الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

Registration n°:
...../...../...../...../.....

جامعة غرداية
University of Ghardaia
والتكنولوجيا العلوم كلية
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science
التطبيقية والعلوم الرياضيات مخبر
Mathematics and Applied Sciences Laboratory

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer Science
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction

Topic

RAG-based Question-Answering System for Algerian Tax law Context

---

Presented by:

Wissal Hamani & Zineb Benyounes

Publicly defended on June 30, 2025

Jury members:

Dr. Slimane Oulad-Naoui	MCB	Univ. Ghardaia	President
Mr. Abderrahmane ADJILA	MAA	Univ. Ghardaia	Examiner
Dr. Slimane Bellaouar	MCA	Univ. Ghardaia	Supervisor
Dr. Attia Nehar	MCA	Univ. Z.A. Djelfa	Co-Supervisor

Academic Year: 2024/2025

Acknowledgment

First and foremost, we express our deepest gratitude to Allah, the Most Merciful and Compassionate, for His infinite blessings, guidance, and strength throughout every stage of this journey. It is by His will that we have reached this significant milestone.

We are sincerely thankful to our esteemed supervisors, **Dr. Slimane Bellaouar** and **Dr. Nehar Attia** for their exceptional mentorship, continuous support, and scholarly insight. Their guidance has been fundamental in shaping our research and refining our ideas. It has been a privilege and honor to work under their supervision.

We would also like to extend our sincere gratitude to **Mr. Abdelfateh Bekkair** for his invaluable support and meaningful contributions throughout this endeavor.

Our deepest gratitude goes to our families, whose unwavering love, patience, and moral support have been a constant source of strength. Their belief in our abilities has encouraged and sustained us throughout this journey.

We are also sincerely grateful to the distinguished members of the thesis jury for their time, thoughtful evaluation, and constructive feedback. We are honored by your presence and deeply appreciate your contribution to the success of this work.

Special thanks are extended to the faculty members, administrative staff, and colleagues at Ghardaïa University for creating a supportive academic environment and for the learning opportunities we have been fortunate to receive.

Finally, we wish to acknowledge all individuals who, in one way or another, contributed to the completion of this thesis. Whether through academic discussions, technical support, or moral encouragement, your efforts have been truly appreciated.

This thesis is the outcome of collective efforts, and we remain profoundly grateful to everyone who has played a part in this accomplishment. May Allah bless you all abundantly for your kindness, guidance, and support.

Dedication

First and foremost, I thank God the source of all strength, wisdom, and guidance without whom none of this would have been possible.

This thesis is dedicated to the pillars of my life those whose unwavering love, strength, and belief have carried me through every step of this journey.

To my beloved **grandfather**, and my cherished **grandmother**, your lives have been a beacon of resilience and wisdom. Your enduring faith in me has quietly lit my path, and your legacy of love and perseverance is woven into every word of this work.

To my incredible **parents** my first mentors and eternal supporters your sacrifices, unconditional love, and steadfast encouragement have been the foundation of all my accomplishments. Your strength gave me courage, and your guidance gave me purpose. I am who I am because of you.

To my dear siblings, **Imen**,**Ali** , and **Asma** your presence has filled my life with joy and meaning. Thank you for walking beside me with patience, laughter, and unwavering support. You've made this journey lighter and more colorful.

To my closest friends , **Kaoutar** and **Kheira** thank you for standing by my side through every challenge. Your companionship transformed obstacles into shared triumphs, and the memories we've built will remain with me forever.

To my dear friend **Zineb**,Through every twist and triumph of this journey, you stood by my side.Your insight, resilience, and quiet strength are reflected in every part of this project.Thank you for being there, every step of the way.

This thesis is not mine alone it belongs to each of you. It is a testament to your love, your faith, and the strength you've given me. From the depths of my heart, thank you.

Hamani Wissal

Dedication

Praise be to God, who taught mankind what it did not know, and by whose grace good deeds are completed. To Him belongs all thanks—first and last, in secret and in public for the blessings He bestowed upon me, guiding and supporting me throughout my academic journey.

To my beloved **mother**, the heartbeat of my soul and the light of my path, my first pillar of support and source of strength your prayers, both whispered and spoken, your sleepless nights for my sake every letter and every achievement is dedicated to you. Your prayers were my greatest support on this journey.

To my dear **father**, my backbone and my pride—from whom I learned the true meaning of responsibility and perseverance, and that ambition knows no limits. You have never withheld your effort or prayers. Thank you for your priceless encouragement and unwavering belief in me.

To my beloved siblings, **Leila**, **Omar**, and **Maria**-my companions on this path and my source of strength who stood by me through every stage and brought joy during times of exhaustion. Thank you for your constant support and endless encouragement. You have always been, and will always be, my rock in this life.

To my cherished **family**, of whom I am proud to be a part, who shaped my identity and offered their support thank you for every proud look, every encouraging word, and every moment of closeness.

To my dear friend **Wissal**, who was a support in difficult moments, and a partner in this journey with all its challenges and achievements. Thank you for your support, your encouragement, and your constant presence in every step you have my heartfelt gratitude.

To everyone who made an impact on this journey, even with just a kind word to those who believed in me and reminded me of my capabilities when I doubted myself thank you from the depths of my heart. You are an inseparable part of this success.
This thesis is for you, and it is complete because of you.

Benyounes Zineb

مـــلـــخـــص

على الرغم من أن النماذج اللغوية الكبيرة تؤدي أداءً جيدًا في الإجابة عن الأسئلة العامة، إلا أن استخدامها في المجالات المتخصصة مثل القانون يواجه عدة تحديات، منها توليد إجابات غير دقيقة أو غير مدعومة بالنصوص القانونية، وصعوبة التعامل مع الأسئلة المعقدة بسبب نقص البيانات المتخصصة عالية الجودة. وتزداد هذه التحديات وضوحًا في السياق القانوني الجزائري، حيث إن النصوص القانونية باللغة العربية غالبًا ما تكون محدودة وضعيفة من حيث الرقمنة. تهدف هذه الرسالة إلى تطوير نظام للإجابة عن الأسئلة القانونية باللغة العربية، يستند إلى قانون الضرائب الجزائري، وذلك من خلال الجمع بين البحث الدلالي الكثيف ونموذج لغوي توليدي. يشمل العمل عدة مراحل: جمع النصوص القانونية من مصدر موثوق، معالجتها مسبقًا، تقسيمها إلى مواد قانونية، تمثيلها باستخدام نماذج مخصصة للغة العربية مثل TREBarA و ،5E وأرشفتها باستخدام SSIAF لتسهيل الاسترجاع. بعد ذلك، يُستخدم نموذج توليدي لصياغة الإجابة اعتمادًا على المادة المسترجعة. تم تنفيذ النظام باستخدام لغة nohtyP في بيئة ، elgooG baloC وتم تقييمه بناءً على جودة الاسترجاع ودقة الإجابات.

أظهرت النتائج التجريبية أن استخدام نموذج 5E في البحث الدلالي حقق نسبة استرجاع بلغت 19%، متفوقًا بشكل كبير على الطرق المعتمدة على الكلمات المفتاحية مثل 52MB . كما أن دمج المحتوى المسترجع مع نموذج توليدي مضبوط أدى إلى إنتاج إجابات أكثر دقة من الناحية القانونية وأكثر سلاسة، خصوصًا في التعامل مع الأسئلة متعددة الطبقات. وتبرز هذه النتائج فعالية الجمع بين البحث الدلالي والتوليد النصي في معالجة التحديات الفريدة للإجابة عن الأسئلة القانونية باللغة العربية في سياق القانون الضريبي الجزائري.

كلمات مفتاحية: الإجابة عن الأسئلة القانونية، الاسترجاع الدلالي، التوليد المعزز بالاسترجاع ، التضمين، معالجة اللغة الطبيعية العربية، قانون الضرائب الجزائري.

Abstract

While large language models perform well in answering general questions, their deployment in specialized domains such as law faces several challenges, including generating inaccurate answers or responses unsupported by legal texts, and difficulty handling complex questions due to the lack of high-quality specialized data. These challenges are even more pronounced in the Algerian legal context, where Arabic legal texts are often limited and poorly digitized. This thesis aims to develop a legal question-answering system in Arabic based on Algerian tax law by combining dense semantic retrieval with a generative language model. The work includes several phases: collecting legal texts from a reliable source, preprocessing them, segmenting them into legal articles, representing them using models adapted to the Arabic language such as AraBERT and E5, and archiving them using FAISS to facilitate retrieval. Then, a generative model is used to formulate the answer based on the retrieved article. The system was implemented using Python in the Google Colab environment and was evaluated based on retrieval quality and answer accuracy.

The experimental results demonstrated that the semantic retrieval approach using the E5 model achieved a recall of 91%, significantly outperforming keyword-based methods such as BM25. Furthermore, the integration of the retrieved content with a fine-tuned generative model led to more legally grounded and fluent answers, especially in handling multi-layered questions. These findings highlight the effectiveness of combining semantic search with generative modeling in addressing the unique challenges of Arabic legal question answering in the Algerian tax context.

Keywords: Legal QA, semantic retrieval, Retrieval-Augmented Generation, Embeddings, Arabic NLP, Algerian tax law.

Résumé

Bien que les grands modèles de langage soient performants pour répondre à des questions générales, leur déploiement dans des domaines spécialisés tels que le droit présente plusieurs défis, notamment la génération de réponses inexactes ou non étayées par des textes juridiques, ainsi que des difficultés à traiter des questions complexes en raison du manque de données spécialisées de haute qualité. Ces défis sont encore plus marqués dans le contexte juridique algérien, où les textes juridiques en arabe sont souvent limités et mal numérisés. Ce mémoire vise à développer un système de questions-réponses juridiques en arabe basé sur le droit fiscal algérien, en combinant une recherche sémantique dense avec un modèle de langage génératif. Le travail comprend plusieurs phases : la collecte de textes juridiques à partir d'une source fiable, leur prétraitement, leur segmentation en articles de loi, leur représentation à l'aide de modèles adaptés à la langue arabe tels que AraBERT et E5, puis leur archivage à l'aide de FAISS pour faciliter la recherche. Ensuite, un modèle génératif est utilisé pour formuler la réponse à partir de l'article récupéré. Le système a été implémenté en Python dans l'environnement Google Colab et évalué en fonction de la qualité de la recherche et de la précision des réponses.

Les résultats expérimentaux ont montré que l'approche de recherche sémantique utilisant le modèle E5 a atteint un rappel de 91%, surpassant largement les méthodes basées sur les mots-clés telles que BM25. De plus, l'intégration du contenu récupéré avec un modèle génératif ajusté a permis de produire des réponses plus juridiques, pertinentes et fluides, en particulier pour les questions complexes à plusieurs niveaux. Ces résultats mettent en évidence l'efficacité de la combinaison de la recherche sémantique et de la génération de texte dans le traitement des défis spécifiques aux systèmes de questions-réponses juridiques en arabe, dans le contexte du droit fiscal algérien.

Mots clés: Réponse aux questions juridiques, Recherche Sémantique, Génération Augmentée par Recherche, Représentations Vectorielles, Traitement Automatique du Langage Naturel en Arabe, Droit Fiscal Algérien.

Contents

List of Figures

List of Tables

Introduction

In recent years, the application of artificial intelligence (AI) in the legal domain has received growing attention, particularly for its potential to improve access to legal information and support legal research. One of the most promising developments in this area is the use of question answering (QA) systems, which aim to enable users to retrieve accurate and contextually appropriate legal information by submitting natural language queries. These systems leverage advancements in natural language processing (NLP) and have been significantly enhanced by the emergence of large language models (LLMs), which are capable of generating fluent and coherent responses across a wide range of topics.

Despite these advancements, applying LLMs in the legal field especially within the Algerian legal context presents notable challenges. A key limitation is the phenomenon of hallucination, where models generate answers that are syntactically plausible, but factually incorrect or ungrounded in legal sources. This issue is particularly problematic in high stakes domains such as law, where users require precise and reliable information. Additionally, the under-representation of Arabic legal texts in training data, combined with the complexity of the Arabic language and legal terminology, further complicates the development of effective QA systems tailored to Algerian law.

Despite the growing interest in Arabic legal text processing in recent years, most existing Legal Question-Answering systems remain primarily designed for English, with limited adaptation to the specific linguistic and legal characteristics of Arabic—particularly within the Algerian context. Early approaches to Legal QA relied heavily on traditional information retrieval techniques such as keyword matching, TF-IDF, and rule-based systems, which often struggled with the complexity and ambiguity of legal language, as noted by (Duong & Ho, 2014). Frequency-based methods such as TF-IDF, BM25 (Rosa et al., 2021), and co-occurrence based techniques (Martinez-Gil et al., 2019) have been widely used to improve relevance ranking, but often fail to capture the deeper semantic meaning of legal texts.

