INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Automation of Information Extraction and Analysis of Portuguese Judgments

Igor da Cunha Caetano

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD José Eduardo de Mendonça Tomás Barateiro, Assistant Professor,
University of Algarve and Iscte - Instituto Universitário de Lisboa

Supervisor:
PhD João Miguel de Sousa de Assis Dias, Assistant Professor,
University of Algarve

September, 2025

[This page is intentionally left blank.]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Automation of Information Extraction and Analysis of Portuguese Judgments

Igor da Cunha Caetano

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD José Eduardo de Mendonça Tomás Barateiro, Assistant Professor,
University of Algarve and Iscte - Instituto Universitário de Lisboa

Supervisor:
PhD João Miguel de Sousa de Assis Dias, Assistant Professor,
University of Algarve

September, 2025

[This page is intentionally left blank.]

To My family for believing in me even when they had no idea of what I was doing. To My girlfriend for always standing by my side. To My grandmother who sadly did not live to see me finish but whose love guided me every step. To my younger self, who bravely (or foolishly) decided to embark on this journey.

[This page is intentionally left blank.]

Acknowledgment

This dissertation represents the result of both rigorous academic inquiry and the unwavering support of remarkable people. I wish to express my heartfelt gratitude to everyone who has guided, inspired, and supported me throughout this journey.

First, I would like to express my gratitude to my supervisors, Professor Doutor José Eduardo de Mendonça Tomás Barateiro and Professor Doutor João Miguel de Sousa de Assis Dias, for the guidance, patience, perspectives, experiences, and knowledge shared throughout this process. Their support and contributions were crucial to the development of this work and for my academic growth.

Secondly, I would like to give a special thanks to ByTheLaw and their support and contributions, without them, this dissertation would not have come to reality.

I would like to express my thanks to the director of the Master's program, Professora Doutora Elsa Cardoso, for her dedication and advice, which contributed not only to the development of this dissertation and my academic growth but also to an enriching overall experience during this Master's program.

To the faculty members and my colleagues of the Master's program, I am grateful for the enriching learning environment and for the countless insights and contributions during classes.

I want to give my sincerest gratitude to my family, friends and specially my girlfriend, for their unwavering support and countless times that they stood by my side motivating me, helping me, and encouraging me throughout my academic journey and the development of this dissertation.

In summary, the research and writing of this Master's dissertation has been enriched by the support, contributions, honesty, and generosity of many remarkable individuals. I hope that this work can make a meaningful contribution to the academic community and the advancement of knowledge in the legal and technological domains.

[This page is intentionally left blank.]

Resumo

A pesquisa jurídica envolve a análise minuciosa de documentos relevantes para identificar princípios jurídicos, regras e legislação aplicáveis que possam apoiar a tomada de decisões jurídicas, a fim de fortalecer ou enfraquecer uma posição jurídica. Esse processo exigente requer que os especialistas invistam uma quantidade considerável de tempo e trabalho, destacando um dos desafios inerentes à área. O crescimento do Processamento de Linguagem Natural (NLP) impulsionou o interesse de muitos investigadores e desenvolvedores em criar e desenvolver ferramentas para o domínio jurídico, a fim de apoiar especialistas jurídicos nessas tarefas, permitindo-lhes trabalhar com mais eficiência. Com o surgimento dos Large Language Models (LLMs), muitos têm desenvolvido ferramentas específicas para o domínio jurídico devido às suas vastas capacidades em vários domínios e tarefas. No entanto, o uso destas tecnologias traz desafios relacionados não apenas à tendência dos modelos de alucinar, mas também à forma como podemos avaliar esses sistemas sem conjuntos de dados anotados. Este estudo tem como objetivo explorar como, através da construção de uma ferramenta baseada em Retrieval Augmented Generation (RAG) de prova de conceito, os LLMs podem ser usados para auxiliar peritos jurídicos na pesquisa jurídica, minimizando as alucinações, e desenvolver uma metodologia de avaliação que possa ser aplicada quando não temos acesso a dados anotados para validar ou avaliar os nossos sistemas.

Palavras-chave: *Legal*, *Large Language Models*, *Retrieval Augmented Generation*, *Documentos Portugueses*

[This page is intentionally left blank.]

Abstract

Legal search entails thoroughly analyzing possible relevant documents to identify applicable legal principles, rules and legislation that can support legal decision-making in order to strengthen or undermine a legal position. This demanding process requires experts to invest a considerable amount of time and labor, highlighting one of the challenges inherent in the field. The growth of Natural Language Processing (NLP) has propelled the interest of many researchers and developers to create and develop tools for the legal domain in order to support legal experts in such tasks, allowing them to work more efficiently. With the emergence of LLMs, many have been developing tools specific for the legal domain due to their vast capabilities across fields and tasks. However, the use of such technologies brings challenges related not only to the models' tendency to hallucinate but also on how we can evaluate these systems without annotated datasets. This study aims to explore how, through building a proof-of-concept RAG-based tool, LLMs can be used for aiding legal experts in legal search while minimizing hallucinations, and to develop an evaluation methodology that can be applied when we don't have access to annotated data to validate or evaluate our systems.

KEYWORDS: *Legal*, *Large Language Models*, *Retrieval Augmented Generation*, *Portuguese Documents*

[This page is intentionally left blank.]

Contents

[This page is intentionally left blank.]

List of Figures

List of Tables

[This page is intentionally left blank.]

List of Acronyms

AI: Artificial Intelligence

LLM: Large Language Model

LLMs: Large Language Models

IR: Information Retrieval

IE: Information Extraction

NLP: Natural Language Processing

NER: Named Entity Recognition

RAG: Retrieval Augmented Generation

QA: Question-Answering

WoS: Web of Science

IEEE: IEEE Xplore

CoT: Chain-of-Thought

[This page is intentionally left blank.]

CHAPTER 1

Introduction

The task of legal search entails thoroughly analyzing possible relevant documents to identify applicable legal principles, rules, and legislation that can support legal decision-making in order to strengthen or undermine a legal position. This demanding process requires experts to invest a considerable amount of time and labor [1], highlighting one of the challenges inherent in the field. With this, a growing interest in leveraging technologies like Natural Language Processing (NLP) has surfaced in order to assist experts by simplifying or automating some processes and tasks due to the nature of the legal domain. Recent advancements in NLP have further propelled exploration and development of innovative solutions, not only for legal search but also for various other tasks in the legal domain. This research focuses on developing a tool capable of automating Information Extraction (IE) and Information Retrieval (IR) of legal documents, specifically Portuguese Judgments, and uses recent NLP technologies like Large Language Models (LLMs) to analyze the extracted information to support legal search. This chapter introduces this study, presenting the motivation, research questions, objectives, and methodology followed in this research.

1.1. Motivation

Traditionally, implementations of NLP technologies in the legal domain have relied on three main approaches: (i) Rule-based methods, where a given system uses manually written rules to extract information from documents; (ii) Information Retrieval (IR) to efficiently retrieve relevant legal documents from various sources; and (iii) Named Entity Recognition (NER) to identify and classify entities present in a given document [2]. In addition to these approaches, some tools or implementations also focus on summarization or even Information Extraction (IE) instead of solely on Named Entity Recognition (NER) [3]. However, it is important to mention that due to the nature of this domain and of these technologies there are limitations and challenges they bring such as the language ambiguity and complexity from legal texts [2].

In the last few years, Large Language Models (LLMs) have emerged and shown great potential across various tasks and fields [4], which has lead many to try and implement them in the legal domain in order to handle some of the challenges. LLMs in the legal domain have been used in many tasks, including legal search, legal consultation [1], Contract Analysis [2] and others. Nevertheless, these models also bring their own limitations and challenges, such as the tendency to hallucinate, which entails in creating information that is not supported by facts [5]. Across fields but, specially in the legal domain, it is

crucial to minimize hallucinations since, in this field, answers should be supported with the proper sources [6].

Due to the tendency of LLMs to hallucinate, the Retrieval Augmented Generation (RAG) framework surfaced as an approach to try to minimize this problem in order to enhance the generation of knowledgeable, accurate and contextually relevant responses [5] [7] [8]. RAG is a framework which involves the use of a retriever and a generator. The retriever is capable of accessing external databases to retrieve the relevant documents for a given query, and the generator, a Large Language Model (LLM) in this case, is responsible of generating the answer to the given question according to the retrieved documents, a system prompt and the contents of the question itself. This leads to a more flexible and cost-efficient use of this technology [9] since there is not a need to fine-tune or re-train a model when we are able to update the document database or the documents themselves [10]. With this, many solutions try to apply and modify this framework while also testing different approaches to execute the retrieval step or testing different generative models or even different embedding models in order to create more semantically and contextually relevant representations of text.

