	Generated responses that are comprehensive and cover all aspects of the anticipated response.
Faithfulness	The response must derive directly from the provided context.
Fluency	The ability of a RAG system to generate readable and clear text.
Context relevance	The context of the retrieved information is relevant to the query.
Response relevance	How relevant the generated response is for the given query.
Precision	The fraction of relevant instances among the retrieved instances.
Recall	The fraction of relevant instances retrieved from the total number of relevant cases.
MRR	The average of the reciprocal ranks of the first correct response to a set of queries.
MAP	The mean of the average precision scores for each query.
METEOR	Combines precision, recall, and sentence fluency to accurately calculate the similarity between automatically generated and reference responses to evaluate the effectiveness of text generation tasks.
BLEU	Measures the overlap between a generated response and a set of reference responses by focusing on the precision of n-grams.
ROUGE	Counts the number of overlapping units, e.g., n-grams, word sequences, and word pairs, between the generated and reference responses considering recall and precision.
Privacy	Safeguarding sensitive data and ensuring compliance with legal confidentiality requirements.
Safety	Safeguarding sensitive data and ensuring compliance with legal confidentiality requirements [62].
Robustness	System resilience against adversarial inputs, ambiguous queries, or edge cases [63].
Bias	Mitigation of unfair biases in retrieval, generation, or external knowledge [64].
Trust	Ensuring reliability, transparency, and accountability in generated outputs [64].

complexity of using in-house embedding storage and LLMs is another problem that may hinder the use of open-source solutions [40], [42], [44], [45], [52]. In retrieval, the computational complexity of multiperspective retrieval can pose challenges for real-time applications in specific scenarios [40]. However, leveraging different techniques, e.g., caching embeddings, can reduce redundant computation and API costs [20]. HyPA–RAG [49] integrates an adaptive retrieval process to minimize unnecessary token usage and computational cost. Generally, a well-established RAG pipeline can improve latency by integrating precomputed and optimized retrieval, reducing the reliance on expensive API calls [41].

B. NO RESPONSE AND HALLUCINATION

Based on the failure points (FP) of the RAG systems presented in the literature [65], legal RAG approaches have addressed most of these FPs. These FPs could lead to one of two challenges, i.e., no response and/or a hallucinated response. This subsection discusses how the proposed legal RAG methods address these challenges.

TABLE 10. Failure points and corresponding rag methods, techniques, and processes.

RAG Failure Point	Methods/Techniques to Overcome	RAG process	Legal RAG method
FP1: Missing Content	Knowledge graph	Retrieval	Chatlaw [25], HyPA-RAG [49], PRO [45]
	Prompting with cited cases	Augmentation	CLERC [43]
	Large dataset	Retrieval source	Chat-EUR-Lex [41]
	Training LLM	Generation	IEM [19]
	Real-time data source	Retrieval source	[47]
	Eval-RAG	After Generation	Eval-RAG [48]
FP2: Missed the Top K	Euclidean norm matching	Retrieval	IEM [19]
	Recursive Character Text Splitter (RCTS)	Chunking	LegalBench-RAG [44]
	Sentence-BERT and the FAISS library	Retrieval	SAVIA [24]
FP3: Not in Context	Query expansion	Retrieval	CaseGPT [39]
	Similarity knowledge containers	Retrieval	CBR-RAG [20]
	Query complexity classifier	Retrieval	HyPA-RAG [49]
	Query rewriting	Retrieval	MVRAG [40]
	Knowledge graph	Retrieval	[22]
	Re-rank & Knowledge graph	Retrieval	PRO [45]
FP4: Not Extracted	Scoring Spans, AMR graph, Confidence Calculation	Retrieval	DRAG-BILQA [18]
	Few-shot prompting	Augmentation	LJP [23]
	Iterative refinement	Generation	CaseGPT [39]
	LLM Fine-tuning	Generation	LLeQA [21]
	Faiss vector store	Retrieval	TaxTajweez [42]
FP6: Incorrect Specificity	Prompt tuning	Augmentation	Chat-EUR-Lex [41]
	NDCG metric	After Generation	[51]
FP7: Incomplete	Knowledge graph	Retrieval	HyPA-RAG [49]
	Structured prompt	Augmentation	PRO [45]