With the rise of machine learning, methods such as Support Vector Machines (SVM), Decision Trees, and hybrid models like Combined SVM and K-Nearest Neighbors (K-NN) (Alcántara Francia et al., n.d.) have been employed to enhance legal text classification and document retrieval (Kim, Xu, Lu, & Goebel, 2017). More linguistically informed approaches, such as Predicate Argument Structure-Based QA (Hoshino et al., 2019), have also been explored. However, these techniques typically require extensive manual feature engineering.

In recent years, deep learning techniques—particularly models like LSTMs—have shown promise in capturing the semantic nuances of legal texts (ADEBAYO et

al., 2016). Nevertheless, these models still face significant limitations when applied to low-resource languages such as Arabic, largely due to the lack of annotated legal corpora and the scarcity of domain-specific pretrained models. Although models like BERT have transformed natural language processing, their use in Arabic legal question answering is still limited. This is mainly because there aren't enough high-quality legal datasets in Arabic, and the legal language itself is often complex and hard to process. Arabic's unique grammar and rich vocabulary also make it more challenging for these models to perform well without careful adaptation.
Retrieval-Augmented Generation frameworks (Wiratunga et al., 2024) have emerged as a promising alternative, combining dense retrieval with generative models to generate contextually relevant responses, thereby addressing many of the limitations of earlier approaches in handling complex legal queries.

The main objective of this study is to address these challenges by developing a Retrieval-Augmented Generation (RAG) based question-answering system that grounds responses in verified Algerian legal documents. By integrating a semantic retrieval mechanism with a generative language model, the system aims to reduce hallucinations and improve the accuracy and relevance of generated answers. This research seeks to contribute both to the advancement of Arabic NLP technologies and to the practical improvement of legal information accessibility in Algeria, particularly through the development of systems capable of understanding and responding to legal queries in a reliable and context-aware manner.

The contribution of this work can be divided into two main aspects. The first focuses on the development of a semantic retrieval system based on the RAG (Retrieval-Augmented Generation) architecture, tailored to the Algerian legal context. To achieve this, official legal documents in Arabic were collected, cleaned, and segmented according to legal articles. These segments were then embedded using a pre-trained multilingual sentence embedding model and indexed using the FAISS library to enable efficient and semantically relevant retrieval. The second aspect involves the construction of a legal question-answering benchmark for the evaluation of the system. A set of question-answer pairs was created based on reliable legal sources, with a particular focus on Algerian tax law. Each question was processed through the RAG system, which first retrieves the most relevant legal articles and then generates a coherent answer using a pre-trained Arabic language model. The goal of this system is to enhance the accessibility and understanding of Algerian tax law for citizens, law students, and legal professionals.

This thesis is organized into three main chapters, in addition to the introduction and conclusion, as follows.

- Chapter 1 (Basic Concepts) introduces the foundational concepts necessary to understand the research problem. It starts with an overview of the Algerian legal context and proceeds to define the key components of Question Answering (QA) systems and Large Language Models (LLMs). A particular emphasis is placed on Retrieval-Augmented Generation (RAG), explaining the architecture, the retrieval and generation phases, the role of embedding models, and its applications. This chapter lays the groundwork for the development of the proposed legal QA system.
- Chapter 2 (State of the Art) presents a comprehensive literature review of existing approaches in Legal Question-Answering. It is divided into three

main parts: traditional methods (including frequency-based, TF-IDF, BM25, and co-occurrence-based techniques), machine learning approaches (such as SVM and KNN), and deep learning-based methods (including neural networks, transformer models, and RAG-based solutions). The chapter also reviews available datasets for Legal QA in both English and Arabic, providing a solid background on prior work in this area.

- Chapter 3 (Proposed Solution and Experiments) details the development process of the proposed Tax-RAG system for the Algerian legal context. It describes the stages of document processing, retrieval, and response generation using LLM. This chapter also introduces the construction of a legal benchmark designed to evaluate the system, including data sources, QA pair creation, and validation. Furthermore, the experimental setup, performance comparisons for both the retrieval and generation phases, and evaluation metrics are thoroughly discussed.
- The thesis concludes with a summary of the contributions, main findings, and perspectives for future improvements and extensions of this work.

Chapter 1

Basic Concepts

1.1 Introduction

This chapter presents the essential concepts needed to understand the development of a legal question-answering system specifically designed for the Algerian legal system.

We will start by exploring the Algerian legal system to better understand the domain of our work. Then, we will examine Arabic question-answering systems before diving into large language models (LLMs), discussing their functionality, and presenting examples of their use.

Finally, we will address the Retrieval Augmented Generation (RAG) approach, showcasing how it combines information retrieval techniques with the capabilities of LLMs to improve the performance and accuracy of question-answering systems.

1.2 Algerian Legal Context

Law is a set of rules that govern the behavior of individuals in society and regulate their relationships. The Algerian legal system is based on a combination of Islamic law, tribal customs, and French legal frameworks formed through colonialism.

The legal system is based on the Algerian Constitution, which is considered the supreme law of the country. The primary sources of Algerian law include legislative laws (enacted by parliament), executive regulations (such as decrees and orders), customary laws, and ratified international treaties.

However, despite this strong legal framework, Algerian citizens and legal professionals struggle to access legal texts efficiently. The lack of a dedicated system to answer the questions of citizens creates barriers to the law. Addressing this challenge is essential to ensure legal transparency and accessibility for all.

In this thesis, we aim to address this issue by focusing specifically on Algerian tax law. This choice is motivated by several key factors: its complexity, its frequent annual updates, and its central importance within the national legal framework.

Tax law governs how the state collects revenue from individuals and companies, defines the obligations of taxpayers, and plays a crucial role in shaping economic policy and financing public services.

In Algeria, tax law is divided into six main categories, as shown in Table 1.1.

Table 1.1: Overview of Algerian Tax Laws.

No.	Tax Law Name	Description
1	Registration Law	Governs the registration of legal documents and contracts, including applicable duties and fees.
2	Business Tax Law	Regulates taxes imposed on businesses and commercial activities in Algeria.
3	Stamp Law	Covers stamp duties applicable to official documents and transactions.
4	Direct Taxes and Similar Law	Defines taxes applied directly to individuals and entities, such as income and property taxes.
5	Indirect Tax Law	Pertains to taxes levied on goods and services, including VAT and customs duties.
6	Tax Procedure Law	Establishes the rules for tax collection, declarations, audits, and taxpayer rights.

Source:Source: Ministry of Finance – Tax Codes

1.3 Question-Answering Systems

Arabic question-answering systems are artificial intelligence systems that aim to understand questions asked in Arabic and provide accurate and relevant answers based on specific databases or information sourcesShaheen & Ezzeldin.

Question-answering systems rely on Natural Language Processing (NLP), Information Retrieval (IR), and Information Extraction (IE) to provide accurate and relevant answers.

The main stages of question-answering systems begin with Question Processing, where the user's query is carefully analyzed to grasp its meaning, context, and intent. This involves understanding the relationships between words and applying various operations such as removing insignificant words to clarify the core of the questionAllam & Haggag Following this, the Document Retrieval stage employs advanced search techniques to locate relevant texts that may contain the answer. These documents are then ranked and classified according to their relevance and importance to the query, often using methods like TF-IDF to ensure precision. Finally, in the Answer Extraction phase, the system meticulously extracts the precise answer from the retrieved documents, providing a focused and accurate response to the user's question.

The following diagram in Figure 1.1 summarizes the three stages.

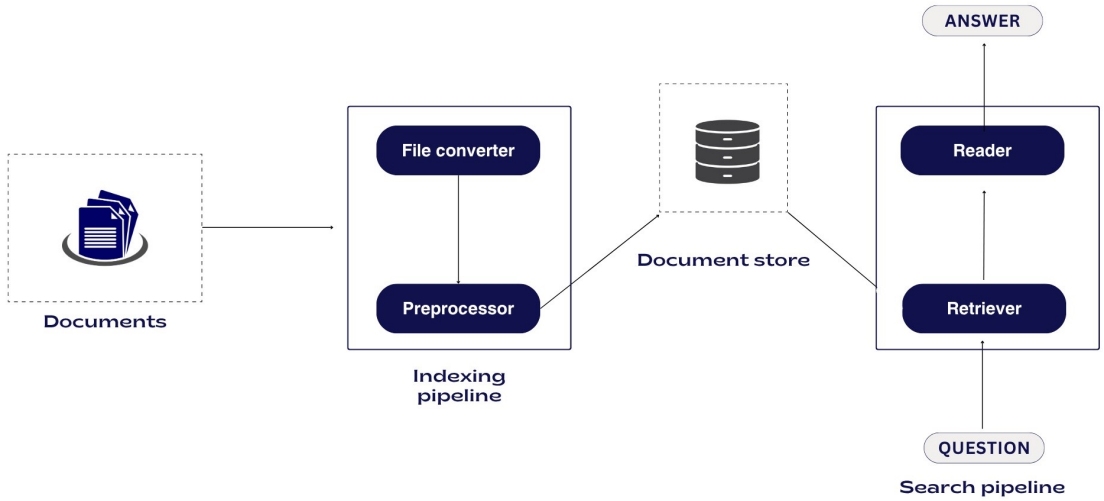

Figure 1.1: Stages of a Question Answering System.

1.4 Large Language Models (LLMs)

Large Language Models (LLMs) are powerful AI models based on Transformer architecture, that contain hundreds of billions of parameters W. X. Zhao et al., such as GPT-3, PaLM, Galactica, and LLaMA. LLMs are trained on massive amounts of text data to understand natural language and solve complex tasks.
In natural language processing, two main approaches are used in developing and implementing large language models (LLMs).

Pre-training: at the heart of every large language model is the pre-training phase, where the model learns to predict the next tokens in a huge data set through self-supervised learning. This stage establishes a significant linguistic and contextual foundationNaveed et al..
Pre-trained machine learning models excel at different tasks, but their performance can be significantly improved for specific applications by fine-tuning task-specific data.

Fine-tuning: Although pre-trained models demonstrate strong generalization to unfamiliar tasks, they often struggle to understand user intent, which may result in inaccurate or unethical responses.
A fine-tuning phase is necessary to address this limitation, where the model is further trained on instruction-based, structured datasets. This process enhances the model's ability to produce safer, more context-aware responses, with minimal additional computational cost Naveed et al..

Figure 1.2 is presented showing how the pretrained model is initially trained on large-scale general data and subsequently fine-tuned on domain-specific .

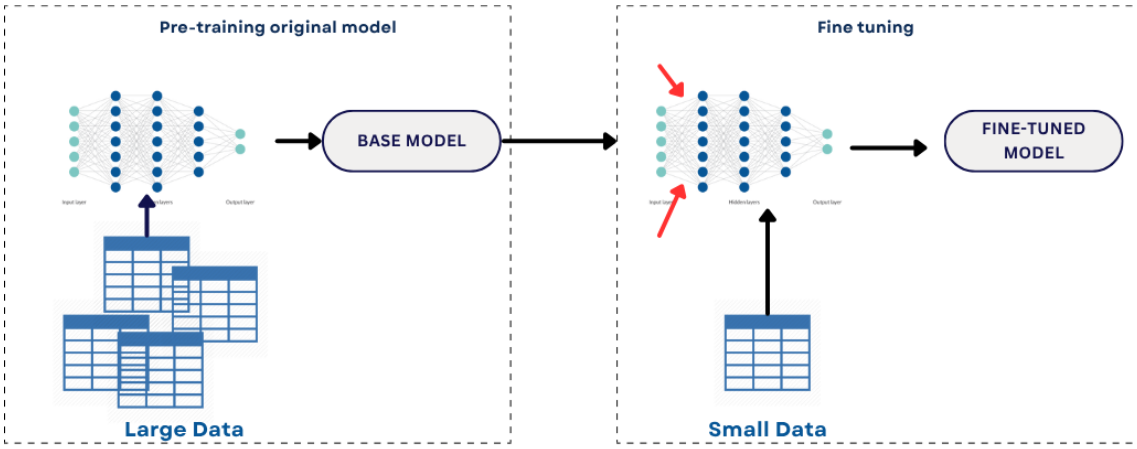

Figure 1.2: Overview of the Pretrained Model and Fine-Tuning.

1.5 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is an emerging technique that has gained importance in applications where high accuracy and up-to-date information are essential. RAG combines information retrieval methods with advanced text generation models. Unlike traditional models that rely on pre-training information, which is often insufficient and outdated, RAG enhances models' ability to provide accurate answers by retrieving relevant external data Zhang & Zhang.

1.5.1 Architecture of RAG

The RAG process starts with collecting data related to the topic under processing, then is divided into small parts, each part specialized in a specific topic to avoid retrieving irrelevant information. Then comes the document embedding process where the data are converted into a vector representation that captures the semantic meaning, then the query is converted into a vector representation with the same embedding model. The system compares the query embedding with the document embeddings and retrieves the most relevant document parts based on similarity. Finally, the initial query and the retrieved text parts are fed into an LLM model to generate a coherent and accurate response.