As mentioned, one of the core features of RAG is the use of a retriever, which retrieves relevant documents from a database for a given query. While certain studies focus on modifying the retrieving step, there is still a need to have a high amount of high quality data in order to allow these models to retrieve accurate and relevant documents for a user's query, leading to a need to have ways to create structured annotated datasets.

1.2. Research Questions and Objectives

As previously mentioned, this study will focus on developing a tool developed for the purpose of supporting Portuguese legal experts in legal search. With this purpose, the objectives of this study are:

(1) Build a proof-of-concept system that has the capacity to extract information from Portuguese legal documents using NLP;
(2) Through the use of LLMs analyze the extracted information and provide a summary of the information;

Given these objectives, two research questions arise:

(1) How can we extract relevant information for legal experts from Portuguese legal documents;
(2) How can we validate and evaluate a question-answering systems for the legal domain without a dataset of questions and answers annotated by experts?

1.3. Research Methodology

For this dissertation, it was opted to follow the Design Science Research Methodology [11] to develop an objective-centered solution based on the objectives mentioned in Section 1.2.

The Design Science methodology abides by the following structure:

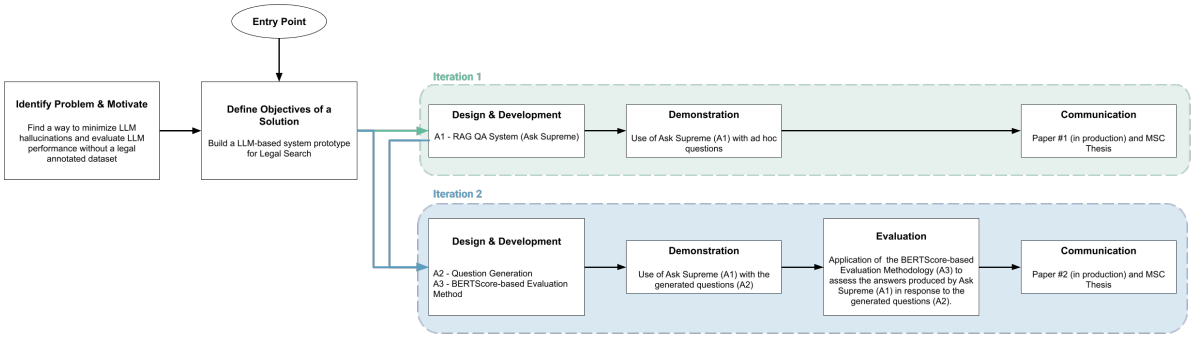

FIGURE 1.1. Design Science Research Methodology Diagram [11]

For the purposes of this study, our entry point is an objective-centered solution, since we are developing a tool with a given objective, in this case, support legal experts through the use of LLMs and the RAG framework.

As detailed in the beginning of this chapter and in Section 1.1, in the legal domain, legal search tends to be a demanding and prolonged process. Such challenge has lead researchers to develop NLP-based systems capable of allowing researchers to execute said process more efficiently.

In order to aid experts in legal search, as mentioned in Section 1.2, our objective is to build a system that leverages NLP technologies such as Retrieval, IE and LLMs capable of extracting, analyzing and summarizing relevant legal information for experts. In our case, we opted to use the Portuguese *Supremo Tribunal de Justiça*'s judgments in order to create a specialized system we could more easily validate.

Through designing and developing our system, we were capable of creating a set of artifacts in order to contribute to a solution for the issue at hand. The contributions produced are (i) a systematic literature review on NLP-based approaches in the legal domain, (ii) the development of an evaluation methodology for question-answering systems without a dataset annotated by experts, (iii) a specialized RAG system capable of receiving an user's request related to our data and answering it based on retrieval and LLM technology.

To demonstrate the effectiveness of the artifacts, we created a validation dataset. Questions were generated by an LLM based on the legal information and the predefined question categories. With our generated questions, through the RAG framework we were capable of taking advantage of the process of retrieval to answer the generated questions while also extracting the retrieved context.

Given this validation dataset, to evaluate our system we took advantage of BERTScore, where we were able to iteratively test and improve our system in order to prove how the system's prototype is capable of aiding experts in the process of legal search.

Finally, the outcomes of this work are intended for scholarly publications, where we expect to publish two articles derived from the literature review and the systems produced alongside this dissertation.

1.4. Document Structure

The remainder dissertation is structured in five chapters, them being:

- **Chapter 2 (Background)**, where we present an overview of the fundamental NLP technologies and techniques in existing approaches applied to the legal domain alongside the key concepts, techniques and processes of the technologies applied during this dissertation.
- **Chapter 3 (Literature Review)** contains the literature analysis of related topics addressed in the dissertation like Large Language Models (LLMs), Information Extraction (IE), Information Retrieval (IR), Retrieval Augmented Generation (RAG) alongside an analysis of the trends present in the literature;
- **Chapter 4 (Design & Development)**, presents the core components and functionalities of our system alongside the decisions made during the development process.
- **Chapter 5 (Results)** shows how the tool and its core functionalities are evaluated by describing the approaches and experiments made and respective results achieved;
- **Chapter 6 (Conclusions)** encompasses the assessment of the tool in terms of complying with the set objectives previously presented and the limitations faced during development alongside with possible future work that can be performed in order to enrich the study;

CHAPTER 2

Background

Before delving into the details of the research made for this dissertation, it is imperative to firstly comprehend the key concepts and processes of NLP based solutions and how they are applied to the legal domain. This section provides an overview of the fundamental NLP technologies and techniques in existing approaches applied to the legal domain, starting with the existing tasks, the datasets used alongside the process of extraction of data, the models and techniques used, and finally the common evaluation methods used.

2.1. Tasks

The first step for the development of any type of tool is the identification of the task or tasks that align with the predefined objectives or requirements of the desired tool. In NLP, depending on the objective, there are many existing tasks; however, for the purposes of this study, only tasks applicable to the specific domain of this research will be presented.

Greco and Tagarelli [12] subdivided several NLP tasks into a set of legal tasks, namely Legal Search, Legal Document Review, and Legal Outcome Prediction.

Legal search or legal research, as mentioned previously, entails thoroughly analyzing possible relevant documents to identify applicable legal principles, rules, and legislation that can support legal decision-making in order to strengthen or undermine a legal position. Normally, this process is lengthy and requires a deep understanding of legislation and how it can be used for a given case, leading this task to be driven by the concepts of relevance in order to obtain legal documents that contain legal information for a legal question of concern.

Legal document review pertains to the analysis and organization of available documents to determine which are sensitive or relevant to the litigation. An example given by Greco and Tagarelli [12] is the following, "*For instance, document review can be intended to negotiate or revise an agreement, ensure that the filings of an attorney's client comply with appropriate regulations, modify a brief for a trial motion, inspect a contract to avoid potential risks, or review client tax documents*". Depending on the case, different criteria might have more relevance, for example for the specific example given previously, the authors mention "*Relevance, responsiveness to a discovery request, privilege, and confidentiality are essential criteria for any document in the review, but also in the analysis of the information to relate key documents to alleged facts or key legal issues in the case*", highlighting the complexity of this process for experts.

Legal outcome prediction, as the name suggests, consists of predicting "*the outcome of a judicial decision based on the relevant facts and laws*" [12]. This prediction can be

applied to a variety of cases, however, there is a higher prevalence of criminal or civil cases for this specific task. Inside this task, Greco and Tagarelli [12] also mention other critical secondary tasks such as *legal precedent prediction*, which involves the prediction of "*passages of precedential court decisions that are relevant to a given legal argument posed in the context of a judicial opinion or a legal brief*", *overruling prediction*, which consists of determining "i*f a statement is an overruling, i.e., a sentence that nullifies a previous case decision as a precedent, by a constitutionally valid statute or a decision by the same or higher ranking court (which establishes a different rule on the point of law ...)*", *case importance and article violation*, which entails "*Predicting the importance of a case can be seen as a regression task, e.g., to measure on a scale from lower scores for key cases, to higher scores for unimportant cases*" and finally, *employment notice prediction*, relating to the prediction of "*the number of months awarded for reasonable notices in employment termination cases*".