The challenge scale is shown in Fig. 7. Here, the FP reflects a hallucinated or no-response challenge. The worst-case scenario of the RAG system is to generate a hallucinated response, and the best-case scenario of an FP for the RAG system is to generate an incomplete answer, as explained in FP7 (Incomplete). Note that the no-response scenario is preferred over the hallucinated response [66]. Most RAG prompts include snippets like "If you don't know the answer, just say that you don't know, don't try to make up an answer." [46]. The challenge scale shown in Fig. 7 can serve as a foundation for developing new evaluation metrics by combining the FP with a degree of hallucination to measure the dependency on external knowledge. In other words, each FP can be measured in a given RAG system. The RAGAs evaluation metric [26] (Section IV) was developed to evaluate the faithfulness, the relevance of the answer, and the relevance to the context of the generated responses. We suggest including the presented FPs [65] when measuring the degree of hallucination and faithfulness. The generated responses should be evaluated against each FP; thus, the overall score is calculated. This evaluation can be examined on the retrieval module and/or the generation module (i.e., the LLM).

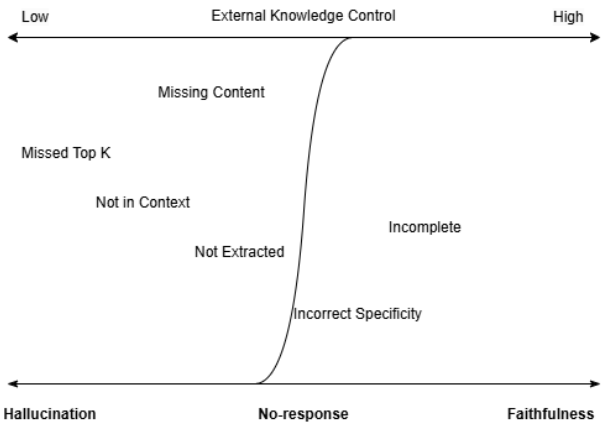

FIGURE 7. Challenge scale and associated FP of the RAG systems [65]. The stronger the dependence on external knowledge in generating responses, the fewer FPs occur and the higher the faithfulness of the generated responses.

1) NO RESPONSE

The most common challenge for a poor RAG system is the inability to generate a response [65] due to a limited number of document chunks passed to the LLM to generate

the response based on the matching algorithm calculated the similarity between the original query and the document chunks in the vector database. This problem can be caused by various factors, including an incomplete dataset, a flawed chunking method, a poor embedding model, a nonrelated matching algorithm, a poor augmentation mechanism, a weak generator, or an inaccurate prompt. Although the no-response scenario indicates that the RAG system cannot produce a response, it provides more robust control over external knowledge compared with scenarios that involve hallucinated responses, as shown in Fig. 7.

2) HALLUCINATION

RAG has been introduced to overcome the hallucination problems of LLMs [14], [67] and RAG itself can be employed to evaluate the performance of LLMs [48]; however, RAG systems are still subject to providing hallucinated responses [65]. For example, a previous study found that GPT-4o generated hallucinated responses [43], and used a prompt with cited cases to overcome this challenge. In another study, the LegalBench dataset was used to assess the legal reasoning capabilities of LLMs in the RAG pipeline to overcome the hallucination problem of RAG systems [44].

Additional methods and techniques implemented in papers reviewed herein to overcome these challenges are listed in Table 10.

