Accurate AI Assistance in Contract Law Using Retrieval-Augmented Generation to Advance Legal Technology

Youssra Amazou, Faouzi Tayalati, Houssam Mensouri, Abdellah Azmani, Monir Azmani
Intelligent Automation and Biomed Genomics Laboratory-FST of Tangier,
Abdelmalek Essaadi University, Tetouan, Morocco

Abstract—**Understanding legal documentation is a complex task due to its inherent subtleties and constant changes. This article explores the use of artificial intelligence-driven chatbots, enhanced by retrieval-augmented generation (RAG) techniques, to address these challenges. RAG integrates external knowledge into generative models, enabling the delivery of accurate and contextually relevant legal responses. Our study focuses on the development of a semantic legal chatbot designed to interact with contract law data through an intuitive interface. This AI Lawyer functions like a professional lawyer, providing expert answers in property law. Users can pose questions in multiple languages, such as English and French, and the chatbot delivers relevant responses based on integrated official documents. The system distinguishes itself by effectively avoiding LLM hallucinations, relying solely on reliable and up-to-date legal data. Additionally, we emphasize the potential of chatbots based on LLMs and RAG to enhance legal understanding, reduce the risk of misinformation, and assist in drafting legally compliant contracts. The system is also adaptable to various countries through the modification of its legal databases, allowing for international application.**

Keywords—AI Lawyer; contract law; legal technology; Retrieval-Augmented Generation (RAG); Large Language Models (LLMs); GPT; chatbots

I. Introduction

Legal services play an essential role in ensuring compliance with regulations and facilitating access to public information [1]. By providing assistance in legal matters, these services help to improve the efficiency, accuracy and accessibility of the legal field. However, despite their importance, challenges persist in understanding complex legal documentation and creating accurate contracts, not least because of the prevalence of misinformation and the complexity of litigation. Integrating AI technologies into legal services offers significant benefits, such as improved efficiency and reduced costs [2]. By standardizing tasks, AI improves the accuracy and consistency of routine tasks, streamlining processes and boosting productivity [3]. In addition, AI technologies are making legal services increasingly accessible to the general public, facilitating access to legal information and services. This increased accessibility contributes to better access to justice, legal transparency and efficient dispute resolution.

In the current context, where understanding legal documents is critical and contract drafting requires precise expertise, exploring innovative solutions based on artificial intelligence has become essential. These solutions aim to streamline processes [4], minimize risks associated with misinformation and litigation, and provide assistance without requiring specialized legal skills. This article aims to present an intelligent chatbot system specifically designed for the legal domain, particularly to help citizens understand contract law without the need for a lawyer.

Chatbots, due to their ability to provide detailed and coherent responses in conversational dialogue [4], are promising tools to address these challenges [5], [6], [7]. This chatbot is designed to enhance the understanding of legal documents, especially contract management law, and facilitate contract drafting by establishing connections between different elements. By integrating external knowledge, it helps minimize discrepancies with current regulations, detect potential errors, and avoid future complications. Proactively, this assistance system enables non-expert users in the legal field to draft contracts by guiding them on key information to include, while ensuring compliance through legal references and reliable databases. Additionally, its ability to provide precise answers to complex questions makes it a valuable tool for legal clarification [8], [9], [10]. To fully harness advancements in artificial intelligence and natural language processing to improve the efficiency and accessibility of our system [11], [12], we rely on large language models (LLMs). These models have achieved significant progress in recent years, demonstrating their ability to generate coherent text on a wide range of topics [13]. A striking example of this advancement is the emergence of the GPT-4 model [14], which has exhibited rudimentary reasoning capabilities, marking a significant leap in the field. However, LLMs face a critical limitation, their inability to access specific and relevant information in real time [15]. In specific domains, such as the provisions of a law in a particular country, these models may lack the necessary facts or details, as such information might not be present in their memory. Additionally, another limitation of LLMs mentioned in the text is the issue of "hallucination," where these models generate statements that appear plausible but are factually incorrect. Despite their impressive fluency and conversational capabilities, this suggests that LLMs can still produce inaccurate or misleading information [16]. This shortcoming poses a significant challenge for practical applications.

To address this challenge, the concept of Retrieval-Augmented Generation frameworks has emerged as a promising solution. These systems enhance LLMs by integrating them with vector databases containing relevant information [17]. When a query is presented by the user, these systems dynamically retrieve and incorporate pertinent data into the LLM's context. This augmentation strategy has proven highly effective [18], as it addresses the shortcomings of LLMs in processing legal data. The synergy between LLMs and RAG systems provides more robust and context-sensitive applications in this domain [19]. This advancement paves the way for potential improvements to our system, thereby offering users greater practical utility.

The general contributions of this article are as follows:

- The proposal of an innovative chatbot for legal assistance, designed to provide basic legal advice accessible to all. This system is particularly aimed at helping citizens understand contract laws without requiring prior legal knowledge. By simplifying access to legal information, it promotes greater democratization of justice.

- The implementation of an advanced technical architecture based on Retrieval-Augmented Generation (RAG). This technology combines document search capabilities with large-scale language models, enabling a deeper contextual understanding of legal texts and ensuring greater accuracy in the responses provided.

- A significant improvement in accessibility and transparency in the legal field. By making complex legal texts more accessible and simplifying access to information, the chatbot helps enhance citizens' understanding and promotes greater transparency in legal processes.

- A remarkable ability to reduce misinformation. Unlike traditional language models, this system relies solely on verified and regularly updated data, providing accurate and relevant responses. This approach overcomes the limitations of traditional AI systems by minimizing the risks associated with incorrect or speculative information.

- Dynamic personalization and adaptability to local laws. By modifying integrated databases and legal rules, the system can be easily adapted to different jurisdictions, offering a flexible solution capable of meeting the specific needs of each country.