Having a basic understanding of which legal tasks are the most commonly seen across studies and implementations, we also need to understand which NLP tasks relate to them and what they entail, with a focus on legal search tasks, since it is the task this dissertation focuses on. Based on the categorization proposed by Greco and Tagarelli [12] and the adaptation of said structure proposed by Siino *et al.* [13], in legal search, we find three main NLP tasks, namely, **Text Entailment**, **Information Retrieval**, and **Question Answering**. In addition to these tasks, based on the objectives of this dissertation, we also included the task of **Information Extraction**, since such a task is relevant to the process of building text-based tools.

Text Entailment aims to determine whether a logical relationship exists between two segments of text, meaning whether a segment of text is inferred from the other [14]. For this task, traditionally, two segments can be classified into one of three categories, for example, assuming we have segments A and B, if A proves B, then its classified as "entailment" or "positive"; if A disproves B then its classified as "contradiction" or negative; and if they have no correlation then its categorized as "neutral" [15].

Information Retrieval is the process of finding or obtaining information within a collection of text or documents given an objective or need, allowing a user to access information from a specific repository through a system. The information can be retrieved at text-level, meaning, retrieving the text inside of a document; at document level, where a set of documents is retrieved; and metadata-level, where the metadata that describes the data of the collection is analyzed and retrieved. There are many ways a user interacts with a similar system, it can be through either a query or an index, which can be formatted according to the specifications of the system, such as numbers or natural language [16].

Question Answering, as the name suggests, involves the process of a user asking the system a question or query expressed in natural language, and consequently, the system will answer the query based on the information it was trained on using natural language. This task can be used alongside information retrieval, allowing the system

to fetch contextual information based on a set of documents the developer wishes to use. Examples of this technology are Frequently Asked Question robots, where given a user's question, they will answer it based on a collection of preset answers, extracting and presenting the answer of the most similar question to the user's query. However, nowadays with the evolution of technologies such as LLMs, we have seen systems using a generation component for answers instead of an extraction component [17].

Information Extraction is the process of extracting and/or identifying structured information such as entities, relationships, events, and sentiments from semi-structured or unstructured text data, and transforming it so it becomes organized, searchable, and machine-readable. In other words, the process of extracting and identifying data from unstructured data and transforming it into structured data, such as, for example, a database table, that allows the systems to more easily use the data and generate insights from it.

Inside Information Extraction there are other sub-tasks derived from it such as **Named-Entity Recognition (NER)** which entails identifying named entities in a text; **Relationship Extraction**, which consists of extracting and categorizing the relationship between entities; **Event Extraction**, which allows a system to identify events such as "meetings" or "appointments", when and where it took place and who was involved; Finally **Sentiment Analysis**, which consists of identifying the sentiment being expressed or communicated by a piece of text [18].

2.2. Datasets

One of, if not the most crucial step to build any data-based tool is the choice of the dataset. This step is crucial because it will influence the model's answer quality and become the basis for all the answers. The choice of the dataset is dependent on the task and objective and in certain cases it is possible to use pre-existing datasets, but on the other hand, specially for the legal domain, there is a need to build our very own dataset since, the applicable legislation is distinct between countries and jurisdictions, which leads many researchers to build datasets specific for their respective country and/or legislation. In this section, some examples of common existing datasets, both general datasets used across domains and specific datasets for the legal domain, are provided.

Firstly, we have SQuAD or Stanford Question Answering Dataset, which consists of a dataset with questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable [19]. This dataset is commonly used for Question Answering-based tasks and is even considered one of the standards for this task.

CUAD or Contract Understanding Atticus Dataset is a corpus of 13,000+ labels in 510 commercial legal contracts that have been manually labeled under the supervision of experienced lawyers to identify 41 types of legal clauses that are considered important in contract review in connection with a corporate transaction, including mergers, acquisitions, and more [19].

LECARD or Chinese Legal Case Retrieval Dataset [20] is a legal case retrieval dataset for Chinese law systems, which contains 107 query cases and 10,700 candidate cases. These queries and results are adopted from criminal cases published by the Supreme People's Court of China with annotations from legal experts.

LLeQA or the Long-form Legal Question Answering dataset is a French native dataset for studying information retrieval and long-form question answering in the legal domain, which contains 1,868 expert-annotated legal questions in the French language, complete with detailed answers rooted in pertinent legal provisions [21].

COLIEE is the Competition on Legal Information Extraction and Entailment, an annual competition run by the International Conference on Artificial Intelligence and Law (ICAIL) [22]. In this competition, there are many tasks that many researchers try to compete in, such as Legal Case Retrieval, Legal Case Entailment, Statute Law Retrieval, and Legal Textual Entailment. For each task, there is a dataset used, such as the Statute Law Retrieval and Legal Textual Entailment datasets, which were built based on the Japanese civil law.

As we can see, there are many datasets, and most of them are used depending on the country where the implementation is being built. In Chapter 3, we will see how during the literature review analysis, researchers chose or built their datasets.

2.3. Embedding models

In order for computational models to understand natural language, we need a way to transform textual data into numerical data. To achieve such a thing, we can use embeddings, which refer to the representation of high-dimensional vectors in lower-dimensional spaces, where each vector represents part of our data. For example, we can represent words such as "king" and "queen" as vectors, and we can make calculations based on those vectors and infer meaning from them based on similarity and/or distance. Given these vectors, machine learning models can process them and understand their patterns and be used for tasks such as classification, clustering, recommendation, text generation, and more [23].

The choice of an embedding model has an impact when implementing it with machine learning models, since, depending on the training data used for these models, the same word can have different representations between models. Consequently, the encoded meaning can also be different, which can lead to better or worse results depending on the objective and available data. Embeddings are computed based on tokens, which are the basic units of text or data that a model processes. Tokens can correspond to full words, sub-word units, characters, sentences, or even larger structures, depending on the model and its tokenization strategy. The choice of which unit to apply is dependent on the task and available data. Some examples of commonly used embedding models are BERT or Bidirectional Encoder Representations from Transformers [24], which considers the context from both sides of a word when making the representation and creates contextualized embeddings, where distinct word embeddings are created for the same word depending on

the context; Word2Vec [25], a prediction-based model which also uses neural networks to learn the associations between words; GloVe or Global Vectors for Word Representation [26] that leverages the co-occurrence matrix of words across the entire corpus.

2.4. Large Language Models

Large Language Models or LLMs are one of the most recent developments in NLP and have made a great impact not only on the field of research and technology but also on culture and in our day-to-day. This section discusses the core functionalities of LLMs and how they work.

2.4.1. Language Models and Transformers

Firstly, we need to understand what language models are. According to Quevedo *et al.* [3], language models or even the task of language modeling consists of predicting upcoming words from a prior word context. We can exemplify this by having a sequence of words, where each word belongs to a set of all possible words, or a vocabulary, and based on the sequence and vocabulary, we can try to predict what is going to be the next word through probability or statistical methods. Advancements in deep learning led to the development of LLMs, which leverage Transformers-based architectures and are trained on a large collection of data, enabling them to generate more coherent natural language across a range of topics. Due to this, these models are more capable of processing vast amounts of information and recognizing complex patterns, allowing them to perform tasks such as summarization, translation, information extraction, and others.

The Transformers architecture has made a great impact both in the NLP field and also the Machine Learning field, leading it to be considered nowadays one of the standards when developing solutions in technological field. The reasons for such architecture to be considered a standard stem from the self-attention mechanism and the ease of access to pre-trained models through the HuggingFace [27] platform. This Neural Network architecture implements an attention mechanism called self-attention, making it possible for the model, through a word in the input, to capture its relationship with other words around it. When a model receives an input, it will first tokenize the input and create embeddings for each token. Given those tokens, the model will process each of them and compute attention scores based on three key concepts: a query, a set of keys, and a value. The query represents a vector for a specific word or token chosen for the respective iteration of the process; the keys represent the vectors for all words or tokens of the input; and the value corresponds to the dot product between the query and a key. Iteratively, a word or token is chosen, and the values are calculated based on the existing keys of the input, giving us the attention score. The attention score is high when two tokens match or are related. This process is done in a Multi-headed attention layer in the Transformers model, so each head applies self-attention and tries to capture different aspects of the relationship [28]. We can see an example in Figure 2.1.