C. COMPLEX QUERY HANDLING

RAG systems may struggle with ambiguous, multi-hop, or vague queries, reducing accuracy in complex reasoning tasks. For example, in [46], the RAG system struggles with Q9 and Q12, which require deeper contextual understanding. CASEGPT [39] showed limitations in addressing unprecedented cases. Including multiple retrieved cases (neighbors) for context in CBR-RAG [20] poses challenges to maintain prompt coherence, affecting the quality of the generated output. In Chatlaw [25], the system struggles with diverse user inputs, mainly when users provide incomplete, deceptive, or misleading information, which can lead to incorrect answers. LegalBench-RAG [44] struggles with tasks that require multi-hop reasoning or handling technical legal jargon, particularly in datasets like MAUD. For PRO [45], multi-hop reasoning or combining information from multiple retrieved documents remains a challenge, leading to incomplete or incorrect responses. TaxTajweez [42] and [51] RAG systems struggle with complex or ambiguous user queries, leading to less relevant or incomplete retrieval results.

D. DEPENDENCE ON RETRIEVAL ACCURACY

As clarified earlier, RAG systems rely heavily on retrieval precision. Errors or irrelevant document fragments affect the outputs. For example, the retrieval precision of CLERC [43] is reduced due to repeated legal terms and irrelevant words in legal documents that mislead retrieval models. In addition, setting a static confidence threshold in DRAG-BILQA [18] may not be adaptable to all questions or complex queries.

Although incorporating KG into the retrieval can improve the accuracy of the retrieval, incomplete KG can result in missing critical information and key relationships within legal texts [25], [49].

In LegalBench [44], general-purpose re-rankers, such as Cohere's model, perform poorly on specialized legal texts due to a lack of domain adaptation. On the other hand, increasing recall improves the chance of retrieving relevant snippets but introduces more noise, reducing precision.

In general, incorporating external knowledge sources into RAG while maintaining coherence and relevance remains complex and challenging.

E. EVALUATION METRICS LIMITATIONS

Current metrics (e.g., BLEU, METEOR, precision, recall) may not fully assess factual correctness and semantic quality. For example, ROUGE struggles to accurately measure the factual correctness and semantic quality of long-form responses [21]. Besides, the METEOR scores remained low due to the complexity and length of the legal content, limiting effective evaluation [18]. However, as revealed in Section IV, there is no standardized evaluation method for RAG systems, and various frameworks utilize different metrics to evaluate RAG systems.

VIII. RESEARCH DIRECTIONS IN LEGAL RAG

Recently, RAG has been applied in the legal domain, and it has exhibited promising benefits and outcomes. This section highlights the potential research directions inspired by the articles reviewed herein.

A. EXPANDING LEGAL DOMAINS

The scope of future legal RAG research should be extended to cover a wider range of legal domains. Legal acts are involved in everyday activities; thus, developing reliable and accurate LLM-driven systems can benefit individuals and law practitioners. Most of the surveyed studies have covered many specific legal domains focusing on narrow subdomains, such as contract analysis and case law retrieval. While other papers, such as Chatlaw [25] and CBR-RAG [20] cover broader legal domains. Legal RAG systems often struggle with cross-jurisdictional generalization due to differences in legal systems, terminologies, and practices. Future research should explore methods to enhance the adaptability of RAG systems across jurisdictions. For example, techniques like multi-task learning (e.g., as demonstrated in LEGAL-BERT [68]) could be extended to RAG systems to improve their performance in diverse legal environments. The open challenge here is how to scale RAG systems to handle the complex interdependencies in multi-domain legal scenarios, and how to generalize domain-specific models without sacrificing performance in individual domains.

B. DEVELOPING AND ENHANCING LEGAL DATASETS

As discussed in Section VII, one of the most common root causes of failure in RAG systems is noise in used datasets.

TABLE 11. Challenges and research directions in legal RAG systems.