The following diagram in Figure 1.3 summarizes these steps.

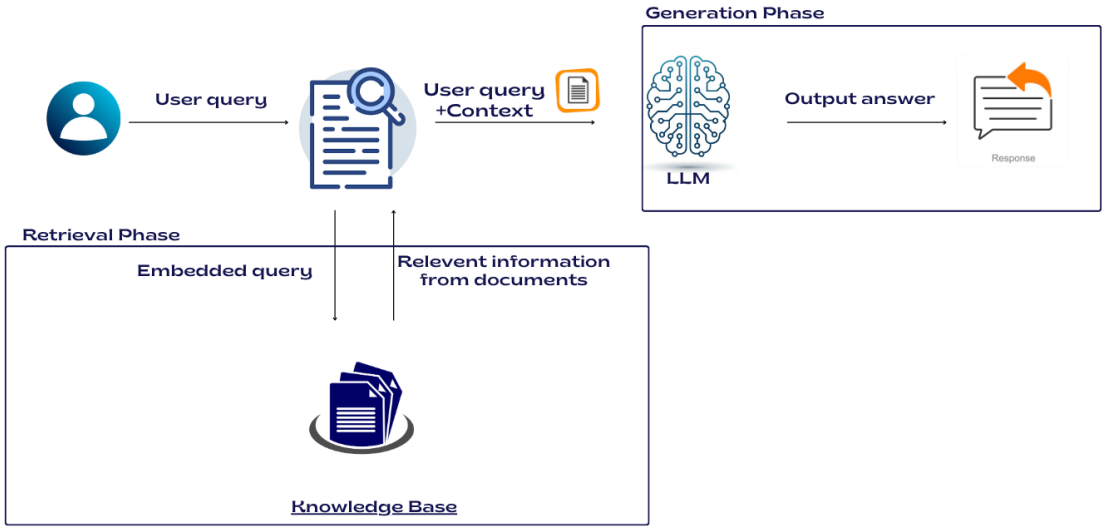

Figure 1.3: Retrieval-Augmented Generation (RAG).

1.5.2 Retrieval Phase

The performance of a RAG system heavily relies on the effectiveness of its retrieval algorithms, which provide context that subsequently feeds into the generative model, retrieving the most relevant information from broad external sources based on the user's query S. Zhao et al..
Retrieval techniques are classified into three main types based on how data is represented and processed:

Sparse Retrieval: This traditional retrieval method relies on keyword matching between queries and documents, and uses statistical models such as TF-IDF or BM25 to rank documents based on keyword overlap. While this approach is fast and effective for queries containing precise terms, it often fail to capture semantic meaning or contextual relationships between words.

Dense Retrieval: This method leverages deep learning models to transform both the query and documents into vector representations in a high-dimensional semantic space. These vectors are then compared using similarity measures (such as cosine similarity) to find the most semantically relevant results. Dense retrieval demonstrates superior performance in capturing context and implicit meanings, especially in specialized or linguistically complex domains.

Hybrid Retrieval: This method combines the strengths of both sparse and dense retrieval techniques to improve overall system performance. By integrating keyword-based matching results with semantic vector similarity, this approach leverage precise term matches and deep contextual understanding. Hybrid retrieval often achieves better accuracy and robustness compared to using either method independently.

1.5.3 Role of Embedding Models in RAG

Embedding models form the foundation of RAG systems by converting text into numerical vector representations that preserve semantic meaning.
These models rely on transforming words and sentences into a multidimensional vector space, where semantic similarity is quantified using metrics like cosine similarity, allowing for accurate retrieval even with different syntax.
Domain-specific models (e.g., for legal or medical domains) demonstrate superior performance in capturing specialized terminology, while general models face challenges such as the need for large computing resources or the difficulty of covering updated vocabulary without retraining.

1.5.4 Generation Phase

After retrieving the most relevant passages and documents from the knowledge base, the answer generation phase begins using language models such as GPT or BERT. This phase is typically implemented using attention-based mechanisms and multi-source fusion techniques.
During this process, the model integrates the retrieved information with the original query by constructing a single input known as a prompt that combines both elements, leveraging its deep linguistic and contextual understanding to produce an output that is linguistically coherent, contextually accurate, and semantically aligned with the retrieved content.

1.5.5 Evaluation Metrics for RAG Systems

We assess the performance of our RAG-based question-answering system within the Algerian legal context, with a particular focus on two critical aspects: the effectiveness of information retrieval and the quality of the generated answers.

Retrieval Metrics

For the evaluation of retrieval performance, the focus is on metrics that effectively measure the relevance, accuracy, diversity, and robustness of the retrieved information.

- **Context precision:** It assesses whether the correct ground truth contexts appear at higher positions in the retrieved list.

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times V_k)}{\sum_{k=1}^{K} V_k}$$

  Where:

  - K is the cutoff rank, representing the top-K retrieved items to consider.

  - V_k is an indicator variable (usually 0 or 1) that reflects the relevance of the item at rank k. For instance, $V_k = 1$ if the item at position k is relevant, and $V_k = 0$ otherwise.

- **Context Recall:** It measures how accurately the retrieved context reflects the ground truth by evaluating its contained facts and claims. It represents the proportion of ground truth claims that are successfully captured within the retrieved context.

$$\text{Context Recall} = \frac{\text{Number of GT claims that can be attributed to context}}{\text{Total number of claims in GT}}$$

  Where:

  - **GT (Ground Truth)** refers to the set of reference or annotated claims that are considered correct and complete based on human judgment or a gold-standard dataset.
  - A **GT claim that can be attributed to context** means the system successfully retrieved evidence (context) that justifies or supports that particular claim.

- **Mean Reciprocal Rank (MRR)** is the average of the reciprocal ranks of the first correct answer for a set of queries.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

  where $|Q|$ is the number of queries and $rank_i$ is the rank position of the first relevant document for the i-th query.

- **F1 Score:** is the harmonic mean of precision and recall. It provides a single metric that balances both concerns, especially useful when there is a trade-off between precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Average Precision (AP):** summarizes a precision-recall (PR) curve into a single value representing the average of all precisions. It approximates the area under the PR curve and ranges between 0 and 1, where a perfect model achieves an AP of 1. Higher values indicate better performance across different classification thresholds.

$$\text{AP} = \frac{1}{R} \sum_{k=1}^{N} P(k) \cdot \text{rel}(k)$$

  Where:

  - R is the total number of relevant documents in the ground truth.

- N is the total number of retrieved documents.
- $P(k)$ is the precision at rank k, calculated as:

$$P(k) = \frac{\text{Number of relevant documents in the top } k}{k}$$

- $\text{rel}(k)$ is a binary indicator function:

$$\text{rel}(k) = \begin{cases} 1 & \text{if the item at rank } k \text{ is relevant} \\ 0 & \text{otherwise} \end{cases}$$

Generation Metrics

In text generation, evaluation goes beyond simply measuring accuracy; it also assesses the quality of the generated text in terms of coherence, relevance, fluency, and alignment with human judgment.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** is a set of metrics used to assess the quality of summaries by comparing them to human-generated reference summaries. It helps measure the content overlap between the generated and reference texts. Different ROUGE variants evaluate various aspects of similarity, such as n-gram overlap (ROUGE-N, ROUGE-W), word subsequences (ROUGE-L, ROUGE-S), and word pair matching between the system-generated and reference summaries.

- **BLEU (Bilingual Evaluation Understudy):** is a metric used to assess the quality of machine-translated text by comparing it to one or more reference translations. It calculates n-gram precision in the generated text relative to the reference text and applies a brevity penalty to prevent excessively short translations. However, BLEU has limitations, as it does not consider the fluency or grammatical correctness of the generated text.

- **BertScore:** It utilizes contextual embeddings from pre-trained transformers like BERT to assess the semantic similarity between generated and reference text. It calculates token-level similarity using contextual embeddings and provides precision, recall, and F1 scores. Unlike n-gram-based metrics, BertScore effectively captures word meaning in context, making it more resilient to paraphrasing and more sensitive to semantic equivalence.

- **Faithfulness:** This metric evaluates whether the generated answer is factually accurate and grounded in the entire retrieved context. An answer is deemed faithful if all of its claims are supported by the retrieved content, and does not introduce hallucinated or unrelated claims. It is calculated by measuring the semantic similarity between the full answer and the full retrieved context using cosine similarity of embeddings.

$$\text{Faithfulness} = \cos(\text{Embedding}_{\text{Answer}}, \text{Embedding}_{\text{Full Context}})$$

- **Answer Relevance:** This metric assesses how topically aligned the generated answer is with the retrieved legal context, by comparing it to each individual chunk (e.g., sentence or paragraph) of the context. A higher relevance score

indicates that the answer covers content discussed in the retrieved material, reflecting the retriever's ability to provide useful and related legal information.

$$\text{Answer Relevance} = \frac{1}{N}\sum_{i=1}^{N}\cos(\text{Embedding}_{\text{Answer}}, \text{Embedding}_{\text{Chunk}_i})$$

1.5.6 Applications of RAG

RAG has deep linguistic understanding capabilities that allow it to provide more reliable and specialized responses. This makes it suitable for many applications across different domains P. Zhao et al., including:

- Question-Answering System: Provide answers in a specific domain using a specialized database.
- Neural Machine Translation (NMT): is the automated process of translating text from one language to another.
- Summarization: It is the process of extracting basic information from long texts.
- Commonsense reasoning: The ability of a machine to make decisions about problems or tasks in a human-like manner.

1.6 Conclusion

In conclusion, this chapter explored fundamental concepts that are essential for understanding and addressing tasks related to Retrieval-Augmented Generation (RAG) and Legal Question Answering (QA). In addition to defining key terms, we examined widely used tools and techniques relevant to our work, including large language models (LLMs), QA system architectures, embedding models, and retrieval mechanisms. We also discussed evaluation metrics commonly used to assess the performance of such systems. In the next chapter, we examine in detail the state of the art of Legal Question Answering.

Chapter 2

State Of The Art

2.1 Introduction

This chapter presents a collection of works related to our topic, starting with research papers on legal Question Answering (QA) systems. These articles are categorized into four approaches: traditional methods, machine learning (ML), deep learning (DL), hybrid approaches, and additional specialized techniques for QA. Furthermore, we review various datasets that have been specifically developed for Question Answering in the legal domain.

2.2 Traditional Approaches for Legal QA systems

In this section, we present traditional approaches used in Legal Question Answering (QA) systems. These methods include ontology-based reasoning, term weighting techniques like TF-IDF and BM25, and statistical co-occurrence analysis.

2.2.1 Ontology-Based Methods

Existing approaches to textual entailment and question answering often lack the legal knowledge and reasoning capabilities required for deep entailment tasks, particularly the ability to provide logical justifications for answers. Kourtin et al.addresses the challenge of enabling machines to process background legal information and reason with it to determine the truth value of implicit (Yes/No) questions. The research is motivated by the need to develop a semi-automatic tool for generating criminal law ontologies and legal reasoning rules, aimed at supporting automated legal question answering systems that can address questions from the USA bar examination. The primary objective is to construct a legal ontology and an accompanying set of rules using an 18-step methodology designed to extract and represent all information necessary for legal analysis. These steps include identifying competency questions, extracting legal concepts, categorizing relevant nouns, distinguishing between classes and instances, building class hierarchies, and identifying both atomic and definable classes. Additional stages involve defining object

and datatype properties, specifying domains and ranges, and formalizing definable classes through OWL axioms and SWRL rules. The process integrates various techniques such as natural language processing for text analysis, the Stanford parser for semantic triple extraction, and tools like WordNet, dependency parsing, natural logic, and OpenIE for relation extraction. The semi-automated tool successfully implements 15 of the 18 steps and was evaluated through task-based assessments, competency questions, and ontology evaluation tools such as the Pellet reasoner and OOPS. Development and testing, conducted on a set of 12 multiple-choice questions and validated on 4 additional ones, demonstrate the feasibility and incremental improvement of the system. Error analysis has traced performance issues to specific stages of the methodology, guiding further refinement.

2.2.2 TF-IDF-Based Methods

Developing a question answering system for non-English languages such as Vietnamese presents significant challenges, primarily due to the limited availability of linguistic tools and resources. However, with the growing volume of Vietnamese legal documents available online, efforts have been made to build systems tailored to this domain. For instance, Duong & Ho,proposed vLawyer, a state of the art question answering system specifically designed for Vietnamese legal texts. This work aims to enhance vLawyer's information retrieval accuracy by information extraction techniques.

vLawyer leverages existing tools such as Apache Lucene[1] is a powerful search library, a high-performance text search engine, for indexing and searching. It also employs term frequency-inverse document frequency (TF-IDF) to assess the significance of terms. The vLawyer system utilizes a similarity-based model to retrieve relevant legal texts. It operates by extracting candidate passages, constructing a term-document matrix, and applying cosine similarity within the Latent Semantic Indexing (LSI) space to identify the most relevant documents. vLawyer demonstrates impressive performance, achieving approximately 70% accuracy in retrieving legal documents.