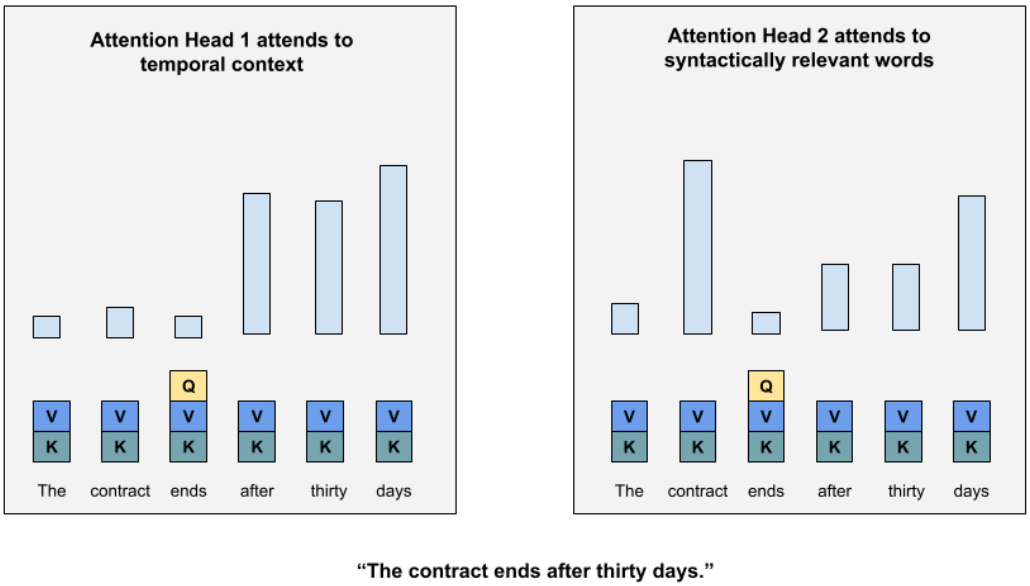

FIGURE 2.1. Transformers' Attention Mechanism Example

Transformers can be used in three different configurations. Encoder-only, where a single attention mechanism is used with bi-directional attention, and is used in tasks where one has immediate access to all words from the text. The bi-directional attention works by instead of only using the previous words to predict a specific word, it also uses the words that come after the predicted word in order to create a representation that considers the whole context. One task where this mechanism is applied in in Mask Language Modeling, which implies masking a random word in a given sentence and training the model to predict that word based on the words around it in the sentence [29] [30]. As we can see here, the focus is to create the best representation possible based on previous and subsequent words in a given sentence.

Decoder-only architecture is used in tasks where one does not have access to all words in the text, because, for instance, they do not exist yet. Hence, the models that employ this architecture are focused on tasks such as text generation [29] [30], where, based on previous words, the model can predict the next word in a sequence. The attention mechanism applied in this architecture is unidirectional, and can only look at preceding words to calculate the representation of the query word.

Finally, Encoder-Decoder architectures combine both attention mechanisms. For instance, it is often used for translation tasks, where the text in the source language is used to create a richer representation using a bi-directional attention mechanism, and the text being generated in the target language is represented using a unidirectional attention mechanism [29] [30].

With this knowledge, we can see how the different Transformers' architectures can be applied to many NLP based tasks. Encoder-only models have a focus on tasks such

as Text Classification, Information Extraction, and Information Retrieval, Decoder-only models have a focus on Text Generation, and Encoder-Decoder models can be used for Text Translation, Question Answering, and Summarization. Some examples of existing models are Google's BERT, one of the most well-known and most used Encoder-only models, OpenAI's GPT (Generative Pre-trained Transformers), a Decoder-only model, and Google's T5, a translation Encoder-Decoder Transformers model [28] [29] [30].

Given this, we now have a grasp of how LLMs work and what their core technologies are. Given this, some LLMs are well known, and many use them in their daily lives. Some examples of LLMs are OpenAI's ChatGPT, which uses different versions of the GPT models mentioned previously and is probably the most recognized model; Google's Gemini/BARD; META's Llama; Anthropic's CLAUDE; Mistral; Deepseek, and others.

2.4.2. RAG and Fine Tuning

Despite the capabilities of these models, due to the statistical methodology applied when generating information, it is rather complex to guarantee and understand the veracity of information, especially when LLMs present it as a fact. When an LLM generates answers with fabricated, false, or inconsistent information, we call these types of answers hallucinations, and these answers are one of the most prevalent limitations of this technology [31]. Considering this occurrence, we can see how hallucinations can impact solutions built for the legal domain, where the factuality of information is significant independently of the task, leading this to not be acceptable in this domain [10]. In order to address such phenomena, different approaches can be applied, such as re-training the model, fine-tuning it, or using the RAG framework.

Re-training consists of repeating the process of model training with or without changing the parameters of this process. The model can be trained for larger periods of time, which can lead to better accuracies, however, the cost of this process is very high [9], and in certain cases, it can not even be feasible due to the amount of processing needed for such a process.

Fine-tuning is the process of adapting a specific model for a specific task and/or domain. Some LLMs such as GPT, Llama and Gemini/Bard are initially trained on a diverse collection of data, which allows them to have a general language understanding. With fine-tuning we can specialize a model for a specific purpose in a chosen domain and/or task by further training a model on a set of documents chosen, leading to the adjustment of its parameters which can lead to higher accuracy and better pattern recognition in the domain of choice. However, this process can be costly [9] both computationally and monetarily, and requires curated datasets for the respective objective, and depending on the amount of data needed for a given objective, this process can either require a lower or greater amount of resources.

Retrieval Augmented Generation or RAG consists of a framework applied to LLMs with the objective of allowing these systems to use up-to-date information without the need to re-train or fine-tune, while trying to minimize hallucinations [9]. The traditional

method for RAG follows a specific set of steps. First, a user gives a query, and through an embedding model, a vector representation of the query is created. Second, through a similarity measure, most commonly the cosine similarity, the top K most relevant documents are retrieved from a database, which contains the vector representations of the data we want to use. Third, the retrieved documents are given to a Generator, a generative model, such as an LLM, that based on those documents and the user query, answers accordingly. Figure 2.2 presents a diagram of how the traditional RAG pipeline works.

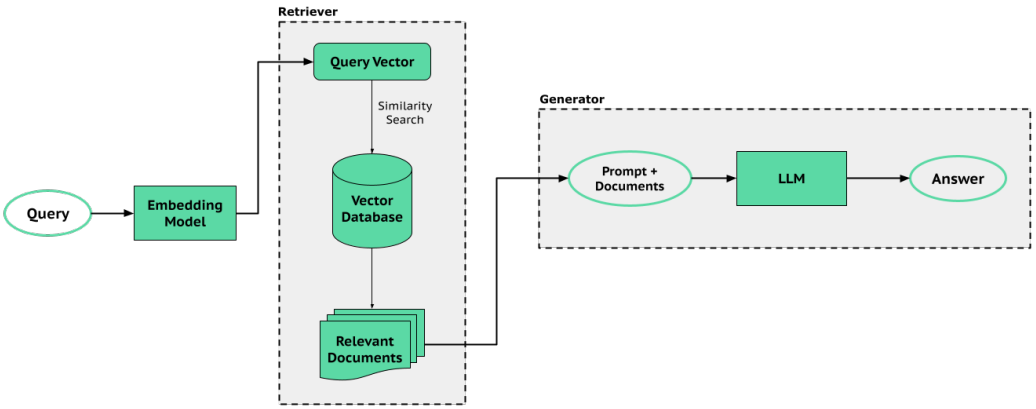

FIGURE 2.2. Traditional RAG Pipeline

This technique allows us to use up-to-date information by updating our database, so for example, in the legal domain, if a certain legislation is updated, we do not need to re-train our LLM if we use RAG, we only update the vector database. With access to tailored data for our task, objective, and domain in the vector database, and by giving the system instructions to answer questions solely based on the retrieved documents, we are capable of minimizing hallucinations. For the legal domain, if our database is populated with legal documents, we are able to increase the reliability of these systems since when it answers a question, it will use valid legal information to formulate that answer.

As we can see, RAG brings a considerable amount of benefits when compared with both re-training and fine-tuning, specially for the legal domain, where information is not static since legislation can and is updated over time, and new cases surge every day. For legal search specifically, it is important to consider all the information available, be it old or new, which translates into a necessity of updating our knowledge base instead of solely relying on new information. With RAG, it is possible to only update the database, instead of changing the parameters of the model, leading us to consider this framework for possible solutions in this field.