Category	Issue	Description
Challenges	Computational cost and complexity	Retrieval and generation processes require significant and costly computations.
	No response and hallucination	RAG systems may generate fabricated or inaccurate outputs when retrieval lacks precision or relevance.
	Complex query handling	RAG may struggle with ambiguous, multi-hop, or vague queries, reducing accuracy in complex reasoning tasks.
	Dependence on retrieval accuracy	RAG performance relies heavily on retrieval precision; errors or irrelevant documents negatively impact outputs.
	Evaluation metrics limitations	Current evaluation metrics may fail to fully assess factual correctness and semantic quality.
Research Directions	Expanding legal domains	Developing reliable and accurate RAG applications for a wider range of legal domains.
	Developing and enhancing legal datasets	Developing robust open-source datasets in different legal domains.
	Multilingual legal RAG	Efficient methods and techniques for multilingual legal corpora.
	Multidimensional approach in legal tech	Multidimensional approaches enhance retrieval and generation capabilities.
	Evaluation metrics	Developing a standardized evaluation method to assess the performance of RAG systems.
	Reinforcement learning to optimize legal RAG applications	Employment of reinforcement learning to optimize RAG applications in the legal domain.

To date, three benchmark datasets have been developed in different legal domains to develop and evaluate legal RAG systems [21], [43], [44]. Therefore, future research should focus on developing robust open-source datasets in different legal domains. As explained earlier in Section V, most of the datasets used in the surveyed papers were developed for the experiment and the specific tasks. Developing a benchmark dataset for the legal domain such as the ALQA dataset which is used in CBR-RAG [20] can help in evaluating the RAG systems in the legal domains. For example, developing huge datasets that can enhance the RAG models as well the re-training of the LLMs [69], [70], [71]. Datasets like LexGLUE [72] have set a precedent for benchmarking legal NLP tasks, but more specialized datasets are needed for RAG systems. These datasets should include annotated legal texts, case law, and statutory provisions to enable fine-tuning and evaluation of RAG models in specific legal contexts.

C. MULTILINGUAL LEGAL RAG

Another promising research direction for non-English researchers is to take advantage of the available non-English legal knowledge in legal RAG systems. Researchers in this field may study efficient methods and techniques for multilingual legal corpora, and legal technology researchers can leverage the recent state-of-the-art methods in this domain [73], [74]. For example, handling code-switching in non-Latin scripts, addressing fluency errors, improving document comprehension, and minimizing irrelevant retrievals for multilingual RAG models are promising researches for multilingual legal RAG [73]. Recent advancements in mBERT and XLM-R [75] provide opportunities to train legal RAG systems on multilingual corpora efficiently. Additionally, datasets such as MultiEURLex [76], which cover EU legal documents in multiple languages, could serve as a foundation for developing multilingual RAG systems.

D. MULTIDIMENSIONAL APPROACH IN LEGAL TECH

Integrating RAG with knowledge graphs, fine-tuning processes, and prompt engineering, as shown in Fig. 2, is becoming a prominent approach in legal technology. This multidimensional approach is expected to enhance retrieval and generation capabilities, and future research and case studies are expected to further enrich the field with more interpretable and reliable LLM-driven applications. For example, the integration of ConceptNet [77] with RAG systems could help bridge the gap between structured and unstructured legal knowledge. Case studies on prompt engineering for specific legal tasks (e.g., drafting legal briefs) [78] could also guide future research in this direction.

E. EVALUATION METRICS

Focusing on developing a standardized evaluation method to assess the performance of RAG systems is a promising research field as long as RAG is in the early stages. Researchers in this field can leverage the latest metrics and approaches [26] as discussed in Sections IV and VII. Developing a comprehensive evaluation framework for legal RAG systems is essential to evaluate their reliability and performance. Building on works like Kwiatkowski et al. [79], who developed human evaluation methods for natural language systems, researchers could devise legal-specific metrics that assess the factual accuracy of generated responses and citation quality in retrieved case law. Open challenges include how to measure the interpretability of RAG systems in high-stakes legal settings, and what new metrics can evaluate the ethical implications of RAG outputs.