2.2.3 BM25-Based Methods

In the study (Rosa et al., 2021)., the effectiveness of a classical retrieval model, BM25, is evaluated for legal case retrieval using the COLIEE 2021 shared task dataset. BM25 (Best Matching) is a widely used probabilistic ranking function in information retrieval that scores documents based on term frequency and inverse document frequency with length normalization Lin et al.. Instead of using sophisticated or computationally expensive neural models, the authors show that a straightforward, well-executed BM25 approach can achieve competitive results in identifying relevant supporting legal cases To address this, each legal case is segmented into overlapping chunks of 10 sentences, with a gap of 5 sentences in between. Segmenting enables finer-grained matching of queries and candidate cases. For every new legal query, BM25 similarity scores are calculated between the chunks of the new case and those of the candidate cases. Then, each candidate is labeled

[1] `https://lucene.apache.org/`

with its best rank in all comparison, best representing the best fit. The algorithm sorts such instances by top ratings and selects top ones by hardcoded thresholds. General retrieval is achieved using Pyserini Lin et al. (2021). Python interface of Anserini, operating on the Lucene search engine. Unaccountably, this simple setup worked remarkably well. The BM25 based system ranked second in COLIEE 2021. It achieved an F1-score of 0.0937 on the test set, with a precision of 0.0729 and a recall of 0.1311 — all well above the median F1-score of 0.0279 reported for other submissions. These are the results based solely on the first segments of each base case, a wise choice to minimize computational load. What is remarkable about this research is the success of a traditional method in a world where neural models increasingly dominate. The results show that if used with proper segmentation and scoring techniques, old models like BM25 are still able to deliver good performance. However, the approach has its limitations, especially scalability since it is computationally expensive to compare all document segments.

2.2.4 Co-occurrence-Based Methods

Legal professionals usually face the daunting task of scanning huge volumes of unstructured legal documents to find precise answers. This process is time-consuming, mentally taxing, and prone to errors, especially when speed is essential. Martinez-Gil et al., aims to fill that gap by offering a system that significantly improves the speed and accuracy of legal information retrieval. By delivering quick and accurate answers to legal queries, the proposed solution helps alleviate the cognitive load on legal experts and enables better decision-making. The centerpiece of this system is a clever method called reinforced co-occurrence, designed to reveal underlying relationships between legal sentences. By examining how particular terms frequently co-occur throughout large legal documents, the system can identify patterns that may not be immediately obvious. These patterns are then used to retrieve the most suitable information to respond to legal queries.

The system operates in a fixed pipeline. It analyzes the input question first, then goes ahead to scan the legal corpus for determining important co-occurrence patterns, next extracts likely answers, normalizes them to ensure consistency, and ranks them to indicate the most accurate and relevant responses.

For the optimization of the system's efficiency, several advanced techniques are utilized. These include text mining and heavy preprocessing tasks such as stop word elimination, lemmatization of words, and removal of frequent adjectives and verbs that add negligibly to value. Normalization is also utilized to handle outliers and deliver high-quality outputs throughout the process. The system was experimented on a legal question dataset and achieved a 65% accuracy rate, outperforming baseline and non-ML methods. Despite this success, the paper highlights some limitations: the system's performance is greatly dependent on the quality of the legal corpus, it is time-consuming to adjust its parameters.

2.3 ML Approaches for Legal QA systems

Classic machine learning can be used for legal question answering by applying traditional NLP techniques and statistical models. Methods like TF-IDF and word embeddings help extract important legal terms, while machine learning models such as SVM are used to classify legal questions or rank relevant documents . In addition, information retrieval techniques improve search results. Some approaches also combine rule-based systems with machine learning for better accuracy. However, these methods can struggle with the complexity of legal language and may not perform as well as deep learning models.

2.3.1 Support Vector Machines (SVM)

To effectively answer yes/no questions from the Japanese Bar Examination, Kim, Xu, Lu, & Goebel developed an advanced legal information retrieval and question-answering system. This system combined TF-IDF and SVM techniques to analyze yes/no questions. The goal was to improve legal information retrieval performance by integrating different methods. The authors also evaluated the system's effectiveness compared to previous standards, achieving high accuracy. The first stage is the information retrieval stage, where the system used the TF-IDF model to retrieve the most relevant articles for a given query. Then the SVM re-ranking model is used to determine the importance of additional features. The second stage is semantic analysis and dependency analysis, which means confirming the yes or no answer by analyzing and understanding a query. The techniques used in this stage are reformulation to expand the query and semantic analysis through the process of query matching embeddings to improve retrieval performance. The third stage is combining the previous methods by combining the obtained results. The SVM model effectively improves accuracy compared to previous methods, achieving, on the experimental dataset, an accuracy of 62.14%, 55.71% for the second phase, and 55.79% for the third phase.

The project by Hoshino et al. focuses on building a legal question answering system that can handle true/false questions from the Japanese bar exam, especially for Task 4 of the COLIEE 2018 shared task. The main challenge is that legal language is often tricky. Legal texts use words that are different from those found in the questions, even when they mean the same thing. Also, many legal questions can't be answered by just reading the articles—they require extra knowledge or common sense, and the vocabulary used in legal documents is a mix of technical terms and everyday language. One of the biggest goals was to build a system that doesn't just give the right answer but can also explain why that answer is correct in a way people can understand. This is important in the legal field where decisions need to be traceable and justifiable. Since the dataset is small (only a few hundred questions), using pure machine learning methods is difficult. That's why this system takes a more rule-based approach, while also including some machine learning techniques. The system is designed to analyze the structure of legal sentences. It uses Predicate Argument Structure (PAS) to break down legal texts and understand the relationship between actions (predicates) and the people or objects

involved. A special synonym dictionary was also created to handle different words that mean the same thing, depending on the context. For example, if two verbs are used differently but refer to the same action with the same object, they can be treated as synonyms. Several modules were used to judge whether a statement is true or false. Each module checks the match between the question and the article in different ways—some focus on exact matches, while others allow for more flexible matching. A machine learning model (SVM) is used to combine these modules and choose the most confident answer based on how many articles support each one. There's also a person estimation feature that helps the system figure out who is doing what in the story-like questions, which makes the reasoning more accurate. The system achieved second place in the COLIEE 2018 shared task showing strong performance. It was especially good at solving simpler questions where the wording of the problem closely matched the legal article—getting over 70% accuracy in those cases. One of the system's modules, called KIS Frame (Knowledge Integration System Frame), gave the best results among all entries in the official run. While the SVM-based combination method worked best overall, its results did vary depending on the year and the types of problems.

2.3.2 Ensemble Methods Combining SVM and KNN

Predicting judicial decisions is a complex and critical task in the legal domain, particularly for the European Court of Human Rights (ECtHR), where the large volume of legal texts poses challenges for manual case analysis. Alcántara Francia et al. focuses on automating the prediction of ECtHR case outcomes using machine learning algorithms. Unlike prior general overviews, the research employs classical text mining techniques—including Bag-of-Words, TF-IDF, and n-gram models—to transform legal documents into structured data for classification.

The models evaluated in this study include Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). SVM was applied with a linear kernel to identify the optimal hyperplane separating the classes in the high-dimensional feature space. KNN classified cases based on the majority class among the k nearest neighbors in the feature space, with k optimized through cross-validation.

Performance was assessed using accuracy and other metrics to determine the most effective approach for this dataset. The results indicate that SVM models achieve the highest accuracy, reaching up to 79%, and in some cases, surpass human expert predictions. These findings highlight the potential of machine learning to enhance efficiency and consistency in predicting legal decisions within the ECtHR context.

2.4 Deep Learning approach for Legal QA systems

Due to the complexity and linguistic density of legal texts, traditional methods are no longer sufficient to provide accurate and reliable answers.

With the advent of deep learning, which analyzes the semantic relationships

between questions and legal texts, has significantly improved the accuracy of legal question answering. Later, transformer models revolutionized the field with their ability to understand context and analyze complex legal texts with greater precision.

2.4.1 Neural Network Architectures in Legal QA

Convolutional neural networks (CNNs) and long short-term memory (LSTMs) are among the first deep learning models to be used in question answering (QA) systems. In this context, Kim, Xu, & Goebel (2017) presented a yes/no legal question answering system using convolutional neural networks (CNNs), targeting questions from the Japanese bar exam. The system consists of three stages: retrieving legal articles using TF-IDF and SVM models; analyzing textual entailment between the question and the article using a CNN enriched with word embeddings (word2vec) and linguistic features such as syntactic roots and negation, and then integrating both stages into a unified framework.
The results showed that the CNN model outperformed traditional methods, achieving an accuracy of 63.87% in textual inference, and the complete system also won first place in the 2015 COLIEE competition.

On the other hand, ADEBAYO et al. presented a system that uses LSTM models, specifically Child-Sum Tree LSTM, to improve legal question answering systems. This model is based on a tree structure that takes into account the relationships between words in a sentence, allowing for a deeper understanding of semantic meaning. Traditional models such as SVM and Random Forest were also tested to compare their performance with deep networks. LSTM has proven effective in understanding legal texts that require in-depth reasoning due to its ability to handle long contexts. This reflects the importance of deep neural networks in improving the performance of legal question answering systems.

2.4.2 Transformer-based Models in Legal QA

Transformer models are considered one of the most important recent developments in the field of answering legal questions, as they enable a deeper understanding of complex legal texts through a self-attention mechanism, unlike traditional methods that rely on retrieving information using keywords.

Among the works that have been completed within the framework of Transformer-based models, the research presented by H.-T. Nguyen et al., which focuses on evaluating the performance of GPT-3.5 and GPT-4 models in legal reasoning tasks using COLIEE Task 4 data, which includes legal questions in English and Japanese formulations.
The study evaluated the performance of GPT-4 and GPT-3.5 in both monolingual and cross-lingual settings. The results showed that GPT-4 outperformed GPT-3.5, particularly in monolingual contexts, while both models faced challenges in cross-lingual scenarios due to linguistic and cultural differences in legal texts. These findings highlight the promise of GPT models in generating accurate legal responses and contribute to the advancement of artificial intelligence applications in the legal domain.

On the other hand, T.-M. Nguyen et al. presented research aiming to enhance the accuracy of legal document retrieval and answering legal questions by leveraging BERT's advanced capabilities in natural language understanding and legal text processing. BERT was trained to create accurate representations of legal questions and articles and was employed in sentence classification and extracting answers for various types of questions. In addition, BERT is integrated with classification models such as LightGBM to improve the ranking of results and increase the accuracy of predictions.
The results showed that BERT had a significant impact in improving the results, as the system achieved first place in document retrieval with an F2 score of 0.94, and ranked second in the question answering test with an accuracy rate of 67%. These results confirm the strength of BERT as one of the most prominent transformer models in legal text processing, with the potential to be improved by training it more deeply on specialized legal data.

A different approach was proposed by Askari et al., who addressed key challenges in retrieving answers in the legal field, namely the large knowledge gap between legal experts and ordinary users, as well as the inconsistent use of formal or informal language among users on legal platforms. These challenges make it difficult for traditional retrieval methods to perform optimally in the legal context.
To help address these limitations, the authors introduced a new benchmark dataset, LegalQA, which includes 9,846 legal questions and 33,670 expert-verified answers; those answers were provided by verified lawyers. The research proposed a new method, CEFS (Cross-Encoder with Fine-grained Structured inputs), which can help with re-ranking by providing fine-grained structured inputs of each of the question's inputs (subject, description, tags), each separated by special tokens.
The system uses a two-part retrieval pipeline. The first part is BM25 based on the candidate answers lexically. The second part is the CEFS, which re-ranks the candidates based on the inputs of the structured question. The CEF model is fine-tuned using the MS MARCO dataset and the LegalQA training dataset. The location-based model (LMD) is also mathematically analysed (as an alternative for the retrieval pipeline's first stage) against the BM25 model.
For the experimental findings, CEFS improves performance, presenting a MAP@1k score of 0.270, compared to 0.236 where CECAT (trained on LegalQA), and 0.109 for MiniLM-MSMARCO. In an ablation study, the question description added most to CEFS's performance, followed by the subject, and then the tags. Even though BM25 performed better than LMD for the retrieval pipeline's first pass, overall BM25 is still a low performer.

2.4.3 Retrieval-Augmented Generation in Legal QA

Researchers have increasingly turned to hybrid approaches that combine retrieval and generation mechanisms to mitigate the limitations of language models in the legal context.
RAG has demonstrated significant effectiveness in improving the accuracy and contextual relevance of responses.