Despite this, it is important to keep in mind that hallucination can not be solved, by any solution since this phenomenon is inherently tied to the fundamental mathematical and logical structure of these models. Even though solutions such as RAG try to minimize

hallucinations, this phenomenon will always exist due to how the process of generation is made and the mathematical foundations present in these systems [31].

2.4.3. Prompting

One of the core aspects of LLMs is prompting, which consists in the use of natural language for guiding an LLM's behavior in order to improve its performance in a specific task or output [32]. This differs from traditional models because LLMs are traditionally trained on a large amount of data without a specific objective in mind, allowing them to be used in a wide variety of tasks. With prompting, we are able to leverage the LLM's capabilities for a determined outcome while not altering the weights of the model.

The process of making a prompt, or prompt engineering, can be challenging since it is an iterative process with the objective of achieving the preferred outcome, as natural language can be interpreted in different ways. In order to do this, there is a set of best practices, however, these will not be the focus of this dissertation. With this knowledge, one important factor to consider is the different techniques we can apply in prompting, such as zero-shot prompting, few-shot prompting, and chain-of-thought prompting [32].

Zero-shot and few-shot prompting are related to a type of learning of the same name, zero-shot learning and few-shot learning, which can be applied to machine learning models, where we can choose to include or not include examples in the process of training. For zero-shot we do not give the model any training examples, allowing the model to generalize to new or unseen outputs when it is impossible or impractical to obtain labeled data for each possible category. For few-shot learning we only give a small number of examples, leading the model to generalize based on a desired output, which allows the use small amounts of data when a bigger amount of data is unavailable [33]. For the prompting techniques, this approach can be applied when writing our prompts, meaning when choosing to build a query we can either write examples of output we desire through few-shot prompting which enables few-shot learning or even in-context learning [34], or we can choose not to include any examples and only write the objective with zero-shot prompting, which enables zero-shot learning [35].

Another prompting technique we can apply is the Chain-of-Thought (CoT) technique, which consists of writing a set of intermediate steps for the LLM to follow when trying to solve a given problem [36]. This technique can also be applied alongside either zero-shot prompting or few-shot prompting, depending on the amount of "reasoning" we want the model to use before answering, which allows the model to handle more complex tasks. In addition to this, for the legal domain, Reji *et al.* [37] have incorporated a variation of the IRAC (Issue, Rule, Application, Conclusion) [38], legal analysis technique with CoT in order to leverage both the technology of LLMs and legal framework so that it is possible to achieve better results when using LLMs in the legal domain.

By applying these techniques, we can guide LLMs to a desired output and have more control over it, which can lead to better performances, however, this iterative process

requires effort in order to find the prompt that leads to our desired output according to the desired specifications.

2.5. Evaluation

One crucial factor in building any tool is testing and evaluation. This process involves using both quantitative and/or qualitative measures to assess the tool's effectiveness, accuracy, and overall performance in relation to its intended purpose. In this section, we will present some of the benchmarks used to evaluate the system, describe the metrics commonly applied, and briefly discuss the relevance of qualitative evaluation methods [39].

2.5.1. Benchmarks

Benchmarks or benchmarking for LLMs pertains to the process of using standardized frameworks for assessing the performance of LLMs. Traditionally, a benchmark consists of a sample of data with a set of tasks, metrics, and a scoring mechanism. For the case of LLMs, they are benchmarked based on a set of capabilities, for example, we could benchmark a model on coding, math, legal knowledge, translation, summarization, common sense, reasoning, etc [40]

Benchmarks are crucial in developing and/or enhancing models since they show us the progress of a model while learning, and since they use quantitative measures, they allow us to more directly analyze where the model excels and where it needs to improve. We can apply many benchmarks simultaneously, and most of the time, when new models or new versions of models are released, they are tested in a high number of benchmarks to analyze the capabilities of the LLM [40].

In addition to being a dataset, SQuAD also works as a benchmark. This benchmark version of the dataset contains a specific subset of the data from the original dataset alongside a set of metrics which is used to assess the capabilities of the models in Question-Answering. Other examples are MT-Bench, which tests how well an LLM can engage in dialogue and follow instructions [40], and MTEB, which tests the performance of text embedding models on diverse embedding tasks [41].

2.5.2. Quantitative Metrics

One way to assess the performance of a given system is to use a variety of quantitative evaluation metrics, where each metric provides different insights into the system's performance. There is an extensive list of applicable metrics, however, for the purposes of this dissertation, only the most commonly used in NLP-based solutions will be addressed.

Firstly, we have classification metrics such as Accuracy, Precision, Recall, and F1-score, which focus on evaluating the system's ability to correctly identify information. In order to understand how these metrics are calculated, we need to understand what a confusion matrix is. A confusion matrix, for classification, is built according to Figure 2.3

Focusing on the metrics mentioned and based on the confusion matrix in Figure 2.3, we have:

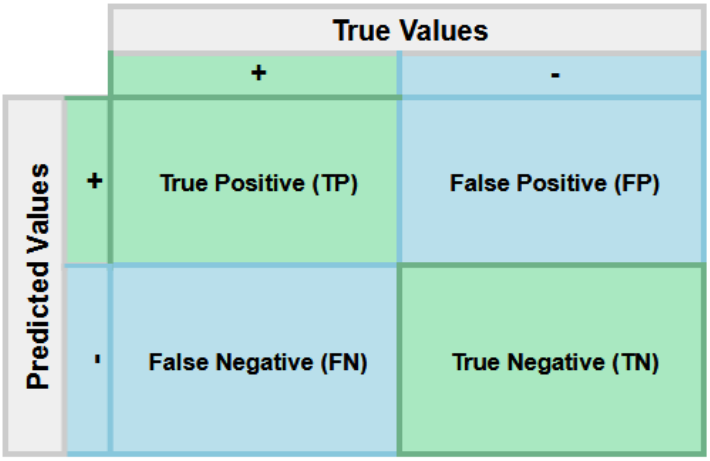

FIGURE 2.3. Confusion Matrix Illustration

(i) Accuracy measures the proportion of correct predictions for all classes among all predictions made, giving us an overall score of how, in general, the model performed in assigning labels [39]. The calculation of this metric is the following:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

(ii) Precision measures the proportion of true positive predictions among all positive predictions made, providing us a score of the positive predictions made [39]. This metric follows this formula:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

(iii) Recall or sensitivity measures the proportion of true positive predictions among all actual positive instances, measuring how often the system identifies the positive instances among all positive samples [39]. It's calculated according to this formula:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

(iv) F1-score consists of the harmonic mean between precision and recall, helping us to identify if a model is becoming too specialized in identifying a specific label [39]. Its formula is the following:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4}$$

Outside of classification measures, we also have regression measures and specific metrics which are used in certain cases, depending on what we want to achieve with the system.

Root Mean Squared Error, or RMSE, which is the square root of the Mean Squared Error (MSE), provides a standard way of measuring the error between the actual value

and the predicted value [42]. RMSE's formula is presented as:

$$RMSE(y, \hat{y}) = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \tag{2.5}$$

For NLP based metrics, one core concept we need to understand is the N-grams technique, which consists of separating sentences into sets of N consecutive words. The following metrics apply N-grams to some extent alongside metrics previously presented.

BLEU, or Bilingual Evaluation Understudy, assesses the quality of a machine-generated text when compared to one or multiple textual references with the use of N-grams and Precision, allowing us to quantify the resemblance between texts [43]. We can see an example of how BLEU can be calculated in Figure 2.4.