F. REINFORCEMENT LEARNING TO OPTIMIZE LEGAL RAG APPLICATIONS

Transformer models are the most widely used approach for embedding and generation in legal RAG systems,

as demonstrated by the literature survey performed in this study. However, it is expected that reinforcement learning can be employed to optimize legal RAG [80], [81] in retrieval and generation modules for different types of real-world applications. For example, methods like Reinforcement Learning from Human Feedback (RLHF) (e.g., as used in OpenAI's GPT models) could be adapted for Legal RAG systems to improve their performance in real-world legal applications. This approach would allow the system to learn from interactions with legal professionals, ensuring it retrieves and generates more relevant and accurate outputs. In addition, the reward-based optimization approach [82] could be applied to fine-tune models for specific legal tasks (e.g., identifying precedents in case law). Building on works like Ziegler et al. [83], which optimized language models for specific human preferences, RL could enable legal RAG systems to better align with legal practitioners' needs. Recent advances in RL for enhancing reasoning capabilities in LLMs, such as those demonstrated in DeepSeek-R1 [84], highlight its potential to align model outputs with domain-specific objectives.

IX. CONCLUSION

This paper has presented an overview of the utilization of RAG in the legal domain. We have covered and analyzed all methods, techniques, and stages of legal RAG. The analysis presented in this paper provides insights into embedding, retrieval, augmentation, and generation techniques. In addition, we have thoroughly investigated IR, as it is the backbone of RAG, and we explained the different evaluation metrics used to assess RAG systems. Furthermore, we have proposed a challenge scale to control the hallucination results of RAG, which is expected to be an initial foundation for developing a new evaluation method.

REFERENCES

[1] K. Mania, "Legal technology: Assessment of the legal tech industry's potential," *J. Knowl. Economy*, vol. 14, no. 2, pp. 595–619, Jun. 2023.

[2] J. B. Rajendra, "Disruptive technologies and the legal profession," *Int. J. Law*, vol. 6, no. 5, pp. 271–280, Jan. 2020.

[3] S. Sharma, S. Gamoura, D. Prasad, and A. Aneja, "Emerging legal informatics towards legal innovation: Current status and future challenges and opportunities," *Legal Inf. Manage.*, vol. 21, nos. 3–4, pp. 218–235, Dec. 2021.

[4] R. Dale, "Law and word order: NLP in legal tech," *Natural Lang. Eng.*, vol. 25, no. 1, pp. 211–217, Jan. 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[6] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[7] X. Chen, J. Zheng, C. Li, B. Wu, H. Wu, and J. Montewka, "Maritime traffic situation awareness analysis via high-fidelity ship imaging trajectory," *Multimedia Tools Appl.*, vol. 83, no. 16, pp. 48907–48923, Nov. 2023.

[8] X. Chen, H. Wu, B. Han, W. Liu, J. Montewka, and R. W. Liu, "Orientation-aware ship detection via a rotation feature decoupling supported deep learning approach," *Eng. Appl. Artif. Intell.*, vol. 125, Oct. 2023, Art. no. 106686.

[9] Z. Zhang, J. Xiong, Z. Zhao, F. Wang, Y. Zeng, B. Zhao, and L. Ke, "An approach of dynamic response analysis of nonlinear structures based on least square Volterra kernel function identification," *Transp. Saf. Environ.*, vol. 5, no. 2, p. 46, Mar. 2023.

[10] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, "Large legal fictions: Profiling legal hallucinations in large language models," *J. Legal Anal.*, vol. 16, no. 1, pp. 64–93, Jan. 2024.

[11] F. Yu, L. Quartey, and F. Schilder, "Exploring the effectiveness of prompt engineering for legal reasoning tasks," in *Proc. Findings Assoc. Comput. Linguistics (ACL)*, Toronto, ON, Canada, 2023, pp. 13582–13596.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 27730–27744.

[13] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 10088–10115.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 9459–9474.

[15] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? Comparing knowledge injection in LLMs," 2023, *arXiv:2312.05934*.

[16] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine tuning vs. retrieval augmented generation for less popular knowledge," 2024, *arXiv:2403.01432*.

[17] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions," 2024, *arXiv:2410.12837*.

[18] Y. Zhang, D. Li, G. Peng, S. Guo, Y. Dou, and R. Yi, "A dynamic retrieval-augmented generation framework for border inspection legal question answering," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Hohhot, China, Aug. 2024, pp. 372–376.