- The article is structured into four main sections. The first section explores the fundamentals of conversational systems and their role in various applications, detailing how Retrieval-Augmented Generation works. It highlights its key components, including the information retrieval mechanism, contextual content generation, and their integration into interactive systems. The second section analyzes existing work in the field of legal digitalization, using contract law as a case study, while presenting automated assistance systems such as chatbots. The third section describes the adopted methodology, outlining the various stages of development of the proposed system, with a focus on the technical specifications that ensure its performance and efficiency in the context of legal assistance. Finally, the fourth section presents the research findings and discusses their implications, emphasizing the system's contributions to improving assistance tools in complex fields such as law.

II. Scope of Study

A. Chatbot System

A chatbot, also known as a conversational agent or dialogue system, is software designed to simulate human conversation through text or voice interactions. Chatbots are widely used to automate routine tasks, save time, and enhance user experience across various industries [20], [21], [22]. Their functionality relies on predefined rules or advanced AI techniques such as Natural Language Processing (NLP), enabling dynamic and contextually appropriate responses. Integrated into chat platforms, chatbots often serve as virtual assistants, capable of handling both structured tasks and informal conversations within their programmed expertise [23], [24]. Recent developments have highlighted their transformative potential in areas such as customer support, where they improve efficiency and availability; education, where they enable personalized learning experiences; healthcare, where they assist with patient communication and preliminary diagnostics; and entertainment, where they create interactive user experiences. As their applications continue to expand, chatbots are reshaping interactions between humans and technology [21], [25]. The performance and effectiveness of chatbots generally rely on three key elements.

- Natural Language Understanding (NLU) is crucial for interpreting users' messages accurately. This involves analyzing natural language to identify intent, extract relevant information, and understand the context of the conversation [26], [27], ensuring that the chatbot can respond appropriately to user needs.

- Response generation involves providing suitable and contextually relevant answers. This can be achieved using predefined rule-based systems, retrieval models that select existing responses [26], [28], or generative models that create unique and personalized responses tailored to the user's input.

- Managing the context of conversations is essential for maintaining coherence in long interactions. By remembering previous exchanges [27], chatbots can adjust their responses according to the evolving needs of the user, improving the flow and relevance of the conversation.

B. Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is an innovative method that combines the power of large language models with dynamic external knowledge retrieval, directly integrated into the text generation process. This approach overcomes several limitations of LLMs, such as outdated knowledge and hallucinations, by anchoring the generated content on relevant, accurate, and up-to-date information from reliable external sources.

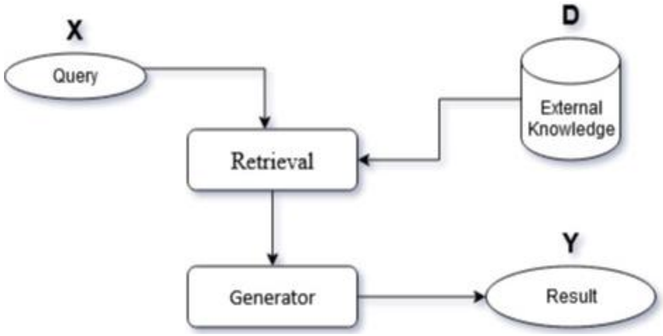


Fig. 1. Retrieval-augmented generation components.

By integrating the robustness of a generative model with the relevance and timeliness of the retrieved data, RAG produces responses that are not only natural and human-like but also contextually appropriate and highly reliable [16], [29]. This fusion of retrieval and generation delivers results that are inaccessible by either component in isolation, positioning RAG as a major advancement in generative models [13], [30]. It allows for the production of high-quality responses, even in less explored fields, while remaining economically advantageous.

The principle of RAG, illustrated in Fig. 1, involves predicting the output y from the input source x, both of which come from a corpus D. Simultaneously, a reference set Z is accessible through data sources. The direct association between a document z ∈ Z and a tuple (x, y) ∈ D is not necessarily known, although it can be established through human annotations [31], [32] or weakly supervised signals [33]. The general framework of RAG includes two main components: (B.1) a document retriever and (B.2) a text generator. The goal of RAG is to train a model that maximizes the probability of y given x and Z. In practice, Z often contains millions of documents, making exhaustive enumeration impossible. Therefore, the first step of RAG is to use document retrievers, such as DPR [29], to narrow the search down to a handful of relevant documents. The retriever takes x and Z as input and produces relevance scores $\{s_1, ..., s_K\}$ for the top K documents $Z = \{z_{(1)}, ..., z_{(K)}\}$. Then, the second step is to use a text generator [34], [35] to produce the desired result y by taking both the input x and the retrieved document set Z as conditions.

1) Information retrieval: Information Retrieval (IR) is a field of information science and computer science that focuses on the representation, storage, and organization of unstructured data to help users find and retrieve relevant information based on their queries [36], [37], [38]. An advanced aspect of IR involves the use of neural document retrievers. These typically use two independent encoders, such as BERT [39], to separately encode the query and the document. They then estimate relevance by calculating a single similarity score between the two encoded representations. For example, in the DPR model [33], the documents Z and contextual queries x are projected into the same dense encoding space. The relevance score q(x, z) for each document z is calculated as the dot product between the document encoding T_z and the query encoding T_x, this allows for efficiently determining which documents are most relevant to a given query, by leveraging advanced natural language processing and machine learning techniques to improve the accuracy and relevance of search results.

2) Information generation: Text generation represents a crucial domain in natural language processing and artificial intelligence, aiming to automatically create readable and coherent texts from existing information [40]. This task, often referred to as Information Generation (IG), relies on advanced machine learning and NLP techniques to produce new content. Text generation models, such as GPT and BERT, have the ability to synthesize data, rephrase information, and generate relevant responses based on the provided contexts. By using deep neural networks, these models analyze vast amounts of textual data to learn the linguistic and contextual structures necessary for text production [41] . This approach enables the creation of varied content, ranging from article summaries to detailed answers to specific questions, continually improving the quality and relevance of the generated texts thanks to advancements in AI and NLP.