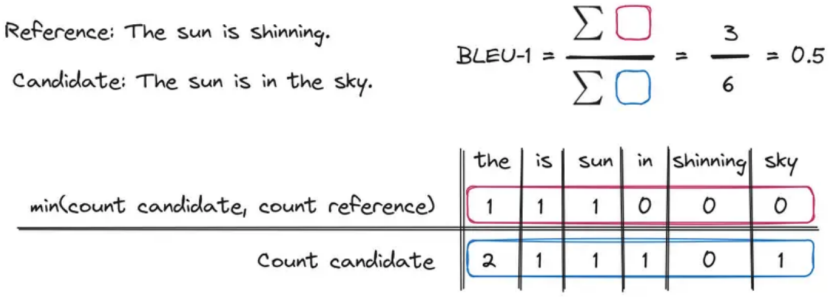

FIGURE 2.4. BLEU Example [43]

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is similar to BLEU, where it is used to assess the effectiveness of machine-generated summaries through N-grams, however, it calculates Recall instead of Precision, leading to this measure focusing on determining how much information from the reference is contained within the generated text [43]. There are multiple variations of ROUGE such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S that employ distinct which employ different techniques in their calculations. We can see an example of ROUGE-N in Figure 2.5.

$$ROUGE - N = \frac{\text{Number of matching n-grams}}{\text{Total n-grams in the reference}}$$

FIGURE 2.5. ROUGE-N Formula [44]

METEOR, or Metric for Evaluation of Translation with Explicit Ordering, differs from the above mentioned metrics by incorporating the F1-score, or the harmonic mean between Precision and Recall, and takes into consideration synonyms, stem forms of words and word order when making its calculations. We can see its formula in Figure

Perplexity is one of the most common metrics used for LLMs. This metric measures how confidently the model is able to predict the sequence of words through the probability distribution for each word generated by the model [43]. Perplexity is calculated in the following way:

$$\text{METEOR} = \underbrace{\text{FMean}}_{\text{Harmonic Mean of Unigram Precision/Recall}} * \underbrace{(1 - \text{Penalty})}_{\text{Word Order Penalty}}$$

FIGURE 2.6. METEOR Formula [45]

$$PPL_{model} = \exp\left(-\sum_{i=1}^{t} \log\left(P_\theta(x_i \mid x_j \neq x_i)\right)\right) \tag{2.6}$$

where $\log(P_\theta(x_i \mid x_j \neq x_i)$ is the log predicted probability of the i^{th} token conditioned on other tokens (j != i) in the sentence according to the model.

BERTScore [46] is a metric used for Transformers-based models, such as BERT, that computes cosine similarity between candidate and reference sentences. The reference sentences represent our standard text, which the candidate, our output text, will compare against by aligning the tokens between sentences and computing the cosine similarity between them, producing a score. The scores are then aggregated to produce Precision, Recall and F1 measures. Here is an example of BERTScore:

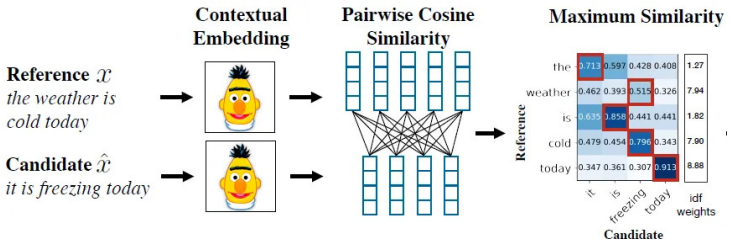

FIGURE 2.7. BERTScore Example [47]

One important consideration to have is that these metrics do not consider the semantic meaning of words, leading these measures to have difficulties analyzing any detail related to the meaning of the words. In certain situations, for example, for problems related to fields where the quality of information and its coherence are imperative, such as in the legal field, other forms of evaluation could be used in addition to metrics in order to have a better system.

2.5.3. Qualitative

In some cases, the use of qualitative techniques either alongside or over quantitative measures is preferred due to the preference for a better understanding of the results and outputs achieved. For instance, with NLP, systems that take advantage of textual data, specially LLMs, are more prone to bias and misinformation in addition to hallucinations [48], leading to the need to evaluate similar systems qualitatively in order to prevent such occurrences, which, for the legal domain is an important consideration [1].

For qualitative techniques, there are not stipulated measures that most follow in order to assess the quality of a system. In such cases, most determine the approach or framework

[1]In future sections, the problems, consequences, and ethical considerations LLMs bring will be discussed more throughly.

according to their respective resources, for example, some might build questionnaires for experts to fill according to their opinion of the system [49], some might ask of experts to evaluate a system according to a set of standards or metrics [50] and many other ways. The conclusion we can take from the use of qualitative techniques for evaluation is that researchers and/or developers should apply such techniques depending on the available resources of the project and its objective and purpose.

CHAPTER 3

Literature Review

3.1. Search Strategy and Inclusion Criteria

This research was conducted until the 1st of January of 2025, following the PRISMA methodology. Three repositories were used, Scopus, Web of Science (WoS), and IEEE Xplore (IEEE), from which 231 articles were obtained. The inclusion criteria that resulted in the articles obtained were: only choose documents from 2021-2025, writing had to be in English, and only conference papers and articles were included.

3.2. Study Selection

To filter the articles and papers obtained, the title, abstract, keywords and introduction were read, from which some exclusion criteria were defined in order to only retrieve the relevant documents related to the objectives and research questions previously stipulated.

3.3. Literature Extraction and Analysis

With the objective of conducting a systematic literature review, the PRISMA guidelines were employed. This set of guidelines aids researchers in conducting systematic literature reviews by providing a clear set of instructions on how to proceed with a literature review in order to gather the relevant information and clearly present it. This allows researchers to more reliably analyze the current state of the concepts, technologies, and/or solutions for their respective research problem. Through this approach, it was possible to more reliably research the relevant information and the most current to the time of this study is being conducting, leading to the enhancement of the quality and reliability of this review and even ensuring the reproducibility of the research process which provides a better overview of the current state of LLMs and other technologies in the legal domain. To better understand how this methodology was applied in the following subsections, the scope of the research will be presented through the research questions, the inclusion and exclusion criteria, and the results of the assessment of the retrieved documents from certain databases.

To extract the necessary documents, the same query was used across every repository, which goes as follows:

> "(TITLE-ABS-KEY("RAG" OR "Retrieval Augmented Generation" OR "LLM" OR "LLMs" OR "Large Language Model" OR "Large Language Models") AND TITLE-ABS-KEY("legal" OR "law" OR "case

law" OR "judicial decisions" OR "legal documents") AND TITLE-ABS-KEY("retrieval" OR "extraction" OR "recognition" OR "summarization")) AND PUBYEAR > 2021 AND PUBYEAR < 2025 AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar"))".

Given the results from the search, 231 articles and papers were found. From this, it was necessary to manually filter documents in order to only keep the relevant literature for this study. Firstly, duplicates were removed, followed by reading the title, abstract, keywords, and in some cases reading the introduction to assess if the article or paper fits by using exclusion criteria to more consistently filter the documents.

Concept	Population	Context	Limitations
Large Language Models	Legal	Retrieval	From 2021-2025
Retrieval Augmented Generation	Law	Extraction	Only articles and conference papers
	Case Law	Recognition	
	Judicial Decisions	Summarization	
	Legal Documents		
70595 documents			
2049 documents			
423 documents			
231 documents			

FIGURE 3.1. Definition of Keywords used to Search

The exclusion criteria used are defined as:

(1) Does not belong to the legal domain;
(2) Does not use textual data;
(3) Objective not related to Legal Search;
(4) Not being publicly accessible;
(5) Being a conference cover;

Figure 3.2 illustrates the screening process applied to the 231 documents retrieved through the PRISMA methodology. First, duplicate records across the three databases were identified, totaling 37 documents, which were subsequently removed. Second, the titles and abstracts of the remaining documents were analyzed to screen for studies that did not meet the inclusion criteria, resulting in the exclusion of 9 studies. Third, the full texts of the remaining 185 documents were thoroughly examined to assess their eligibility, applying the predefined exclusion criteria to determine which studies should be excluded from the review. Specifically, 47 studies were excluded for not belonging to the legal domain, one study did not utilize textual data, 79 studies had objectives unrelated to legal search, 14 studies were not accessible, and one record represented a conference cover. In total, 142 studies were excluded at this stage, leaving 43 documents for inclusion in the review of this dissertation, which will be presented and analyzed in the following sections.