In this vein, Wiratunga et al. presented a system that combines case-based reasoning (CBR) with retrieval-augmented generation (RAG) with the aim of improving the accuracy of answers provided by large language models (LLM) in the

field of answering legal questions. The proposed approach, CBR-RAG, seeks to enhance the retrieval process by using a database of legal cases, where each legal case is represented by its basic components: question, supporting text, legal entities, and answer.
This approach was tested using three types of text representations: BERT, LegalBERT, and AnglEBERT, and three retrieval strategies were adopted: intraclass retrieval, interclass retrieval, and hybrid retrieval.
The results showed that CBR-RAG outperforms traditional RAG in retrieval accuracy and improving the quality of legal answers, as it performed best when using hybrid retrieval with AnglEBERT (k=3), which resulted in an increase in accuracy by 1.94% compared to other methods.

Among the RAG-based systems, Kalra et al. proposed HyPA-RAG to address the shortcomings of large language models (LLMs) in handling legal queries. HyPA-RAG consists of three main components: First, a Query Complexity Classifier (QC) that dynamically adjusts RAG parameters to minimize token usage without sacrificing accuracy. Second, it relies on a hybrid retrieval strategy that employs multiple techniques: dense retrieval based on semantic similarity, token retrieval using BM25 for keyword matching, and knowledge graph-based retrieval to exploit the semantic structure between legal entities. Finally, a specific evaluation framework was developed to measure the accuracy of responses and their relevance to the legal context. Experimental results, particularly on queries related to New York City's LL144 law, showed that HyPA-RAG offers superior performance in terms of accuracy and coherence with legal texts, highlighting the importance of intelligently combining Transformer and multiple retrieval techniques in this sensitive field.

Islamic legal rulings guide Muslims' daily lives, generating many questions that require expert Muftis. However, with Muslims comprising about 25% of the global population and a limited number of qualified Muftis available, the demand for reliable fatwas exceeds supply. This gap motivates the development of AI-based automated Question-Answering systems to provide accurate and efficient responses. One such system isAlotaibi et al. (2022).
KAB is an advanced query answering system specifically designed to address criminal jurisprudence questions in Islamic law. This system combines effective retrieval techniques with prior knowledge, utilizing trusted and authoritative Islamic jurisprudential references to provide accurate and reliable answers. KAB It processes the input query by using an existing FAQ database to match it with previously answered questions and retrieve the relevant metadata. Implementing a generative model to generate solutions based totally on consumer input, KAB leverages a massive dataset to teach it and make sure its accuracy in answering. KAB used to be carefully evaluated the use of BERTScore and METEOR, attaining precision (0.6), recall (0.4), F1 rating (0.48), and METEOR rating of 0.037. These metrics spotlight its capability to grant applicable and in-depth answers.

2.5 Datasets for Legal Question Answering

Legal question answering is a difficult task because it involves complex rules and specific legal language. To help with research in this area, several datasets have been created. In this section, we describe some of the main datasets used for legal question answering and explain how they were built.

JEC-QA

The researchers introduced JEC-QA, the largest legal-domain question-answering dataset. This dataset addresses the challenges of answering legal questions, which require complex reasoning and deep comprehension.

The methodology involved constructing a knowledge base with 26,365 questions sourced from official legal examinations and authoritative texts, including the National Judicial Examination in China, the National Unified Legal Professional Qualification Examination Counseling Book. To improve retrieval accuracy, the researchers employed information retrieval techniques such as ElasticSearch and BERT-based topic classification.

Seven baseline models were evaluated, and the best-performing model achieved only 28.63% accuracy, while human performance 64.21% accuracy for novices and 81.12% accuracy for experts, highlighting a significant gap in the models' ability to perform multi-step reasoning and understand legal concepts. JEC-QA provides a strong foundation for future research to enhance models' capabilities in addressing complex legal questions Zhong et al..

LLeQA: Long-form Legal Question Answering

The main challenge addressed in the research done byLouis et al. is the absence of specialized datasets for lengthy legal questions since most legal question answering systems are limited to short and inadequate answers, while the legal domain demands more in-depth analysis and complicated reasoning founded on elaborate legal texts.
The LLeQA dataset was created with the aim of facilitating research in legal AI and making it possible to create systems that can give complete and correct legal answers. The dataset strives to enhance the comprehension of linguistic models of legal documents by offering abundant data for training and testing deep learning models, thereby working to make legal information more accessible to nonprofessionals.
The LLeQA dataset is made up of 1,868 French legal questions that were created and annotated according to the most frequent legal questions in Belgian law. Every question has a corresponding annotated long answer from 27,942 legal articles in Belgian legislation. The dataset also contains annotated paragraphs that refer to the legal sources that back up each answer, which increases the validity of the interpretation of the legal models that have been applied.

CUAD: AnExpert-Annotated NLP Dataset for Legal Contract Review

It is expensive to obtain annotations from legal experts, and many law firms spend time reviewing contracts. The Contract Understanding Atticus (CUAD) dataset was developed specifically to evaluate legal contracts and also, as a bench-

mark dataset for training and evaluation, the set features a variety of agreements that are accurately described. The dataset consists Hendrycks et al. of more than 500 contracts and more than 130,000 expert annotations across 41 categories, enabling better training and evaluation of natural language models in the legal domain. The CUAD dataset is used by legal experts to ensure quality, and to apply different Transformers models to legal question answering tasks. Preliminary results from experiments conducted using natural language models have shown better performance, with models such as DeBERTa-xlarge achieving 80% accuracy, 44% recall, and 47.8% area under the precision-recall curve (AUPR). This suggests that while there is room for improvement, the results are promising and suggest that CUAD could serve as a valuable benchmark for advancing legal NLP research.

FALQU: Finding Answers to Legal Questions

Many existing datasets suffer from a lack of diversity in the questions and domains they ask, such as COLIEE-2015 and JEC-QA. To overcome this, the authors Mansouri & Campos present the FALQU dataset, a new test set based on the Law Stack Exchange (LawSE), which contains realistic legal questions from different domains and countries.
The database contains 9,880 questions and 34,145 answers, extracted from 24,187 questions in LawSE after removing duplicates and invalid questions. Divided into a training set (8,892 questions) and a test set (988 questions), FALQU was evaluated by several information extraction models including TF-IDF, BM25, and BERT. This research provides an open dataset that can be used to develop legal answering systems based on answer generation rather than just answer retrieval.

2.6 Conclusion

This chapter provided a comprehensive overview of research on legal question-answering, covering studies that use traditional, hybrid, and deep learning approaches. Key legal QA datasets were also highlighted for their significant role in advancing the field. In the next chapter, we will explore experiment and implementation.

Chapter 3

Proposed Solution and Experiments

3.1 Introduction

This chapter outlines the methodology and experimental framework adopted for the development and evaluation of the Tax-RAG question-answering system, specifically designed to process Arabic legal texts. The discussion begins with a detailed description of the system development process, covering document preprocessing stages, the retrieval mechanism, and the integration of pre-trained large language models for answer generation. Examples of expected output are included, along with a discussion of system limitations and the evaluation metrics employed. Subsequently, the construction of a legal benchmark used for performance assessment is examined, highlighting its significance, structural composition, and the procedure followed to collect and validate question–answer pairs.
Finally, the experimental results are presented and analyzed, addressing both retrieval and generation performance, and offering key insights derived from the evaluation outcomes.

3.2 Development process of Tax-RAG Question Answering System

Building an effective RAG-based system to answer tax-related questions requires a carefully structured approach that combines document processing, retrieval, and answer generation models. In our application, we focused on improving each stage to ensure accuracy, efficiency, and suitability in tax inquiries. The following is a detailed explanation of the process.

3.2.1 Document Processing and Retrieval

The foundation of any RAG system lies in its ability to process and retrieve relevant information efficiently. In our case, we dealt with both tax documents

and legal texts, which present unique challenges due to their complexity and frequent updates. To tackle these challenges, we implemented a multi-stage document processing pipeline:

Document Loading & Preparation

Primary sources, such as the Algerian Tax Code of 2025, were collected and converted from PDF to a clean and structured text format. This process removed images and inconsistencies, standardizing the text to ensure clarity and consistency.

Text Splitting & Embedding

The legal texts were segmented at the article level, aligning with the structure of Algerian legal documents (e.g., "Article 1:", "Article 2:"). Article-based splitting preserves semantic boundaries and ensures meaningful chunking. The resulting segments were then embedded using multilingual models to generate dense vector representations. These embeddings were converted to float32 format, normalized using L2 normalization to allow cosine similarity, and stored in a FAISS [1] index (IndexFlatIP) that resides in memory for efficient similarity-based retrieval (Johnson et al., 2019; Douze et al., 2024).

Retrieval Optimization

To achieve maximum accuracy, our retrieval method combines BM25 for exact lexical matching and Jaccard for capturing basic semantic overlap with more advanced semantic models like E5-large or Arabic-Retrieval. This combination allows the system to balance precision and contextual understanding, resulting in strong performance across a wide range of legal queries.

Table 3.1 provides an overview of the embedding models utilized in our retrieval pipeline, including the dimensionality of the vector representations generated by each model.

[1] `https://github.com/facebookresearch/faiss`

Table 3.1: Embedding models used in the retrieval phase of our system.

Embedding Name	Model	Dimension
AraBERTv2	`aubmindlab/bert-base-arabertv02` (**?**)	768
Arabic-Retrieval-v1.0	`omarelshehy/Arabic-Retrieval-v1.0` (Reimers & Gurevych, 2019)	768
E5-Large	`multilingual-e5-large` (Wang et al., 2024)	1024
MiniLM	`all-MiniLM-L6-v2https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`	384
legal-embedding-model	`hiieu/halong_embedding` (Reimers & Gurevych, 2019)	384
E5-Small	`multilingual-e5-small` (Wang et al., 2024, 2022)	384

3.2.2 Response Generation Using Pre-trained LLMs

The final stage of the pipeline involves answer generation using a pre-trained Large Language Model (LLM). The retrieved articles are combined into a carefully designed prompt to limit the model's output to legally-based information. This step aims to generate coherent and contextually accurate responses that preserve the legal phrasing and intent, allowing users to access precise legal insights without directly consulting the original legal texts.

Context Construction & Prompt Design

In this system, the retrieved articles are first cleaned to remove excess whitespace and formatting inconsistencies. A legal reference number is then assigned to each article to facilitate tracking. The articles are subsequently compiled and organized to form a cohesive legal context. The **format_context()** function ensures that the context remains within an optimal display length, preserving the focus and relevance of the content.
A structured Arabic prompt template is used that explicitly directs the model to build answers based on the context provided. This prompt integrates the user's question with the context using a standardized Arabic format tailored to the Algerian legal domain. For example, the prompt begins with a directive such as:
أجب بدقة عن السؤال المطروح بالاعتماد فقط على المواد القانونية التالية:
followed by the retrieved legal articles and the user's question. This structured formulation ensures that the model interprets the query within a legal framework and generates responses that align with the tone, structure, and intent of formal legal discourse.

LLM Integration into the RAG Pipeline

Several large language models, including AraBERT and DeepSeek, as listed in the table 3.2, were tested as part of the development of a Retrieval-Augmented Generation (RAG)-based system for answering legal questions. These evaluations were conducted to select the most suitable model capable of balancing linguistic fluency with strict compliance with the legal context.

Table 3.2: Generation models used in our system.

Generator Name	**Model**	**Dimension**
AraGPT2-base	`aubmindlab/aragpt2-base` Antoun et al. (2021)	768
AraBERTv2	`aubmindlab/bert-base-arabertv2`	768
DeepSeek-Coder 1.3B	`deepseek-ai/deepseek-coder-1.3b-base` `https://huggingface.co/deepseek-ai/deepseek-coder-1.3b-base`	2048
BERT Multilingual QA	`henryk/bert-base-multilingual-cased-finetuned-polish-squad1` `https://huggingface.co/henryk/bert-base-multilingual-cased-finetuned-polish-squad1`	768
XLM-R Large QA	`AlexKay/xlm-roberta-large-qa-multilingual-finedtuned-ru` `https://huggingface.co/AlexKay/xlm-roberta-large-qa-multilingual-finedtuned-ru`	1024

The selected LLM processes input after a preprocessing stage involving tokenization and the application of attention masks, ensuring that the model attends to the most relevant parts of the input sequence. The answer generation is performed by the LLM, using the beam search decoding technique [2], which help in selecting the most likely sequence of words from multiple possible outputs. To maintain response naturalness and avoid repetitive phrases, n-gram repetition penalties are applied during decoding.

Following generation, the resulting text undergoes a final and meticulous post-processing phase aimed at refining the quality of the answer. This includes removing any remnants of the original prompt, normalizing Arabic punctuation and whitespace, and filtering out any special characters that may be automatically generated by the models. The **postprocess_answer()** function plays a pivotal role in this stage, ensuring that the output aligns with professional standards for legal answers.