III. Related Work

The advancements made in the field of artificial intelligence have led to a notable increase in the development and deployment of chatbots for various tasks [42]. Many research efforts have focused on the creation of chatbots and the study of their applications in different fields, such as business support and education. For example, some studies have addressed the technical advancements in chatbot development to enhance their capabilities and effectiveness [43]. This section of the article examines existing research on the development and application of chatbots, highlighting the growing interest in chatbots and their evaluation in the legal field. The development of AI-powered chatbots for legal advice has garnered considerable attention in recent years. Previous studies have explored the application of chatbots in the legal domain, focusing on various aspects such as the type of chatbots, their capabilities, and their limitations [4], [44]. For example, [45] compared two types of chatbots, generative and intent-based, to provide legal advice specific to Indian laws, evaluating their performance based on factors such as response quality and user experience. Similarly, [46] assessed the capabilities of ChatGPT and Gemini chatbots for contract drafting, highlighting the need for human intervention due to the limitations of chatbots in understanding the specifics of the legal system and linguistic conventions.

Other studies have focused on the design and development of legal chatbots, which aim to automatically converse with users, determine the need for legal advice, grant access to legal rights, bridge the communication gap between clients and lawyers, and generate documents for legal activities [47]. In the same vein, [48] took this concept a step further by proposing a chatbot specifically designed for smart contracts. This chatbot can assist non-technical users in specifying and generating code for smart contracts, thus highlighting the potential of chatbots to facilitate the creation and management of complex legal documents.

These studies highlight the potential of chatbots to provide accurate legal information to users. However, despite the existence of these studies, there is still a need to improve chatbot systems in the legal field. The revolution of generative AI and its growing use in this sector emphasize this necessity. Indeed, human intervention remains essential due to the sensitivity of legal information and to avoid potentially severe consequences. The issues of hallucination in large language models used for chatbots particularly underscore the importance of these improvements. Without proper human intervention, chatbots may provide incorrect or inadequate information, leading to significant negative impacts. Therefore, further research and development are essential to strengthen the reliability and effectiveness of chatbots in the legal field. Many researchers have explored different approaches to improve chatbot capabilities. In this context, several studies have focused on enhancing chatbot systems to address this issue, through the integration of a retrieval-based response generation model, which has outperformed models based solely on retrieval or generation, offering increased fluency, improved contextual relevance, and greater diversity in responses [49]. This technology is Retrieval-Augmented Generation, which involves retrieving texts from a relevant external corpus for the task, and then providing them to the large language model [29], [50]. RAG improves the performance of LLMs by integrating the retrieved texts via cross-attention [29], [51], [52] or by directly inserting the retrieved documents into the prompt. Large language models, advanced systems for natural language processing, are trained on massive datasets to process and generate text [34]. Although they are designed for tasks such as machine translation, summarization, and conversational interactions, RAG models can also be used for information retrieval [13]. Retrieval-augmented generation has demonstrated notable success in several natural language processing tasks requiring deep knowledge, such as answering general knowledge questions accurately, fact-checking with high precision, and answering questions within specific domains [17], [29]. Furthermore, retrieval-augmented generation reduces misinformation often produced by large language models and non-RAG chatbots.

IV. Materials and Methods

In this section, we introduce the AI assistance model tailored for contract law, which represents a groundbreaking advancement in efficiently accessing legal knowledge. This model begins with the collection of legal datasets for contracts. Next, by segmenting the documentation into manageable "chunks" and employing sophisticated techniques such as vector representation for storage and similarity searches, it streamlines the process of retrieving pertinent information. Fig. 2 likely provides a visual depiction of this innovative process, aiding in the understanding of its intricacies. This process is further elaborated in the following sections: Data Collection and Document Pre-processing (section A), Document Embedding and Storage (section B), and Search Similarity and Leveraging Answers (section C).

A. Data Collection and Document Pre-processing

The first step in designing this system involves collecting data related to contracts. For example, we used the official documentation of the law governing contracts in Morocco. We found that contract management in Morocco is called the "Code of Obligations and Contracts (COC)," promulgated by the Dahir of August 12, 1913, which serves as the foundation of contract law in the country [53]. This code governs the formation, execution, and nullity of contracts, specifying that contracts must comply with conditions of consent, capacity, lawful object, and cause. It encompasses various types of contracts, such as sales, leases, and mandates, defining the obligations and responsibilities of the parties involved. Additionally, the COC provides mechanisms for remedies in cases of non-performance or nullity of contracts and includes provisions on contractual and extra-contractual liability, thereby ensuring legal security and clarity in contractual relationships in Morocco.

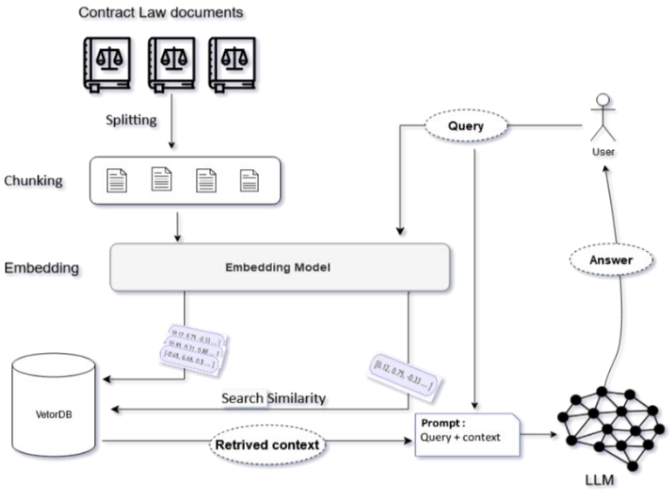


Fig. 2. Overview of the workflow for AI assistance in contract law.

The dataset derived from this documentation includes several key features, as illustrated in Fig. 3, which shows the distribution of articles by section.

- Articles: Each entry corresponds to a specific legal article, categorized by its thematic section.
- Sections: The dataset covers critical areas of contract law, including General Obligations, Electronic Contracts, Quasi-Contracts, Torts and Liabilities, Contractual Terms, and Solidarity.
- Language: The dataset is entirely in French, ensuring consistency with Moroccan legal texts.