[19] A. Nikolakopoulos, S. Evangelatos, E. Veroni, K. Chasapas, N. Gousetis, A. Apostolaras, C. D. Nikolopoulos, and T. Korakis, "Large language models in modern forensic investigations: Harnessing the power of generative artificial intelligence in crime resolution and suspect identification," in *Proc. 5th Int. Conf. Electron. Eng., Inf. Technol. Educ. (EEITE)*, Chania, Greece, May 2024, pp. 1–5.

[20] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, and B. Fleisch, "CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering," 2024, *arXiv:2404.04302*.

[21] A. Louis, G. Van Dijck, and G. Spanakis, "Interpretable long-form legal question answering with retrieval-augmented large language models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 22266–22275.

[22] R. Venkatakrishnan, E. Tanyildizi, and M. A. Canbaz, "Semantic interlinking of immigration data using LLMs for knowledge graph construction," in *Proc. ACM Web Conf. Companion*. Singapore: Springer, May 2024, pp. 605–608.

[23] R. Shui, Y. Cao, X. Wang, and T.-S. Chua, "A comprehensive evaluation of large language models on legal judgment prediction," in *Proc. Findings Assoc. Comput. Linguistics*, Singapore, 2023, pp. 7337–7348.

[24] M. Visciarelli, G. Guidi, L. Morselli, D. Brandoni, G. Fiameni, L. Monti, S. Bianchini, and C. Tommasi, "SAVIA: Artificial intelligence in support of the lawmaking process," in *Proc. 4th Nat. Conf. Artif. Intell.* Naples, Italy: CINI, May 2024.

[25] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan, "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2023, *arXiv:2306.16092*.

[26] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," 2023, *arXiv:2309.15217*.

[27] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for AI-generated content: A survey," 2024, *arXiv:2402.19473*.

[28] M. Mitra and B. B. Chaudhuri, "Information retrieval from documents: A survey," *Inf. Retr.*, vol. 2, pp. 141–163, Apr. 2000.

[29] S. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *Proc. SIGIR*, B. W. Croft and C. J. Van Rijsbergen, Eds., London, U.K.: Springer, Aug. 1994, pp. 232–241.