Overall, the generation module was designed to maintain a balance between the expressive fluency of modern Arabic models and the precision required by legal content, which is essential to ensure the reliability of answers in contexts such as legal and tax consultations.

Figure 3.1 provides a visual representation of the architecture used in our Tax-RAG system.

[2] `https://telnyx.com/learn-ai/beam-search-algorithm`

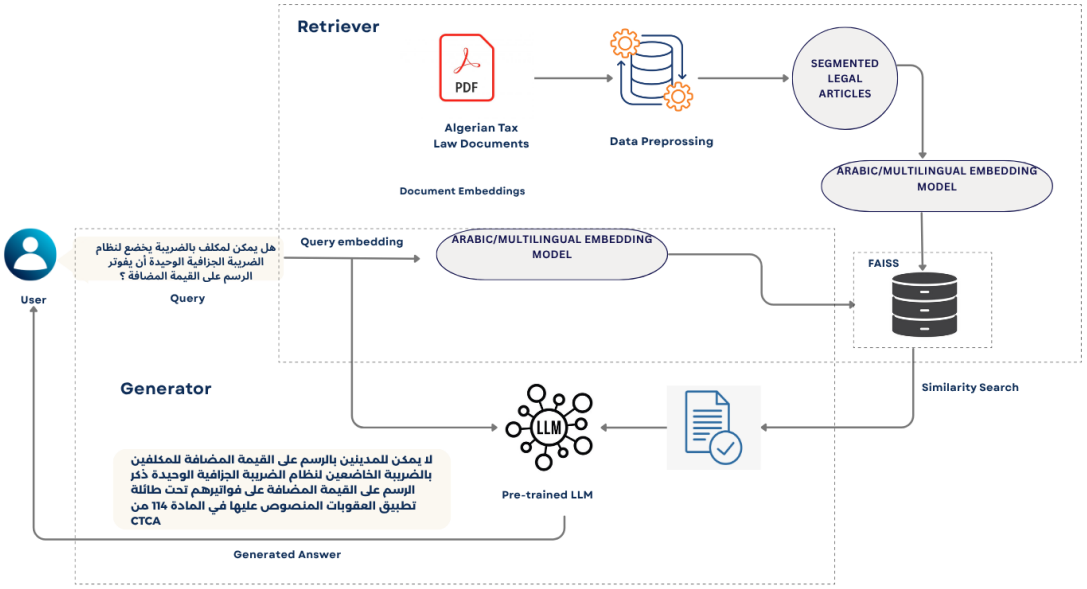

Figure 3.1: Architecture Diagram of Our RAG System.

3.3 Construction of the Legal Benchmark

In any Question Answering system, especially in the legal domain, data quality is paramount. Regardless of how advanced the models may be, they cannot give accurate answers if the data is incorrect or not well-structured. In this section, we will describe the steps taken to construct our Legal Benchmark Dataset that serves as the basis for evaluating our system for the Algerian legal context. Our goal is to create a dataset that not only reflects the complexity of legal documents, but also ensures that the information is accurate, relevant, and applicable to the real world.

3.3.1 Purpose and Importance of the Benchmark

Given the complexities and specificities of legal legislation, establishing a specialized legal standard is an essential step toward developing a quality assurance system tailored to the Algerian context.
This standard aims to provide a solid basis for assessing system quality. By presenting a carefully selected set of Algerian legal texts, accompanied by precise and contextually relevant questions and answers, this standard ensures that the system's assessment is realistic and rigorous. It also allows us to measure the system's ability not only to retrieve relevant information but also to accurately interpret complex legal language.

3.3.2 Benchmark Components and Structure

We designed a benchmark that comprised three core components to support a comprehensive and reliable evaluation of AI models in the legal domain. The benchmark included the following columns:

- **Questions:** It contains a set of questions formulated to cover a wide range of topics in income tax law.
- **Reference answers:** Expert-authored answers that serve as the gold standard for evaluation. These responses provide a basis for evaluating system-generated answers in terms of legal accuracy, completeness, and coherence.
- **Relevant Legal Articles:** It includes, for each question, legal materials from enacted legislation, directly linked to the answer based on expert annotation. This component is used to assess the system's ability to retrieve contextually appropriate legal content in response to a given question.

3.3.3 Sources of Legal Information

To ensure the accuracy and credibility of the dataset, legal texts were collected exclusively from trusted and official sources, including:

1. **The website of the Algerian Directorate General of Taxes (DGI)**[3]: Used to obtain up to date information on personal income tax, particularly regarding salaries and wages, as well as official explanatory materials provided for individual taxpayers.
2. **Publicly available legal documents and reports in PDF format**[4]: These sources were carefully reviewed and verified for authenticity before inclusion in the dataset.

We focused on laws related to income tax, as this topic is both widely relevant to citizens and rich in legal details suitable for QA training.

3.3.4 Question/Answer Formulation and Validation

To formulate the questions, our primary objective was to simulate a broad range of queries that might be posed by laypersons and professionals in the field of tax law. To achieve this, we employed a multi-step process:

- **Question Drafting:** We began by brainstorming a variety of questions, taking into account different viewpoints. This step involved thinking from the perspective of the average individual seeking to understand the basics of tax law, as well as the perspective of someone with specialized knowledge in tax law. we also referred to existing questions from legal forums, official documents, to ensure relevance and coverage of real-world information needs.
- **Answer Extraction:** After formulating the questions, we focused on collecting accurate and reliable answers. This stage included researching reliable sources, legal frameworks, and expert insights to ensure answers are factually accurate and contextually appropriate.

[3] `https://www.mfdgi.gov.dz/particuliers/irg-traitements-salaires`
[4] `https://www.mfdgi.gov.dz/legislation-fiscale-ar/codes-fiscaux-ar#512-776-2025`

- **Expert Review:** Finally, to verify the accuracy and relevance of the questions and answers, we requested the opinions of four experts in the field of tax law.

Table 3.4 describes the key statistics of the benchmark, providing insight into the structure and variability of the dataset. These statistics help to evaluate the complexity of the questions, the scope of reference articles, and the richness of the answers.
Additionally, Table 3.3 presents an illustrative example consisting of a few sample questions, their corresponding answers, and related reference articles to clarify the structure of the benchmark.

Table 3.3: Descriptive statistics of the Q/A benchmark in the tax law domain.

Benchmark Parameter	**Value**
Number of Questions	100
Question Length (min / max / avg)	5 words / 35 words / 16.7 words
Relevant Legal Articles per Question (min / max / avg)	1 / 5 / 2.3
Answer Length (min / max / avg)	19 words / 97 words / 41.5 words

Table 3.4: Example illustrating the structure of the benchmark dataset.

Question	**Relevant articles**	**Reference Answer**
هل التشريع الجبائي يسمح للإدارة الجبائية حق الاطلاع قصر تأسيس وعاء الضريبة؟	المادة 45 من قانون الإجراءات الجبائية	يسمح التشريع الجبائي لأعوان الإدارة الجبائية حق الاطلاع قصد تأسيس الضريبة ومراجعتها بتصفح الوثائق المحاسبية المادة 45 من قانون الإجراءات الجبائية
ماهو أجل تسجيل و دفع الحقوق؟	المواد 58، 81 و84 من قانون التسجيل	أجل التسجيل: يجب أن تسجل العقود التي تتعلق بالشركات خلال شهر من تاريخ إبرامها. رسوم العقود تدفع قبل التسجيل، لكن يمكن تقسيمها على ثلاث دفعات إذا قُدمت ضمانات كافية. تُدفع الدفعة الأولى عند التسجيل، والدفعتان المتبقيتان خلال عشرين يوماً من كل استحقاق سنوي، مع زيادة فائدة بنسبة 5%.
ما هي طرق الطعن التي يمنحها القانون للمكلف بالضريبة؟	المواد 71، 79، 80، 81 مكرر، 82، 89 من قانون الإجراءات الجبائية	يبدأ المكلف بالطعن بشكوى نزاعية لدى الإدارة الجبائية. إذا تم رفضها كلياً أو جزئياً، يمكنه الطعن أمام لجان الطعن، ثم إذا لزم الأمر، أمام المحكمة الإدارية بواسطة دعوى قضائية.
ما يُقصد بالحدث المنشئ للرسم على القيمة المضافة؟	المادة 14 من قانون الضريبة على القيمة المضافة	الحدث المنشئ للرسم هو الواقعة التي ينشأ عنها الدين الضريبي. يختلف حسب نوع العملية: بالنسبة للبضائع هو التسليم القانوني أو المادي، وبالنسبة للخدمات هو قبض الثمن. أما في حالة الصادرات أو الواردات فيرتبط بالإجراءات الجمركية، بينما في الصفقات العمومية يعتمد على تحصيل الثمن أو مرور سنة من التسليم.

3.4 Experimental Results and Discussion

In this section, we present and analyze the experimental results of applying our RAG-based system to the Algerian tax law corpus. We assess the effectiveness of both the retrieval and generation components using appropriate evaluation metrics and discuss the main observations derived from the results.

3.4.1 Development Environment

We present the tools and environment used to carry out our experiments, highlighting the key resources that supported our work and ensured reliable and consistent results.

Google Colab

Google Colab, short for Google Colaboratory, is a cloud-based platform that makes it super easy to write and run Python code right in the browser. Built on the Jupyter Notebook interface, it is especially popular among students, researchers, and data science enthusiasts because it is free to use and offers access to powerful computing resources—including GPUs and TPUs—without needing any special hardware. Colab is perfect for machine learning and data science projects, making it easier to experiment, visualize data, and collaborate with others in real time. Whether you are just starting out or working on advanced ML models, Google Colab provides a flexible and accessible environment that helps bring ideas to life.

Python

Python[5] is a powerful and easy-to-learn programming language known for its simplicity and readability. It lets developers focus on solving problems without getting bogged down by complex syntax. Python is flexible and widely used in many areas, from web development to data science, making it a popular choice for both beginners and experienced programmers.

Python Libraries

`Transformers`[6] is built to follow the typical NLP machine learning pipeline: data processing, model application, and prediction. While the library offers tools for training and development, the report by Wolf et al. (2020) focuses mainly on the core model specifications.

`Pandas`[7] The pandas library, in development since 2008, aims to bridge the gap between Python—an all-purpose language for systems and scientific computing—and specialized statistical computing platforms and database languages. While it

[5] `https://www.python.org/`
[6] `https://huggingface.co/docs/transformers/index`
[7] `https://pandas.pydata.org/`

provides comparable functionality to these tools, pandas goes beyond by offering unique features like automatic data alignment and hierarchical indexing, which are not typically available in other libraries or computing environments in such an integrated form.

`NumPy`[8] is the fundamental package for numerical computing in Python. It provides support for large multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

`PyTorch`[9] is an open-source machine learning framework that provides a flexible deep learning model building environment. It is used for tensor computation, automatic differentiation, and neural network implementation.

`Scikit-learn` is a Python machine learning library that includes simple and efficient tools for data mining and data analysis, such as algorithms for classification, regression, clustering, and dimensionality reduction.

3.4.2 Retrieval Performance Evaluation

In this section, we evaluate the performance of various retrieval models on an Arabic legal question-answering task, focusing specifically on the context of Algerian tax law. We compare three types of models.

Sparse Retrieval (Keyword-Based Methods) includes models that rely on exact term matching between the query and the documents. BM25 is a probabilistic ranking function that assigns higher scores to documents containing query terms with higher frequency, making it effective for keyword-based retrieval. Another method, Jaccard Similarity, measures the overlap between sets of words in the query and in the document, capturing lexical similarity through shared terms.

Semantic Retrieval leverages neural embeddings to capture deeper meaning beyond surface-level word matching. For Arabic-specific models, AraGPT2 is a pre-trained BERT-based model tailored for Arabic language understanding, which enables it to grasp syntactic and semantic nuances. Another custom approach, referred to as Arabic Retrieval, utilizes embeddings specifically trained using Arabic legal text to improve relevance in legal QA scenarios.

Multilingual and general-purpose models also contribute to semantic retrieval. E5-large and E5-small are state-of-the-art embedding models developed by Microsoft that generate powerful text representations for cross-lingual and monolingual retrieval. MiniLMv2 is a lightweight transformer that maintains strong performance while being computationally efficient. IBM Granite is a domain-aware foundational model designed to handle retrieval tasks across diverse fields. In addition, a Legal Embedding Model may be used when specifically tuned for legal texts, further enhancing domain relevance.

Hybrid Retrieval Combining sparse and semantic techniques aims to blend the strengths of both keyword-based and semantic approaches. For example, combining BM25 with Arabic Retrieval allows keyword hits to guide and boost semantic similarity, especially effective in morphologically rich languages like Arabic. Sim-

[8] `https://numpy.org/`
[9] `https://pytorch.org/`

ilarly, integrating BM25 with E5 large embeddings balances lexical precision and semantic generalization, offering improved retrieval performance in complex legal queries.