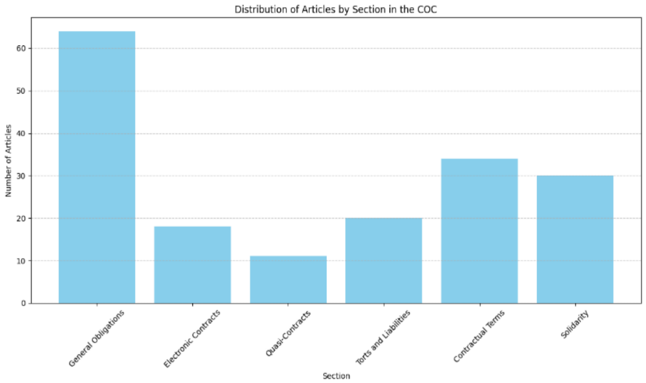


Fig. 3. Distribution of Articles by Section in the COC.

We would like to note that the documentation used for the implementation of this system is digital and in PDF format, which facilitated the collection and analysis of the necessary information. To efficiently extract the content from the documents as text, we used OCR technology, which converts text images into usable digital text. This tool is particularly well-suited for such technical documents. Each document is processed by OCR and then divided into text segments of up to 4,000 characters using a recursive text splitter. The splitting points are determined using separators such as double line breaks, single line breaks, periods, exclamation marks, question marks, spaces, and, as a last resort, no specific separator. Moreover, there is no overlap between segments, resulting in independent and coherent text chunks, thereby facilitating analysis while minimizing information loss.

B. *Document Embedding and Storage*

This step involves converting the segments prepared in the first phase into a format suitable for queries and storing them in a database. To achieve this, we use the LLM-Embedder model, known for its ability to capture complex semantic relationships in texts. This model, characterized by its precision in analyzing linguistic nuances, is ideal for extracting relevant and context-rich information from documents [54]. Using this approach, we obtain vector representations that faithfully reflect the semantic content of the documents. These results are then stored in a vector database, enabling efficient search and retrieval based on similarity queries. Finally, each text associated with an embedding is recorded with a pointer, ensuring precise retrieval based on the corresponding embedding. Fig. 4 illustrates the process of creating this database.

C. *Leveraging Answers*

To answer user questions or queries, we follow a procedure that begins with the embedding of the user's question using the same embedding model as that used for the knowledge base. We then use the resulting embedding vector to query the vector database index, selecting three vectors (Top-k = 3) to determine the amount of context to retrieve. The vector database then performs a nearest-neighbor (ANN) search against the embedding provided, returning the most similar context vectors as illustrated in Fig. 6. It is through this similarity search that we extract the relevant documents from the database. For this task, we opted for the use of Facebook AI Similarity Search (FAISS), recognized for its efficient and scalable similarity search capabilities, particularly suited to the management of large datasets [55], [56]. FAISS offers significant advantages, particularly in terms of speed [57]. We then associate these vectors with the corresponding chunks of text, and transmit the question and the retrieved context to the LLM via a prompt. We ask the LLM to use only the context provided to answer the question, while ensuring that the answers respect the limits laid down for this type of sensitive information. If the context contains no usable data, the system will return a standard “Not applicable” message, as illustrated in Fig. 5, which includes the activity diagram describing the entire process.

To select the best large-scale language model (LLM) for generating chatbot responses, we carried out a comparative analysis of the results obtained using GPT-4 Turbo and Llama 3. GPT-4 Turbo, developed by OpenAI, offers improved speed and accuracy, in-depth understanding of natural language and advanced personalization options. Its outstanding performance places it among the best LLMs available, and its enriched contextual memory makes it particularly effective in delivering consistent, relevant user interactions. For its part, Meta Llama 3, a family of models developed by Meta, is considered the current state of the art and is available in 8 billion and 70 billion parameter versions, pre-trained or adjusted on instruction. The instruction-tuned Llama 3 models are optimized for specific chat use cases and outperform many others. Table I below contains the detailed characteristics of each model used in this study.

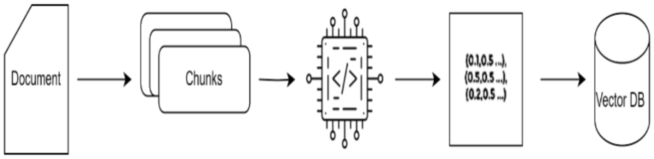


Fig. 4. Vector database creation process.

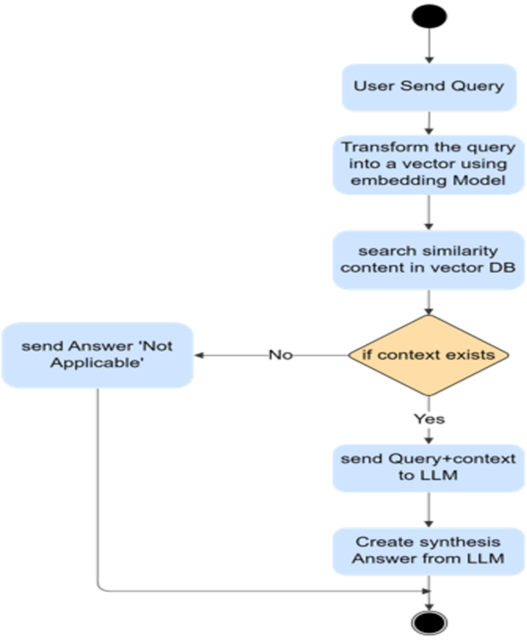

User Send Query
Transform the query into a vector using embedding Model
search similarity content in vector DB
if context exists
No
send Answer 'Not Applicable'
Yes
send Query+context to LLM
Create synthesis Answer from LLM

Fig. 5. Activity diagram for leveraging answers process.

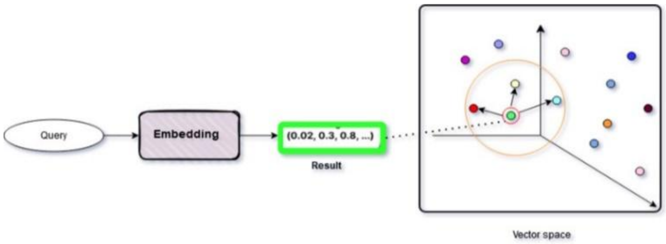


Fig. 6. Vector similarity search process with Top-k = 3.