[30] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.
[31] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," 2020, *arXiv:2004.04906*.
[32] H.-T. Nguyen, M.-K. Phi, X.-B. Ngo, V. Tran, L.-M. Nguyen, and M.-P. Tu, "Attentive deep neural networks for legal document retrieval," 2022, *arXiv:2212.13899*.
[33] M. Grohe, "word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data," in *Proc. 39th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, Jun. 2020, pp. 1–16.
[34] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2023, *arXiv:2312.10997*.
[35] C.-M. Chan, C. Xu, R. Yuan, H. Luo, W. Xue, Y. Guo, and J. Fu, "RQ-RAG: Learning to refine queries for retrieval augmented generation," 2024, *arXiv:2404.00610*.
[36] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on RAG meeting LLMs: Towards retrieval-augmented large language models," 2024, *arXiv:2405.06211*.
[37] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, and X. Huang, "Searching for best practices in retrieval-augmented generation," 2024, *arXiv:2407.01219*.
[38] E. Mollard, A. Patel, L. Pham, and R. Trachtenberg, "Improving retrieval augmented generation," Lab. Phys. Sci. (LPS), Univ. Maryland, College Park, MD, USA, Tech. Rep., Aug. 2024.
[39] R. Yang, "CaseGPT: A case reasoning framework based on language models and retrieval-augmented generation," 2024, *arXiv:2407.07913*.
[40] G. Chen, W. Yu, and L. Sha, "Unlocking multi-view insights in knowledge-dense retrieval-augmented generation," 2024, *arXiv:2404.12879*.
[41] M. Cherubini, F. Romano, A. Bolioli, L. De, and M. Sangermano, "Improving the accessibility of EU laws: The Chat-EUR-Lex project," in *Proc. 4th Nat. Conf. Artif. Intell.* Naples, Italy: CINI, May 2024.
[42] M. A. Habib, S. M. Amin, M. Oqba, S. Jaipal, M. J. Khan, and A. Samad, "TaxTajweez: A large language model-based chatbot for income tax information in Pakistan using retrieval augmented generation (RAG)," in *Proc. Int. FLAIRS Conf.*, vol. 37, May 2024, pp. 1–12.
[43] A. B. Hou, O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, and B. Van Durme, "CLERC: A dataset for legal case retrieval and retrieval-augmented analysis generation," 2024, *arXiv:2406.17186*.
[44] N. Pipitone and G. H. Alami, "LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain," 2024, *arXiv:2408.10343*.
[45] J.-M. Chu, H.-C. Lo, J. Hsiang, and C.-C. Cho, "Patent response system optimised for faithfulness: Procedural knowledge embodiment with knowledge graph and retrieval augmented generation," in *Proc. 1st Workshop Towards Knowledgeable Lang. Models (KnowLLM)*, Bangkok, Thailand, 2024, pp. 146–155.
[46] M. E. Mamalis, E. Kalampokis, F. Fitsilis, G. Theodorakopoulos, and K. Tarabanis, "A large language model agent based legal assistant for governance applications," in *Proc. Int. Conf. Electron. Government*, Jan. 2024, pp. 286–301.
[47] I. Bošković and V. Tabaš, "Proposal for enhancing legal advisory services in the Montenegrin banking sector with artificial intelligence," in *Proc. 28th Int. Conf. Inf. Technol. (IT)*, Zabljak, Montenegro, Feb. 2024, pp. 1–6.
[48] C. Ryu, S. Lee, S. Pang, C. Choi, H. Choi, M. Min, and J.-Y. Sohn, "Retrieval-based evaluation for LLMs: A case study in Korean legal QA," in *Proc. Natural Legal Lang. Process. Workshop*, Singapore, 2023, pp. 132–137.
[49] R. Kalra, Z. Wu, A. Gulley, A. Hilliard, X. Guan, A. Koshiyama, and P. Treleaven, "HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications," 2024, *arXiv:2409.09046*.
[50] T.-H.-G. Vu and X.-B. Hoang, "User privacy risk analysis within website privacy policies," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, Da Nang, Vietnam, Aug. 2024, pp. 1–6.
[51] R. Nai, E. Sulis, I. Fatima, and R. Meo, "Large language models and recommendation systems: A proof-of-concept study on public procurements," in *Natural Language Processing and Information Systems* (Lecture Notes in Computer Science), vol. 14763. Cham, Switzerland: Springer, 2024, pp. 280–290.
[52] A. Chouhan and M. Gertz, "LexDrafter: Terminology drafting for legislative documents using retrieval augmented generation," 2024, *arXiv:2403.16295*.
[53] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," 2023, *arXiv:2401.00396*.
[54] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," 2024, *arXiv:2404.16130*.
[55] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 2, pp. 1–37, Feb. 2021.
[56] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, and C. J. Xue, "Retrieval-augmented generation for natural language processing: A survey," 2024, *arXiv:2407.13193*.
[57] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," 2024, *arXiv:2405.07437*.
[58] (Jun. 2024). *The Ultimate Guide to Evaluate RAG System Components: What You Need to Know*. Accessed: Nov. 11, 2024. [Online]. Available: https://myscale.com/blog/ultimate-guide-to-evaluate-rag-system/
[59] L. M. Amugongo, P. Mascheroni, S. G. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *Preprints*, Jul. 2024, doi: 10.20944/preprints202407.0876.v1.
[60] S. Sivasothy, S. Barnett, S. Kurniawan, Z. Rasool, and R. Vasa, "RAGProbe: An automated approach for evaluating RAG applications," 2024, *arXiv:2409.19019*.
[61] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012.
[62] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, and J. Tang, "The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)," 2024, *arXiv:2402.16893*.
[63] W. Li, J. Li, R. Ramos, R. Tang, and D. Elliott, "Understanding retrieval robustness for retrieval-augmented image captioning," 2024, *arXiv:2406.02265*.
[64] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. S. Yu, "Trustworthiness in retrieval-augmented generation systems: A survey," 2024, *arXiv:2409.10102*.
[65] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proc. IEEE/ACM 3rd Int. Conf. AI Eng.-Softw. Eng. AI*, Apr. 2024, pp. 194–199.
[66] W. Yu, H. Zhang, X. Pan, K. Ma, H. Wang, and D. Yu, "Chain-of-note: Enhancing robustness in retrieval-augmented language models," 2023, *arXiv:2311.09210*.
[67] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," 2021, *arXiv:2104.07567*.
[68] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," 2020, *arXiv:2010.02559*.
[69] P. Henderson, M. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho, "Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 29217–29234.
[70] J. Niklaus, V. Matoshi, M. Stürmer, I. Chalkidis, and D. E. Ho, "MultiLegalPile: A 689GB multilingual legal corpus," 2023, *arXiv:2306.02069*.
[71] M. Ostendorff, T. Blume, and S. Ostendorff, "Towards an open platform for legal information," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries*, Aug. 2020, pp. 385–388.
[72] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," 2021, *arXiv:2110.00976*.
[73] N. Chirkova, D. Rau, H. Déjean, T. Formal, S. Clinchant, and V. Nikoulina, "Retrieval-augmented generation in multilingual settings," 2024, *arXiv:2407.01463*.
[74] S. R. El-Beltagy and M. A. Abdallah, "Exploring retrieval augmented generation in Arabic," *Proc. Comput. Sci.*, vol. 244, pp. 296–307, May 2024.
[75] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[76] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," 2021, *arXiv:2109.00904*.
[77] R. E. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, Feb. 2017, pp. 1–7.
[78] J. Lee, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 24824–24837.
[79] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. V. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 7, Aug. 2019, pp. 453–466.
[80] M. Kulkarni, P. Tangarajan, K. Kim, and A. Trivedi, "Reinforcement learning for optimizing RAG for domain chatbots," 2024, *arXiv:2401.06800*.
[81] Z. Wang, S. Xian Teo, J. Ouyang, Y. Xu, and W. Shi, "M-RAG: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions," 2024, *arXiv:2405.16420*.
[82] Y. Wu, E. Mansimov, S. M. Liao, R. Grosse, and J. Ba, "Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 1–8.
[83] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2019, *arXiv:1909.08593*.
[84] DeepSeek-AI et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," 2025, *arXiv:2501.12948*.