3.4.3 Generation Performance Evaluation

A range of generative models was tested, including Arabic-specific models such as AraGPT2 and AraBERT, as well as multilingual models like DeepSeek, XLM-R Large, and BERT Multilingual QA. These models were evaluated in combination with the top performing retrieval methods identified during the retrieval phase, namely hybrid retrieval (Arabic Retrieval with BM25) and semantic retrieval (Arabic Retrieval and E5-large). Additionally, the models' performance was compared against a baseline scenario without the use of the retrieval phase.

3.4.4 Results and Discussion

In this section, we present a detailed analysis of the experimental results obtained throughout the study, highlighting key findings, and evaluating model performances. The experiments focused on assessing the capabilities of embedding-based retrieval and generative models in processing Arabic legal texts within the Algerian context. The discussion focuses on how these models support the development of effective question-answering systems by enabling accurate retrieval of relevant legal articles and generating contextually appropriate answers.

Retrieval phase

Table 3.5: Evaluation scores at various cutoffs (k = 1, 2, 3, 5) for all models, grouped by model type.

		Sparse		Semantic						Hybrid	
@k	**Metric**	**BM25**	**Jaccard**	**Arabic Retrieval**	**E5-large**	**E5-small**	**legal-model**	**MiniLMv2**	**AraBERT**	**BM25+E5-large**	**BM25+AR**
1	P@1	0.7442	0.5116	0.7326	0.7907	0.7209	0.7093	0.3140	0.2674	0.767	0.7209
	R@1	0.6516	0.4365	0.6516	0.6982	0.6284	0.6303	0.2684	0.2422	0.669	0.6284
	F1@1	0.6778	0.4568	0.6739	0.7243	0.6545	0.6526	0.2818	0.2500	0.697	0.6545
	MRR@1	0.7442	0.5116	0.7326	0.7907	0.7209	0.7093	0.3140	0.2674	0.767	0.7209
	AP@1	0.6516	0.4365	0.6516	0.6982	0.6284	0.6303	0.2684	0.2422	0.669	0.6284
2	P@2	0.4767	0.3837	0.4884	0.5000	0.4826	0.4419	0.2267	0.1628	0.517	0.4826
	R@2	0.7975	0.6308	0.8207	0.8324	0.7975	0.7607	0.3726	0.2699	0.861	0.8091
	F1@2	0.5756	0.4585	0.5911	0.6027	0.5795	0.5411	0.2713	0.1950	0.624	0.5833
	MRR@2	0.8023	0.5988	0.8198	0.8430	0.8023	0.7791	0.3779	0.2907	0.849	0.7965
	AP@2	0.7422	0.5494	0.7481	0.8440	0.7742	0.7471	0.3583	0.2822	0.786	0.7364
3	P@3	0.3643	0.2752	0.3605	0.3643	0.3527	0.3217	0.1783	0.1434	0.368	0.3450
	R@3	0.8939	0.6710	0.8706	0.8823	0.8629	0.8086	0.4438	0.3493	0.894	0.8387
	F1@3	0.4979	0.3738	0.4897	0.4955	0.4820	0.4440	0.2444	0.1957	0.501	0.4686
	MRR@3	0.8295	0.6105	0.8275	0.8547	0.8178	0.7868	0.4012	0.3217	0.857	0.8043
	AP@3	0.7844	0.5664	0.7786	0.8852	0.8251	0.7825	0.3825	0.3216	0.805	0.7544
5	P@5	0.2326	0.1884	0.2302	0.2279	0.2256	0.2093	0.1302	0.1023	0.233	0.2116
	R@5	0.9201	0.7534	0.9142	0.9026	0.9046	0.8629	0.5266	0.3944	0.926	0.8517
	F1@5	0.3566	0.2883	0.3533	0.3494	0.3470	0.3261	0.2009	0.1558	0.357	0.3258
	MRR@5	0.8347	0.6262	0.8351	0.8605	0.8260	0.7944	0.4174	0.3333	0.862	0.8066
	AP@5	0.7952	0.5894	0.7914	0.8971	0.8481	0.8045	0.4124	0.3315	0.816	0.7581

The evaluation results presented in Table 3.5 compare a variety of sparse, semantic, and hybrid retrieval models based on their performance in different metrics and cutoff points (@k = 1, 2, 3, 5). E5-large consistently outperforms the other tested models, particularly at lower cutoffs. At @k = 1 and @k = 3, it achieved the

highest precision and recall, both reaching 0.7907, demonstrating its strong ability to retrieve relevant results at the top of the ranking. The Arabic Retrieval model also performs well at @k = 5, achieving a high recall of 0.9142, while BM25 slightly surpasses it with a recall of 0.9201, outperforming both Jaccard and embedding-based models at this level.

The hybrid model BM25 + Arabic Retrieval showed excellent overall performance, especially at higher cutoffs, suggesting that combining sparse and semantic retrieval strategies significantly enhances performance when broader result coverage is required, such as retrieving the top five results. In contrast, the MiniLMv2 and AraBERT models delivered relatively weak results across all metrics, highlighting their limitations in capturing semantic relevance in the legal context. This underperformance can be attributed to two main factors: the limited capacity of these models (due to their small size and limited pre-trained data) and the nature of the data they were trained on, which may not align well with the legal domain. Overall, these findings emphasize the effectiveness of large-scale semantic models and hybrid strategies for legal question-answering tasks.

Generation phase

Table 3.6: Evaluation metrics for generator models with various retrieval methods.

Retrieval	Generator	Relevance	Faithfulness	BERTScore			ROUGE			BLEU
				P	R	F1	1	2	L	
Base line	Aragpt2	/	/	0.5949	0.4445	0.5088	0.0000	0.0000	0.0000	0.0600
	AraBERT	/	/	0.5668	0.4526	0.5033	0.0000	0.0000	0.0000	0.1600
	DeepSeek	/	/	0.5205	0.4267	0.4690	0.0000	0.0000	0.0000	0.0400
	Multilingual	/	/	0.5963	0.4284	0.4653	0.1176	0.0000	0.1176	0.3600
	XLM-RoBERT	/	/	0.5915	0.5387	0.5639	0.0000	0.0000	0.0000	0.0293
Arabic Retrieval	Aragpt2	0.7274	0.8502	0.6714	0.7253	0.6967	0.1869	0.0659	0.1617	0.0385
	AraBERT	0.7519	0.8346	0.6251	0.7199	0.6687	0.2227	0.0724	0.1921	0.0388
	DeepSeek	0.7020	0.8290	0.6780	0.7205	0.6981	0.1959	0.0749	0.1720	0.0450
	Multilingual	0.2572	0.2573	0.6928	0.5440	0.6080	0.0347	0.0122	0.0343	0.0007
	XLM-RoBERT	0.2358	0.2194	0.6904	0.5572	0.6156	0.0394	0.0173	0.0385	0.0038
E5-large	Aragpt2	0.9071	0.9588	0.6647	0.7134	0.6877	0.1602	0.0479	0.1349	0.0309
	AraBERT	0.8922	0.9456	0.6037	0.7133	0.6535	0.1455	0.0343	0.1195	0.0179
	DeepSeek	0.8999	0.9521	0.6645	0.7087	0.6853	0.1702	0.0580	0.1459	0.0352
	Multilingual	0.8048	0.8057	0.6941	0.5459	0.6101	0.0400	0.0152	0.0396	0.0006
	XLM-RoBERT	0.8179	0.8188	0.6926	0.5557	0.6155	0.0328	0.0103	0.0311	0.0007
Hybrid: Arabic-Retrieval + BM25	Aragpt2	0.7888	0.9527	0.7187	0.7622	0.7398	0.1779	0.0919	0.1708	0.0411
	AraBERT	0.6779	0.9156	0.6828	0.7931	0.7338	0.2712	0.1273	0.2486	0.0769
	DeepSeek	0.7559	0.8277	0.6875	0.7438	0.7138	0.2531	0.1273	0.2315	0.0750
	Multilingual	0.2606	0.2598	0.6959	0.5469	0.6114	0.0385	0.0153	0.0381	0.0009
	XLM-RoBERT	0.2478	0.2301	0.6914	0.5586	0.6170	0.0424	0.0215	0.0422	0.0043

The results presented in Table 3.6 indicate that the baseline models, which do not include any retrieval, show poor performance across all metrics. The generated answers lack context and factual grounding, resulting in low Relevance and Faithfulness scores, as well as extremely low BLEU and ROUGE scores, often close to zero. This highlights the critical importance of incorporating relevant legal content during the generation process.

When using Arabic retrieval methods, performance improves significantly, especially for the AraGPT2 and AraBERT models. These models demonstrate strong Faithfulness and Relevance due to the retrieval of domain-specific legal content. However, metrics such as BLEU and ROUGE remain relatively low. This discrepancy suggests that although the generated answers are accurate and contextually appropriate, they do not lexically match the reference answers, as they may use different words or sentence structures to convey the same meaning.

The E5-large retriever exhibits strong performance across both semantic and surface-level metrics. This is likely due to its multilingual training and high-quality embeddings, which enable it to retrieve diverse and semantically rich content. Despite this, the BLEU and ROUGE scores do not always align with Faithfulness, reinforcing a common challenge in generation evaluation: these metrics rely on exact n-gram overlaps and may unfairly penalize correct answers that are phrased differently from the reference.

The hybrid setup, combining Arabic retrieval with BM25 and paired with AraGPT2, achieved the highest BERTScore F1 with 0.7398, reflecting a strong semantic similarity to the references. It also maintains high Faithfulness at 0.9527. This suggests that combining lexical retrieval techniques like BM25 with semantic retrieval based on dense Arabic embeddings helps capture both precise and diverse content effectively. However, for some other generators, such as Multilingual or XLM-RoBERT, this hybrid retrieval approach may introduce conflicting or redundant information, which can slightly reduce coherence.

Relevance measures how topically aligned the generated answer is with the retrieved legal context. It is calculated by comparing the answer with individual parts of the context. This reflects how well the retriever provided relevant information. Faithfulness measures whether the answer is factually accurate with respect to the entire retrieved context. It ensures that the answer is grounded in the source content and does not include unsupported or hallucinated information. BERTScore measures semantic similarity between the generated and reference answers and is tolerant to paraphrasing. ROUGE and BLEU focus on n-gram overlap, or word-for-word similarity, and thus penalize paraphrased answers even when they are factually correct.

In summary, some model combinations excel in Faithfulness but score low on BLEU and ROUGE because they produce accurate but differently worded answers. Others may score higher in ROUGE and BLEU by matching the reference wording more closely, but are not necessarily more faithful. The most balanced setup, AraGPT2 combined with hybrid retrieval, achieved both high factuality and strong semantic similarity, indicating that it aligns best with both human judgment and automated evaluation metrics. Notably, some of the answers generated by this setup that demonstrated its effectiveness are presented in Table 3.7.

Table 3.7: Examples of questions, contexts, and answers from Tax-RAG

Question	Retrieved context	Generated answer	Reference answer
هل يمكن للخاضعين للنظام الضريبة الجزافية الوحيدة ان يختاروا الحصول لنظام الحقيقي والنظام المبسط ومتى يكمن طلب ذلك؟	المادة 3: يمكن المكلفين بالضريبة الخاضعين لنظام الضريبة الجزافية الوحيدة أن يختاروا، حسب الحالة، الخضوع للضريبة حسب نظام الربح الحقيقي أو النظام المبسط للمهن غير التجارية. ويبلغ الاختيار للإدارة الجبائية قبل أول فبراير...	لا يوجد في قانون الضرائب الجزافية ما يمنع المكلفين بالضريبة الجزافية من التقدم بطلب للحصول على هذا النظام، إلا إذا كانوا قد حصلوا على موافقة مسبقة من الإدارة الضريبية وإذا لم يحصلوا على الموافقة المسبقة فإن عليهم تقديم طلب للحصول على هذه الموافقة إلى الإدارة الضريبية خلال ثلاثة أشهر من تاريخ حصولهم على الموافقة.	يمكن للمكلفين بالضريبة الخاضعين للنظام الضريبة الجزافية الوحيدة ان يختاروا حسب الحالة الخضوع للضريبة حسب النظام الحقيقي او النظام المبسط للمهن غير التجارية عند تبليغ الإدارة الجبائية قبل 10 فيفري من السنة الموالية التي يرغب فيها المكلف تطبيق هذا النظام المادة 30 من قانون الإجراءات الجبائية
ما هو الإجراء المطلوب من المكلفين بالضريبة عند بداية نشاطهم التجاري أو المهني؟	المادة 381: يجب على المكلفين بالضريبة الخاضعين للضريبة على أرباح الشركات أو للضريبة على الدخل الإجمالي أو للضريبة الجزافية الوحيدة أن يقوموا في الثلاثين (03) يوما من بداية نشاطهم، لدى مصلحة الوعاء التابعين لها، باكتتاب تصريح مطابق للنموذج الذي تقدمه الإدارة. بالنسبة للتجار، يُعتد " ببداية النشاط " التاريخ المدون في السجل التجاري المسلم للمكلف بالضريبة...	إذا لم يكن المكلف بالضريبة قد قدم التصريح المذكور سابقا، فإن هذا التصريح لا ينطبق عليه أحكام الفقرة السابقة. إذا لم يقدم التصريح خلال مهلة ثلاثين (03) يوما من تاريخ تقديم التصريح، فإنه لن يكون له الحق في المطالبة بالتعويض عن الضرر الذي لحقه جراء التأخير في تقديم التصريح.	يجب على المكلفين الخاضعين للضريبة وفقاً للمادة 381 من قانون الضرائب المباشرة تقديم تصريح بالوجود لدى مصلحة الوعاء المختصة خلال ثلاثين (03) يوماً من بدء نشاطهم. ويتعين على التجار تحديد تاريخ بدء النشاط وفقاً للسجل التجاري، بينما يعتمد غير التجار على التاريخ المذكور في الوثيقة القانونية المانحة لترخيص النشاط. ويشمل التصريح المطلوب تقديم نسخة قانونية من شهادة الميلاد، مع الإشارة إلى البيانات الأساسية مثل الأسماء والعناوين التجارية والمقرات. كما يلزم تقديم نسخ من عقود الدراسات أو الأشغال في حالة الأجانب، مع ضرورة إرفاق تصريح شامل بجميع وحدات المؤشأة في حالة تعدد الفروع. ويتم تقديم هذه المستندات وفق النموذج الرسمي الذي توفره الإدارة الضريبية.