TABLE I. FEATURES OF LARGE LANGUAGE MODELS UTILIZED IN THIS STUDY

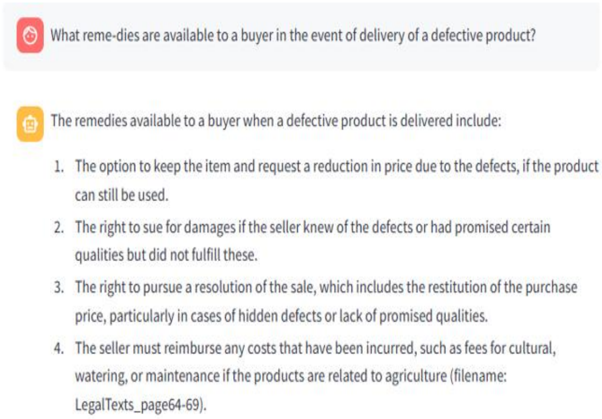
Model	Organization	context length	Knowledge cutoff date
GPT 4 - Turbo	OpenAI	128K tokens	April 2023
Llama 3	Meta	8K tokens	December 2023

V. RESULTS

A. Question-Answers Results of AI Assistance

The results of implementing the contract law assistance system, based on datasets from Moroccan law as an example, demonstrate remarkable performance in providing accurate and contextualized responses to users. Built on an architecture combining search and generation assisted by generative artificial intelligence, this system has proven its effectiveness in handling complex legal information.

For instance, when faced with a question such as, "What remedies are available to a buyer in the event of delivery of a defective product?", the model provided a detailed response enumerating the remedies available to the buyer, as illustrated in Fig. 7 below, relying entirely on information contained in the Moroccan COD law documentation. The large language model plays a crucial role by synthesizing the data provided by the system's prompt and ensuring that the response is coherent and contextualized based on the extracted information. This example demonstrates the chatbot's ability to extract and articulate specific and relevant information from foundational documents. However, when the system cannot find any relevant context to answer a question, as illustrated in Fig. 8, we do not allow the large language model to generate a response based solely on its general knowledge. In such cases, the system returns a response like "Not applicable" to avoid any misinformation or inaccuracies. This approach ensures that the responses provided are always based on reliable and contextualized information derived from the foundational documents.

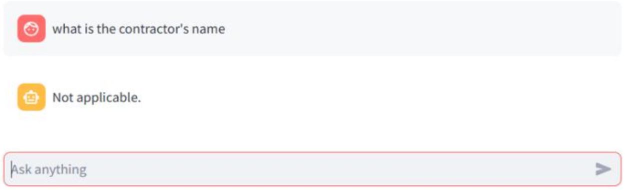


What reme-dies are available to a buyer in the event of delivery of a defective product?

The remedies available to a buyer when a defective product is delivered include:

1. The option to keep the item and request a reduction in price due to the defects, if the product can still be used.
2. The right to sue for damages if the seller knew of the defects or had promised certain qualities but did not fulfill these.
3. The right to pursue a resolution of the sale, which includes the restitution of the purchase price, particularly in cases of hidden defects or lack of promised qualities.
4. The seller must reimburse any costs that have been incurred, such as fees for cultural, watering, or maintenance if the products are related to agriculture (filename: LegalTexts_page64-69).

Fig. 7. Example of the system's answer.



what is the contractor's name

Not applicable.

Fig. 8. Example of the system's response in the absence of relevant context in the used documents.

To study the performance of the large language models integrated into the pipeline of this system, we collected responses provided by GPT-4 Turbo and Llama 3 to various questions, as mentioned in the Methodology section. Table III below presents the results, allowing for a direct comparison of the two models' performance in terms of response quality and response time. These data provide a solid foundation for further analysis in the following section. Finally, it is important to note that the final version of this chatbot will use the LLM that performs best during the evaluation.

B. Performance Evaluation of Answer Generation

To evaluate the results obtained by each LLM in the RAG pipeline of this study we chose the RAGAS framework introduced by Es et al. In 2023. The Retrieval-Augmented Generation Evaluation Framework (Ragas), represents a breakthrough in the evaluation of Retrieval-Augmented Generation systems. In essence, Ragas equips practitioners with state-of-the-art tools, rooted in the latest research, to rigorously scrutinize text generated by LLMs. By leveraging Ragas, deep insights can be gained into the effectiveness of their RAG pipeline. At the heart of the Ragas framework lies its ability to evaluate RAG systems on a component-by-component basis, dividing the pipeline into its constituent parts: the Retriever component and the Generator component. This granular approach provides a nuanced understanding of system performance, identifying strengths and weaknesses at each stage of the process [58]. In the context of this study, our focus is on the evaluation of the Generator component within the RAG pipeline. This component plays a crucial role in the generation of text based on retrieved context, thus requiring careful examination. Ragas offers two main evaluation indicators designed specifically to assess the effectiveness of the Generator component:

- Faithfulness: This metric assesses the factual consistency of the generated response in relation to the context provided. It determines the extent to which the statements made in the response can be logically deduced from the context [58]. Scores range from 0 to 1, with higher values indicating greater fidelity.
- Answer Relevance: evaluates the relevance of the generated answer to the question asked. It calculates the average cosine similarity between the original question and synthetic questions derived from the answer [58]. Higher scores mean greater relevance, indicating that the answer closely matches both the question and the context.

The comparative evaluation of the performance of GPT-4 Turbo and Llama 3, as shown in Table II and graphically represented in Fig. 9, highlighted notable differences in response time, fidelity, and relevance of the generated answers.