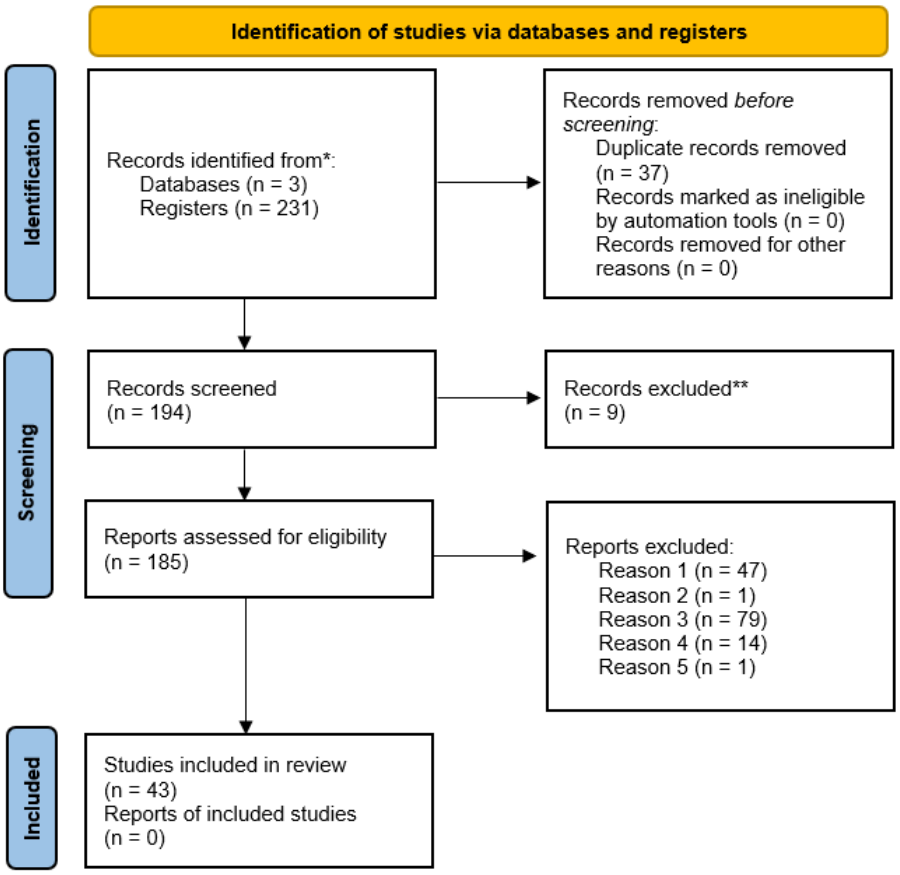

FIGURE 3.2. PRISMA Flow Diagram

Given the new number of articles and papers, it was opted to analyze the results in order to visualize patterns present in the research. Since some articles did not contain keywords, the author assigned each document a set of keywords associated with topics and technologies implemented. The keywords used were LLM, RAG, QA (Question-Answering), Extraction, Retrieval, Summarization, Ethics, Traditional, Recognition, ICL (In-Context Learning), and KG (Knowledge Graphs). These keywords were chosen since they represent the great majority of techniques used across the different studies. "LLM" and "RAG" are separated in order to identify studies that solely implement solutions based on LLMs without RAG and solutions based on the RAG framework respectively; "Retrieval" represents papers that apply or test different retrieval techniques; "Extraction" represents papers that use IE techniques and "Recognition" represents the documents that primarily implement NER and similar techniques focused on extracting one key piece of information; "Ethics" represents articles that describe in certain detail ethical concerns, challenges and limitations of their implementations; "Traditional" represents the documents that directly mention and use traditional NLP techniques or solutions and compare them to their respective solution. "QA," as mentioned, focuses on Question-Answering solutions.

This allows the creation of visualizations and more easily have a set of keywords that represent each article, which we can see in Figures 3.3 and 3.4.

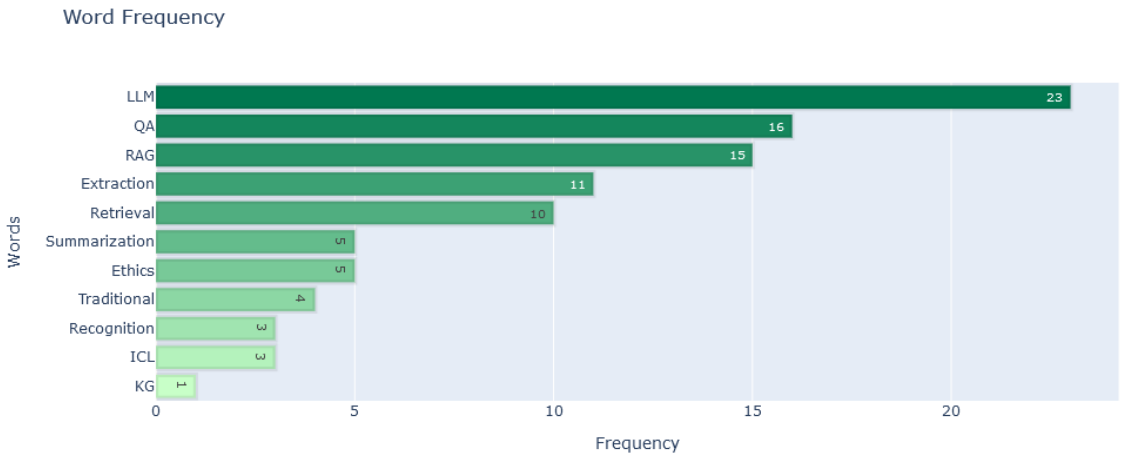

FIGURE 3.3. Word frequency of Keywords

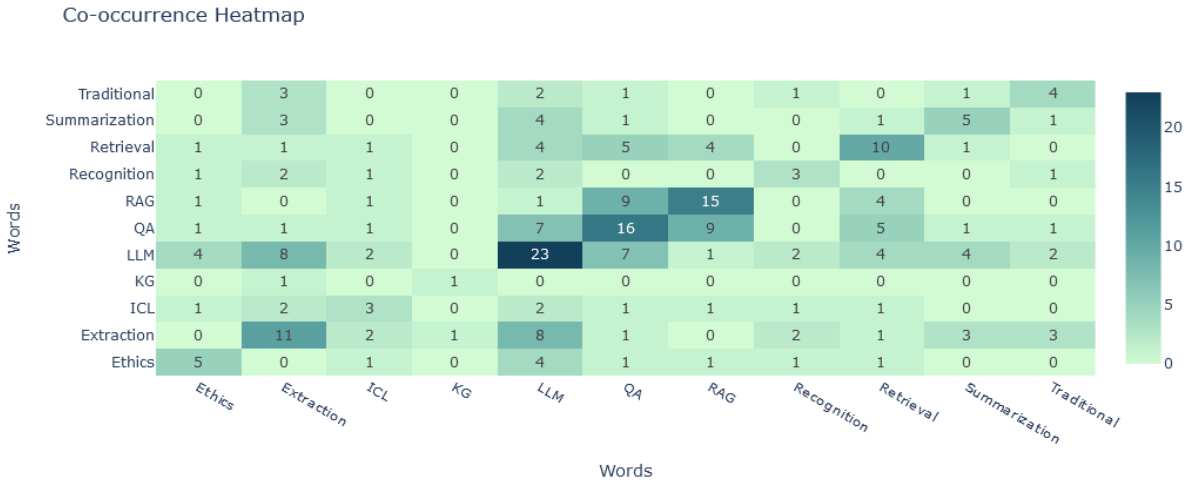

FIGURE 3.4. Frequency Heatmap of Keywords

As we can see on Figures 3.3 and 3.4, there is a high predominance of studies focused on LLMs, and there is also a great amount of studies that implement either RAG or Question-Answering (QA). Alongside these, there seems to exist a moderate amount of articles that implement both QA with either RAG or LLMs. Retrieval techniques have also been shown to be used across implementations due to their moderate amount across the previously mentioned keywords. Contrary to this, there also seems to be a reduced amount of literature obtained that applies either Knowledge Graphs, In-Context Learning, and Recognition techniques. The Ethics label shows to be more associated with LLM applications, which indicates that the challenges and limitations are focused on this type of technology.

3.4. Literature Review Discussion

As introduced in Chapters 1 and 2, the legal domain has increasingly adopted NLP technologies to address many tasks across the domain, especially in tasks such as legal search, legal document review, and legal outcome prediction. With the emergence of LLMs, recent studies have explored new possibilities and challenges in applying these models to legal contexts, often combining them with existing NLP methodologies or integrating them into more complex frameworks such as RAG.

This section presents the findings of the literature review, focusing on how recent works have applied LLMs and related NLP technologies to legal tasks. The structure of this review mirrors the one established in the Chapter 2, and is organized as follows:

- **Datasets**, where the datasets used across the studies and the different strategies and approaches used in order to build the tools laid out by the authors will be presented;
- **Large Language Models**, where it will be explained the capabilities of LLMs, which models were used across studies, which techniques were applied, and the objectives and presentation of their respective approaches;
- **Evaluation**, where it will be discussed how many authors assessed the quality of their tools;
- **Limitations** will show the limitations that authors faced with LLMs, how they tried to circumvent them, and what their impact is when developing solutions based on LLMs

By following this structure, the review aims to provide a comprehensive synthesis of how state-of-the-art techniques are currently being explored and evaluated in the legal NLP domain and what their limitations are.