MAHD HINDI received the B.Sc. degree in information systems technology from Abu Dhabi University, Abu Dhabi, United Arab Emirates. He is currently pursuing the M.Sc. degree with the College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates. His current research interests include LLMs and LLM-driven solutions.

LINDA MOHAMMED received the B.Sc. degree in electrical and electronic engineering (electronic systems software engineering) from the University of Khartoum, Sudan, in 2020. She is currently pursuing the M.Sc. degree in software engineering with United Arab Emirates University, United Arab Emirates. Her current research interests include AI, machine learning, and data science.

OMMAMA MAAZ received the B.Sc. degree in computer engineering from the University of Sharjah, Sharjah, United Arab Emirates, in 2022. She is currently pursuing the M.Sc. degree with the College of Information Technology, United Arab Emirates University, United Arab Emirates. Her current research interest includes the IoT systems.

ABDULMALIK ALWARAFY (Member, IEEE) received the Ph.D. degree in computer science and engineering from Hamad Bin Khalifa University, Doha, Qatar. He is currently an Assistant Professor with the College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates. His current research interests include the application of artificial intelligence techniques across various domains, including wireless and the IoT networks, as well as edge and cloud computing. He is a member of the IEEE Communications Society. He served on the technical program committees of many international conferences. In addition, he has been a reviewer of several international journals and conferences.