TABLE II. EVALUATION METRICS FOR RESPONSE GENERATION

Model	Faithfulness	Answer Relevance	Response time
gpt-4-turbo	1	0,971	3.12
Llama3	0.84	0.79	0.86

TABLE III. RESPONSES OBTAINED FOR EACH MODEL UTILIZED

Questions	**Answers**		**Response time**	
	gpt-4-turbo	***Llama3***	***gpt-4-turbo***	***Llama3***
What are the essential elements of a valid contract?	The essential elements for the validity of obligations deriving from a declaration of will are: 1) the capacity to obligate, 2) a valid declaration of will concerning the essential elements of the obligation, 3) a certain object that can form the subject of an obligation, and 4) a lawful cause to obligate (Article 2, Dahir (9 ramadan 1331) formant Code des obligations et des contrats).	The essential elements of a valid contract include the identity of the contracting parties, the title of the property, the number and date of the building permit, and the description of the property being sold (Dahir.pdf, page 69).	3.10	1.21
What are the defects in consent that can render a contract null and void?	Defects in consent that can render a contract null and void include error, fraud (dol), and duress (violence). These defects impact the validity of the consent given and can lead to the annulment of the contract (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 6).	The defects in consent that can render a contract null and void are error, violence, fraud, and dol (Dahir.pdf, page 50).	2.95	0.66
What are the conditions for claiming damages in the event of non-performance of a contract?	To claim damages for non-performance of a contract, the following conditions must be met: 1) there must be a failure to perform the obligation (non-performance), 2) the non-performance must be attributable to the debtor (fault must be proven unless presumed), and 3) there must be a causal link between the non-performance and the damage incurred. Additionally, the claimant must prove the damage and the amount of loss (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 18).	The conditions for claiming damages in the event of non-performance of a contract are not explicitly stated in the provided text. However, it mentions that in case of resale, the seller must be notified by registered letter with acknowledgment of receipt, and the resale must be done under the same conditions as the preliminary contract (Dahir.pdf, page 71).	3.46	0.80
What happens if one of the parties fails to perform a contract?	If one of the parties fails to perform a contract, the other party may demand specific performance, if applicable, or seek the resolution of the contract in court. Additionally, the non-breaching party may claim damages for losses incurred due to the breach (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 25).	If one of the parties fails to perform a contract, the other party may be entitled to damages for the prejudice suffered due to the non-performance or delay in performance of the obligation (Dahir.pdf, page 29).	4.00	0.67
What are the seller's obligations under a in a contract of sale?	The seller's obligations in a contract of sale primarily include: 1) delivering the thing sold, and 2) guaranteeing the buyer's peaceful possession and enjoyment of the thing sold, free from defects and third-party claims (Articles 498, 499, Dahir (9 ramadan 1331) formant Code des obligations et des contrats).).	Based on the general concept of a contract of sale, the seller's obligations typically include delivering the goods, ensuring the goods are of acceptable quality, and providing any necessary documentation.	3.34	1.06

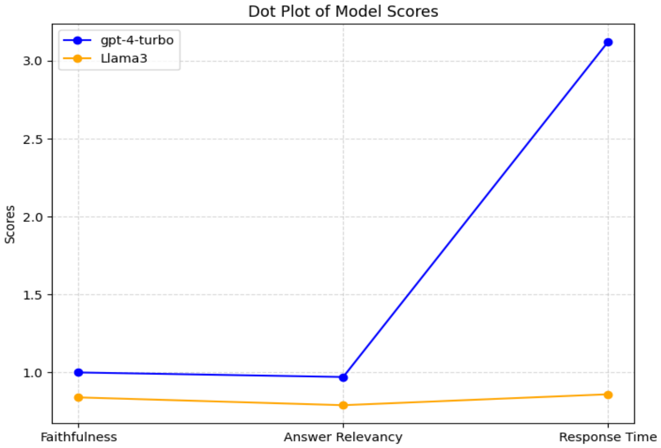


Fig. 9. Comparison of GPT-4 turbo and llama 3 across evaluation metrics.

The average response time for GPT-4 Turbo is 3.12 seconds, whereas Llama 3 stands out with a remarkably reduced response time of 0.86 seconds. This superior speed of Llama 3 is attributed to the use of Groq technology during its integration into our system. Groq is an advanced technology that optimizes query processing and efficiently meets the demands of high-responsiveness environments. However, this improvement in speed for Llama 3 appears to have come at the cost of a slight reduction in answer quality. The data presented in Table II reveal that GPT-4 Turbo outperforms Llama 3 in terms of fidelity (1 versus 0.84) and answer relevance (0.971 versus 0.79). The high fidelity of responses generated by GPT-4 Turbo indicates a better ability to align with the facts and the provided context, ensuring a reliable and accurate user experience.

VI. DISCUSSION

Based on the results obtained, the GPT-4 Turbo model was selected for definitive integration into the RAG process of our assistance system, specifically in the legal domain, and more precisely in contract law. This field demands a very high level of fidelity in responses, given the sensitivity and rigor associated with legal inquiries. The exceptional performance of GPT-4 Turbo in terms of fidelity (score of 1) and response relevance (score of 0.971) makes it the most suitable model to meet the critical needs of users in this demanding context. Additionally, the system has been designed to rely on legal texts as the primary source during the data augmentation phase for answer generation. This strategy ensures enhanced contextual accuracy and better alignment with the specific needs of users in the legal field. GPT-4 Turbo, with its ability to produce factually aligned responses while adhering to precise contextual frameworks, has proven ideal for handling complex queries while efficiently leveraging augmented data. While the Llama 3 model offers undeniable advantages in terms of speed, the priority given to response fidelity and contextual relevance in this specific domain fully justifies the choice of GPT-4 Turbo. This decision strengthens the reliability and robustness of the system, particularly in an environment where information accuracy is critical and any error could lead to significant consequences.

The results of this study illustrate the effectiveness of Retrieval-Augmented Generation systems in legal assistance, particularly in contract management. By combining legal vector databases with large language models, the developed chatbot delivers accurate and contextually appropriate responses. This system surpasses existing chatbots in this domain by adding an intelligence layer based on official legal data, such as national laws and regulations, ensuring strict compliance with current legislation. Unlike standalone LLMs, which may generate incorrect or non-compliant information, this chatbot leverages relevant legal data to avoid hallucinations or errors. For example, the data used in this study is based on Moroccan legislation, ensuring compliance with the national legal framework.