3.5 Conclusion

This chapter presents the methodology and experiments behind the Tax-RAG system, a question-answering model for Arabic legal texts in the Algerian tax domain, based on Retrieval-Augmented Generation (RAG). A custom benchmark of expert-validated question-answer pairs was developed for evaluation. The retrieval phase tested sparse, semantic, and hybrid methods, with the hybrid (BM25 + Arabic-Retrieval) performing best. AraGPT2 proved most effective in the generation phase. However, automated evaluation metrics like BLEU and ROUGE sometimes misaligned with human judgment due to paraphrasing. The study high-

lights the system's potential for legal and tax-related applications.

General Conclusion and Perspectives

This thesis addresses the challenge of developing a reliable legal question-answering system tailored to the Algerian context, where access to accurate legal information is limited—especially for non-experts—due to linguistic complexity, scattered sources, and lack of digitization. This issue is particularly significant as it promotes legal awareness, public trust, and transparency, aligning with broader goals of digital governance. The work's contribution lies in designing a Retrieval-Augmented Generation (RAG)-based system tailored to the Algerian legal context in Arabic, which leverages advanced semantic search and generative models to produce accurate, legally grounded responses—offering a practical solution to a socially and technologically relevant problem.

This system preprocesses legal documents splitting them into articles, which are then converted into vector embeddings using models such as E5-large and Arabic-Retrieval. These vectors are indexed with FAISS for efficient retrieval. Retrieved articles guide a pre-trained large language model (LLM) in generating accurate answers. Additionally, a benchmark based on the Algerian tax code evaluates both retrieval and generation.

Experimental results highlighted that dense embeddings and prompt design improve answer fidelity, while generative models vary in producing law-compliant responses. Confirming the value of specialized embeddings and domain-sensitive generation in creating practical legal applications.

Despite these contributions, the study faces some limitations. Most notably that the current application is restricted to tax law, excluding other legal fields, limiting its effectiveness as a comprehensive legal aid tool. Linguistically, the system handles Modern Standard Arabic efficiently, it struggles with colloquial queries and the bilingual nature of Algerian legal practice, where French is frequently used. Additionally, the system's overall performance heavily largely on the accuracy of retrieved articles and the performance of pre-trained language models were not specifically tailored to Algerian legal texts, which may affect reliability and domain relevance.

Scientifically, the project demonstrates the effectiveness of the Retrieval-Augmented Generation (RAG) approach in resource-constrained, domain-specific natural language processing (NLP) scenarios, particularly within the legal domain. From a practical perspective, the system provides a valuable tool for improving access to legal information, thereby promoting transparency and improving legal literacy in the region. It could also have important applications in legal education and assisting

lawyers, researchers, and public institutions.

Future prospects for this work point to several promising directions. First, the system could be extended to cover additional branches of Algerian law, such as civil, commercial, or administrative law. Second, fine-tuning the language model on Algerian legal corpora would likely enhance performance and reduce ambiguity in answer generation. Another important avenue involves developing automated mechanisms to keep the legal database up-to-date in response to legislative changes. Finally, implementing multilingual support—particularly in both Arabic and French—would ensure greater accessibility and usability for the diverse linguistic landscape of the Algerian legal community. Additionally, future work could explore multi-modality to enable the system to process not only textual data but also other forms such as scanned legal documents, images, or audio inputs, thereby broadening its applicability.

Bibliography

ADEBAYO, K. J., Di Caro, L., Boella, G., & BARTOLINI, C. (2016). An approach to information retrieval and question answering in the legal domain..

Alcántara Francia, O. A., Nunez-del Prado, M., & Alatrista-Salas, H. (n.d.). Survey of text mining techniques applied to judicial decisions prediction. *Applied Sciences*, *12*(20). Retrieved from `https://www.mdpi.com/2076-3417/12/20/10200` doi: 10.3390/app122010200

Allam, A., & Haggag, M. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences*, *2*, 211-221.

Alotaibi, S. S., Munshi, A. A., Farag, A. T., Rakha, O. E., Sallab, A. A. A., & Alotaibi, M. (2022). Kab: Knowledge augmented bert2bert automated questions answering system for jurisprudential legal opinions. *IJCSNS International Journal of Computer Science and Network Security*, *22*(6), 346–356. Retrieved from `https://doi.org/10.22937/IJCSNS.2022.22.6.44` (PDF created on June 11, 2022, using Acrobat Distiller 22.0 (Windows)) doi: 10.22937/IJCSNS.2022.22.6.44

Antoun, W., Baly, F., & Hajj, H. (2021). AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the sixth arabic natural language processing workshop* (pp. 196–207). Kyiv, Ukraine (Virtual): Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2021.wanlp-1.21`

Askari, A., Yang, Z., Ren, Z., & Verberne, S. (2024). Answer retrieval in legal community question answering. In N. Goharian et al. (Eds.), *Advances in information retrieval* (pp. 477–485). Cham: Springer Nature Switzerland.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., … Jégou, H. (2024). The faiss library.

Duong, H.-T., & Ho, B.-Q. (2014). A vietnamese question answering system in vietnam's legal documents. In K. Saeed & V. Snášel (Eds.), *Computer information systems and industrial management* (pp. 186–197). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). CUAD: an expert-annotated NLP dataset for legal contract review. *CoRR*, *abs/2103.06268*. Retrieved from `https://arxiv.org/abs/2103.06268`

Hoshino, R., Taniguchi, R., Kiyota, N., & Kano, Y. (2019). Question answering system for legal bar examination using predicate argument structure. In K. Kojima, M. Sakamoto, K. Mineshima, & K. Satoh (Eds.), *New frontiers in artificial intelligence* (pp. 207–220). Cham: Springer International Publishing.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

Kalra, R., Wu, Z., Gulley, A., Hilliard, A., Guan, X., Koshiyama, A., & Treleaven, P. C. (2024). HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. In S. Kumar et al. (Eds.), *Proceedings of the 1st workshop on customizable nlp: Progress and challenges in customizing nlp for a domain, application, group, or individual (customnlp4u)* (pp. 237–256). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.customnlp4u-1.18/ doi: 10.18653/v1/2024.customnlp4u-1.18

Kim, M.-Y., Xu, Y., & Goebel, R. (2017). Applying a convolutional neural network to legal question answering. In M. Otake, S. Kurahashi, Y. Ota, K. Satoh, & D. Bekki (Eds.), *New frontiers in artificial intelligence* (pp. 282–294). Cham: Springer International Publishing.

Kim, M.-Y., Xu, Y., Lu, Y., & Goebel, R. (2017). Question answering of bar exams by paraphrasing and legal text analysis. In S. Kurahashi, Y. Ohta, S. Arai, K. Satoh, & D. Bekki (Eds.), *New frontiers in artificial intelligence* (pp. 299–313). Cham: Springer International Publishing.

Kourtin, I., Mbarki, S., & Mouloudi, A. (2021). A legal question answering ontology-based system. In B. Bekavac, K. Kocijan, M. Silberztein, & K. Šojat (Eds.), *Formalising natural languages: Applications to natural language processing and digital humanities* (pp. 218–229). Cham: Springer International Publishing.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., & Nogueira, R. (2021). Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international acm sigir conference on research and development in information retrieval* (p. 2356–2362). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3404835.3463238 doi: 10.1145/3404835.3463238

Louis, A., van Dijck, G., & Spanakis, G. (2024). Interpretable long-form legal question answering with retrieval-augmented large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22266-22275. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/30232 doi: 10.1609/aaai.v38i20.30232

Mansouri, B., & Campos, R. (2023). *Falqu: Finding answers to legal questions.* Retrieved from https://arxiv.org/abs/2304.05611

Martinez-Gil, J., Freudenthaler, B., & Tjoa, A. M. (2019). Multiple choice question answering in the legal domain using reinforced co-occurrence. In S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Database and expert systems applications* (pp. 138–148). Cham: Springer International Publishing.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., … Mian, A. (2024). *A comprehensive overview of large language models.* Retrieved from https://arxiv.org/abs/2307.06435

Nguyen, H.-T., Yamada, H., & Satoh, K. (2024). *Gpts and language barrier: A cross-lingual legal qa examination.* Retrieved from https://arxiv.org/abs/2403.18098

Nguyen, T.-M., Nguyen, X.-H., Mai, N.-D., Hoang, M.-Q., Nguyen, V.-H., Nguyen, H.-V., … Vuong, T.-H.-Y. (2023). *Nowj1@alqac 2023: Enhancing legal task performance with classic statistical models and pre-trained language models.* Retrieved from https://arxiv.org/abs/2309.09070

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1410/ doi: 10.18653/v1/D19-1410

Rosa, G. M., Rodrigues, R. C., Lotufo, R. A., & Nogueira, R. (2021). Yes, BM25 is a strong baseline for legal case retrieval. *CoRR*, *abs/2105.05686*. Retrieved from https://arxiv.org/abs/2105.05686

Shaheen, M., & Ezzeldin, A. M. (2014). Arabic question answering: Systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, *39*(5), 4357–4374. (Review Article - Computer Engineering and Computer Science) doi: 10.1007/s13369-014-1062-2

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., … Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., … Fleisch, B. (2024). Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. In J. A. Recio-Garcia, M. G. Orozco-del Castillo, & D. Bridge (Eds.), *Case-based reasoning research and development* (pp. 445–460). Cham: Springer Nature Switzerland.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.6/ doi: 10.18653/v1/2020.emnlp-demos.6

Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, *13*(5). Retrieved from https://www.mdpi.com/2227-7390/13/5/856 doi: 10.3390/math13050856

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., … Cui, B. (2024). *Retrieval-augmented generation for ai-generated content: A survey.* Retrieved from https://arxiv.org/abs/2402.19473

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024). *Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely.* Retrieved from https://arxiv.org/abs/2409.14924

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J.-R. (2025). *A survey of large language models.* Retrieved from https://arxiv.org/abs/2303.18223

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). Jec-qa: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9701-9708. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6519 doi: 10.1609/aaai.v34i05.6519

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de La Recherche Scientifique

كلية العلوم والتكنولوجيا

Faculté des Sciences et de la Technologie

قسم الرياضيات و الإعلام الآلي

Département des Mathématiques & de l'Informatique

جامعة غرداية

Université de Ghardaia

شهادة الترخيص بالإيداع

أنا الأستاذ : أولاد النوي سليمان

بصفتي رئيس و المسؤول عن تصحيح مذكرة الماستر الموسومة ب:

RAG-based Question-Answering System for Algerian Tax law Context

والمُنجزة من طرف الطالبتين:

1. الطالبة :حماني وصال
2. الطالبة :بن يونس زينب

الشعبة: إعلام آلي التخصص: الأنظمة الذكية لاستخراج المعارف تاريخ المناقشة: 30/06/2025

أُشهد بموجب هذا أن الطالبتين قد قامتا بجميع التصحيحات المطلوبة من طرف لجنة المناقشة، وأن النسخة الإلكترونية مطابقة تمامًا للنسخة الورقية، وفقًا للمعايير المعتمدة.

إمضاء المسؤول عن التصحيح

مصادقة رئيس القسم

مساعد رئيس قسم الرياضيات والإعلام الآلي المكلف بالتدريس والتعليم في الطور بوشقوف اسماء

جامعة غرداية قسم الرياضيات والإعلام الآلي كلية العلوم والتكنولوجيا