The GPT-4 Turbo and Llama3 models integrated into the development process of this system showed distinct results: GPT-4 Turbo stands out for the richness of its responses, despite being slower, while Llama3 offers superior speed but with less depth in complex cases. When the sought information is not explicitly available in the legal texts, the system uses the "Not applicable" option to avoid providing incorrect or unfounded answers. This system represents a significant advancement in access to justice, facilitating the understanding of laws and the creation of compliant contracts while ensuring essential transparency and reliability. However, it cannot replace human expertise in complex legal cases. Future improvements are planned, such as automating legal updates, integrating multimodal functionalities, and continuously evaluating performance through frameworks like RAGAS. It is crucial to clarify the system's limitations to avoid any misinterpretation, emphasizing that this chatbot is not intended to replace professional legal advice.

VII. Conclusion

This work highlights the growing impact of AI-based technologies, particularly advanced language models combined with Retrieval-Augmented Generation systems, in the field of legal assistance. By relying on official documentation and rigorous information management, the developed system represents a significant step forward in improving access to justice by providing responses tailored to specific legal requirements.

This approach overcomes some of the limitations of traditional chatbots and LLMs, ensuring greater accuracy in the information provided. The system also stands out for its flexibility, as it can be adapted to different countries simply by modifying the legal databases, ensuring its relevance to various legislative contexts. However, while this system is a powerful tool for automating certain legal tasks, it does not replace human expertise, particularly in more complex cases. Nevertheless, it paves the way for the democratization of access to legal information, emphasizing the importance of regular updates and continuous vigilance to avoid errors in ever-evolving contexts.

Future developments will focus on automating legal updates, integrating multimodal capabilities such as speech recognition and document analysis, improving explainability by providing explicit legal references, and enhancing adaptability to different legal frameworks. By tackling these challenges, future iterations of this system will significantly enhance the accessibility, accuracy, and usability of AI-driven legal assistance.

References

[1] L. TARANENKO and N. CHUDYK-BILOUSOVA, “The Role of Legal Service for Contractual Work Organization in Social and Medical Spheres,” University Scientific Notes, 2021, doi: 10.37491/unz.80.9.

[2] B. Alarie, A. Niblett, and A. H. Yoon, “How artificial intelligence will affect the practice of law,” 2018. doi: 10.3138/utlj.2017-0052.

[3] S. B. Shedthi, V. Shetty, R. Chadaga, R. Bhat, B. Preethi, and P. K. Kini, “Implementation of Chatbot that Predicts an Illness Dynamically using Machine Learning Techniques,” International Journal of Engineering, Transactions B: Applications, vol. 37, no. 2, pp. 312–322, Feb. 2024, doi: 10.5829/IJE.2024.37.02B.08.

[4] F. Firdaus, R. A. Rajagede, A. Sari, S. Hanifah, and D. A. Perwitasari, “Digital Assistant for Pharmacists Using Indonesian Language Based on Rules and Artificial Intelligence,” International Journal of Engineering, vol. 37, no. 9, pp. 1746–1754, 2024, doi: 10.5829/ije.2024.37.09c.04.

[5] S. Perez-Soler, S. Juarez-Puerta, E. Guerra, and J. De Lara, “Choosing a Chatbot Development Tool,” IEEE Softw, vol. 38, no. 4, pp. 94–103, 2021, doi: 10.1109/MS.2020.3030198.

[6] R. P. Karchi, S. M. Hatture, T. S. Tushar, and B. N. Prathibha, AI-Enabled Sustainable Development: An Intelligent Interactive Quotes Chatbot System Utilizing IoT and ML, vol. 1939 CCIS. 2023. doi: 10.1007/978-3-031-47055-4_17.

[7] A. Savanur, M. Niranjanamurthy, M. P. Amulya, and P. Dayananda, “Application of Chatbot for consumer perspective using Artificial Intelligence,” in Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021, 2021. doi: 10.1109/ICCES51350.2021.9488990.

[8] S. Meshram, N. Naik, M. Vr, T. More, and S. Kharche, “Conversational AI: Chatbots,” in 2021 International Conference on Intelligent Technologies, CONIT 2021, 2021. doi: 10.1109/CONIT51480.2021.9498508.

[9] W. Sanjaya, Calvin, R. Muhammad, Meiliana, and M. Fajar, “Systematic Literature Review on Implementation of Chatbots for Commerce Use,” in Procedia Computer Science, 2023, pp. 432–438. doi: 10.1016/j.procs.2023.10.543.

[10] T. Jindal, L. N. U. Ishika, P. Sharma, and G. Kaur, Chatbots benefications towards the education sector. 2023. doi: 10.4018/978-1-6684-8671-9.ch006.

[11] M. W. Ashfaque, S. Tharewal, T. Malche, S. I. Malikb, and C. N. Kayte, “Analysis Of Different Trends In Chatbot Designing And Development: A Review,” in ECS Transactions, 2022, pp. 7215–7227. doi: 10.1149/10701.7215ecst.

[12] R. Negi and R. Katarya, “Emerging Trends in Chatbot Development : A Recent Survey of Design, Development and Deployment,” in 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023, 2023. doi: 10.1109/ICCCNT56998.2023.10307280.

[13] H. Naveed et al., “A Comprehensive Overview of Large Language Models,” Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.06435

[14] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, 2020.

[15] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large Language Models Struggle to Learn Long-Tail Knowledge,” in Proceedings of Machine Learning Research, 2023.

[16] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.07567

[17] Z. Levonian et al., "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.03184

[18] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[19] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," 2024. [Online]. Available: www.aaai.org

[20] W. S. Nsaif, H. M. Salih, H. H. Saleh, and B. Talib, "Conversational Agents: An Exploration into Chatbot Evolution, Architecture, and Important Techniques," in Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 2024, pp. 246–262. doi: 10.55549/epstem.1518795.

[21] N. Biswas, S. Biswas, and S. Maity, Analysis of chatbots: History, use case, and classification. 2024. doi: 10.4018/979-8-3693-1830-0.ch004.

[22] K. B. Prakash, A. J. S. Kumar, and G. R. Kanagachidambaresan, Chatbot. 2021. doi: 10.1007/978-3-030-57077-4_9.

[23] P. Kandpal, K. Jasnani, R. Raut, and S. Bhorge, "Contextual chatbot for healthcare purposes (using deep learning)," in Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020, 2020, pp. 625–634. doi: 10.1109/WorldS450073.2020.9210351.

[24] G. K. Ahirwar, Chatterbot: Technologies, tools and applications, vol. 913. 2020. doi: 10.1007/978-981-15-6844-2_14.

[25] C. Ionuț-Alexandru, Experimental Results Regarding the Efficiency of Business Activities Through the Use of Chatbots, vol. 276. 2022. doi: 10.1007/978-981-16-8866-9_27.

[26] P. Suta, X. Lan, B. Wu, P. Mongkolnam, and J. H. Chan, "An overview of machine learning in chatbots," International Journal of Mechanical Engineering and Robotics Research, vol. 9, no. 4, pp. 502–510, 2020, doi: 10.18178/ijmerr.9.4.502-510.

[27] M. Ahmed, H. U. Khan, and E. U. Munir, "Conversational AI: An Explication of Few-Shot Learning Problem in Transformers-Based Chatbot Systems," IEEE Trans Comput Soc Syst, vol. 11, no. 2, pp. 1888–1906, 2024, doi: 10.1109/TCSS.2023.3281492.

[28] A. Chizhik and Y. Zherebtsova, "Challenges of building an intelligent chatbot," in CEUR Workshop Proceedings, 2021, pp. 277–287.

[29] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[30] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.03299

[31] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Of Wikipedia: Knowledge-powered conversational agents," in 7th International Conference on Learning Representations, ICLR 2019, 2019.

[32] Y. Mao et al., "Generation-augmented retrieval for open-domain question answering," in ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2021. doi: 10.18653/v1/2021.acl-long.316.

[33] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703.

[34] W. Yu, "Retrieval-augmented Generation across Heterogeneous Knowledge," in NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop, 2022. doi: 10.18653/v1/2022.naacl-srw.7.

[35] S. S. Sonawane, P. N. Mahalle, and A. S. Ghotkar, Information Retrieval, vol. 104. 2022. doi: 10.1007/978-981-16-9995-5_4.

[36] E. Tzoukermann, J. L. Klavans, and T. Strzalkowski, Information Retrieval, vol. 9780199276. 2012. doi: 10.1093/oxfordhb/9780199276349.013.0029.

[37] M. Erritali, "Information retrieval: Textual indexing using an oriented object database," Indonesian Journal of Electrical Engineering and Computer Science, vol. 2, no. 1, pp. 205–214, 2016, doi: 10.11591/ijeecs.v2.i1.pp205-214.

[38] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[39] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[40] B. Li, P. Yang, Y. Sun, Z. Hu, and M. Yi, "Advances and challenges in artificial intelligence text generation," Frontiers of Information Technology and Electronic Engineering, vol. 25, no. 1, pp. 64–83, 2024, doi: 10.1631/FITEE.2300410.

[41] D. Hijam, S. Gottipati, and S. Fardeen, "Telugu Text Generation with LSTM," in 2024 3rd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2024, 2024. doi: 10.1109/ICSTSN61422.2024.10670921.

[42] D. Weber-Wulff, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, and L. Waddington, "Testing of Detection Tools for AI-Generated Text."

[43] J. Casas, M. O. Tricot, O. Abou Khaled, E. Mugellini, and P. Cudré-Mauroux, "Trends & methods in chatbot evaluation," in ICMI 2020 Companion - Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020. doi: 10.1145/3395035.3425319.

[44] J. Ng, E. Haller, and A. Murray, "The ethical chatbot: A viable solution to socio-legal issues," Alternative Law Journal, vol. 47, no. 4, 2022, doi: 10.1177/1037969X221113598.

[45] M. Wyawahare, S. Roy, and S. Zanwar, "Generative vs Intent-based Chatbot for Judicial Advice," in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024, 2024. doi: 10.1109/IATMSI60426.2024.10502550.

[46] P. Giampieri, "AI-Powered Contracts: a Critical Analysis," International Journal for the Semiotics of Law, 2024, doi: 10.1007/s11196-024-10137-z.

[47] A. Kumar, P. Joshi, A. Saini, A. Kumari, C. Chaudhary, and K. Joshi, Smart Chatbot for Guidance About Children's Legal Rights, vol. 681. 2023. doi: 10.1007/978-981-99-1909-3_35.

[48] I. Qasse, S. Mishra, and M. Hamdaqa, "IContractBot: A Chatbot for Smart Contracts' Specification and Code Generation," in Proceedings - 2021 IEEE/ACM 3rd International Workshop on Bots in Software Engineering, BotSE 2021, 2021. doi: 10.1109/BotSE52550.2021.00015.

[49] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[50] B. Peng et al., "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.12813

[51] S. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," in Proceedings of the 39th International Conference on Machine Learning, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, May 2022, pp. 2206–2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[52] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models." [Online]. Available: https://github.com/

[53] W. Lex, "Code des obligations et des contrats, notamment les articles 77, 79, 80, 81, 84, 264 et 435, (Dahir du 12 août 1913 (9 ramadan 1331))," 1913.

[54] P. Zhang, S. Xiao, Z. Liu, Z. Dou, and J.-Y. Nie, "Retrieve Anything To Augment Large Language Models," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.07554

[55] C. and S. D. Danopoulos Dimitrios and Kachris, “Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud,” in Embedded Computer Systems: Architectures, Modeling, and Simulation, M. and J. M. Pnevmatikatos Dionisios N. and Pelcat, Ed., Cham: Springer International Publishing, 2019, pp. 373–386.

[56] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08734

[57] C. Mu, B. Yang, and Z. Yan, “An Empirical Comparison of FAISS and FENSHSES for Nearest Neighbor Search in Hamming Space,” Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.10095

[58] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.15217.