Table B.1 in Appendix B, presents a comprehensive list of all the articles where their main objective was developing a tool, system or solution. This table does not include other studies that did not develop a system since, for the purpose of analyzing the trends across datasets, models, techniques, and evaluation methods, such documents would not contain that respective information. However, these other studies, according to our review, present reviews, discussions, or testing of the best practices or between existing techniques, which will be discussed through this chapter nonetheless.

3.4.1. Datasets

For the purpose of this study, for each of the documents retrieved, they were categorized into a set of labels as follows: "Public Datasets" comprised of datasets publicly available, or used in conferences; "Extracted Documents from Public Repositories" consists of an approach based on how the authors extracted legal documentation from publicly available repositories and created a dataset based on the specific documents extracted; and "Annotated / Augmented Data either by Humans or LLMs" represents the use of either a public dataset or extracted documentation with addition of either gold standard

annotations, or the use of LLM-generated information such as questions and/or answers to enhance the dataset quality, or the addition of references or other types of information made by humans. We can see the results of this categorization in Table 3.1.

TABLE 3.1. Datasets from Review

Origin of Dataset	# of Studies	Citations
Extracted Documents from Public Repositories/Websites	18	[51], [52], [49], [53], [10], [54], [50], [55], [56], [57], [58], [59], [60], [61], [5], [7], [62], [8]
Public Datasets	11	[63], [64], [65], [66], [50], [67], [68], [37], [2], [69], [70]
Annotated/Augmented Data either by Humans or LLMs	8	[9], [71], [72], [62], [73], [74], [59], [55]

For categories "Public Datasets" and "Annotated / Augmented Data either by Humans or LLMs", we can see a relatively low amount of studies which used such approaches, where we see 11 studies and 8 studies respectively. As for "Extracted Documents from Public Repositories" it is apparent how this approach is the approach of choice, with 18 studies adopting it.

Such a phenomenon was expected due to the nature of the legal domain, as previously mentioned, legal information differs between countries and jurisdictions, which leads researchers to adopt an approach where they extract and utilize only the information relevant to their specific study, considering both the country in which the solution was developed and the study's objectives.

3.4.2. Large Language Models

LLMs, as mentioned, have surfaced recently and have greatly impacted many aspects in our society. They possess a considerable amount of possible uses, specially in the legal domain where they can aid experts in many tasks. In some articles, authors such as Anh *et al.* [2], Quevedo *et al.* [3] and Greco and Tagarelli [12] present some of the NLP specific tasks and how these technologies are used in legal tasks such as legal search, legal document analysis and legal outcome prediction. Anh *et al.* [2] mentions the evolution over the years of NLP methodologies in legal tasks and what the common strategies are when developing NLP based solutions for the legal domain. Such approaches were the use of Rule-based methods, IR and NER to build their solutions. Quevedo *et al.* [3] expands on the aforementioned tasks, mentioning tasks such as summarization but also IE as a whole instead of solely NER. Greco and Tagarelli [12] delve into much detail on how Transformers-based technologies are used in certain legal tasks, such as the ones previously mentioned, and, in addition, share many NLP strategies used to create implementations inside these tasks. Moreover, the authors also mention how GPT-based methods are being used, giving examples of implementations and how they impacted the legal domain, noting some studies which show the potential of LLMs and how they are able to be comparable with other Transformers models. This study also remarks on the limitations that LLMs bring and how the abuse of use of this technology has shifted

the way people work, leading to the development of countermeasures in order to debunk many of the risks these technologies bring, such as misinformation and data security. Also, the studies mention how these technologies are contributing to a rise in interest for researchers, leading to more studies testing the capabilities and limitations of these systems across domains.

In order to better understand how these technologies are being used, firstly, the capabilities of LLMs in the legal domain will be presented. Secondly, we will present the techniques employed across the studies relative to the models used and the training and prompting techniques applied. Lastly, we will also delve into the approaches and techniques used for both Question-Answering, RAG and Information Extraction systems, respectively.

3.4.2.1. *Capabilities*

Some applications where LLMs can be used are presented by Anh *et al.* [2], where they mention the possibility of these models being used in tasks like Legal Document Summarization, Contract Analysis and Generation, Legal Question Answering, Legal Text Classification, Legal Information Extraction and Legal Reasoning. Lai *et al.* [1] also present some of the characteristics these models have in the domain and how they are applicable, for example, in Legal Consultation, where "*...LLMs can interact with users. Therefore, users can ask legal questions to the model and receive answers and suggestions based on the training data*". Quevedo *et al.* [3] present legal NLP tasks like Summarization, Information Extraction, Question Answering and Information Retrieval, that were analyzed between 2015 and 2022 and presents the traditional methods used and some examples LLM-based, showing us this way the great potential these models have in the legal domain. Some studies, such as the one made byM, P, and M [58], discussed the possible use for LLMs specifically in summarization, showing the potential of these models in legal text summarization be it abstractive or extractive and how they are at least comparable to other state-of-the-art deep learning models.

In some situations, such as in the study made by Adhikary *et al.* [74], a way to automatically annotate legal documents through sequence labeling with LLMs was developed as a way to augment the number of training samples. This solution shows how, even if LLMs are not the core technology being applied, they still show potential in augmenting data. Another examples of said approach can be seen in the studies made by Moreira *et al.* [72], Cho *et al.* [73], Ryu *et al.* [55], Hoang *et al.* [67], and Espírito Santo *et al.* [59], where they utilize LLMs to generate samples for their respective datasets, showcasing the possible usages of this technologies outside of the NLP tasks mentioned previously.

There are also in addition to the aforementioned studies, examples of researchers experimenting with this technology for other purposes. S *et al.* [75] use Llama2-7b [76], Flan [77] and Claude2 models for summarizing legal text for legal analysis. The best results the authors achieved were with Llama2-7b in the metrics used, but there is still space for improvement, maybe with the addition of more prior information to the models. Yao *et*

al. [50] present an Artificial Intelligence (AI) Agent, where LLMs are used in various steps during development, like, pre-processing, synthetic data generation for data quality augmentation and CoT output generation. They also applied other approaches for generating answers, such as systems like IR on a Knowledge base, a Recommendation System with re-ranking of relevant documents and a Reinforcement Learning system based on human feedback in order to guarantee higher quality on LLM answers. In terms of results, the authors state that by using qualitative analysis through experts' questions and answers, and metrics, the model was comparable to other tested models. Chauhan *et al.* [61] created an AI Agent based on LLMs that can answer questions and recognize the intention and entities, and uses retrieval to create more accurate answers. The results from this study was a 65% accuracy value, showing potential but also space for improvement. Quevedo *et al.* [3] use Language Models and Large Transformers to show that these solutions still have space for improvement and even show that some models exclusively trained on specific domain text do not necessarily outperform general-purpose models across legal tasks, and suggest the use of these models with other state-of-the-art techniques.

With this great variety of tasks these models are capable of doing, many studies apply LLMs in different ways for different purposes, however, due to the objectives of this study, the tasks that will be presented will focus on Information Extraction, Information Retrieval and Question Answering since these tasks are also the most associated with Legal Search.

3.4.2.2. *Models and Techniques*

In Section 2.4, some examples of Large Language Models were given. Although there is a great number of existing models, in the process of research, it was only possible to see some models across the literature, which we can see in Table 3.2.

TABLE 3.2. Models from Review

Models	# of Studies	Citations
GPT Variants	20	[64], [65], [71], [72], [52], [49], [53], [10], [50], [73], [55], [67], [68], [74], [57], [58], [59], [60], [8], [5]
LLaMa	13	[9], [64], [63], [51], [54], [70], [74], [56], [58], [59], [60], [69], [75]
Gemini / Bard / Gemma	2	[52], [60],
Mistral	7	[9], [66], [56], [37], [60], [69], [7],
Others	9	[64], [51], [53], [62], [50], [67], [59], [2], [75]

For this dissertation, it was opted to divide the models between three of the most well-known models, such as the variations of GPT, Llama [78] and Gemini/BARD [79]. In addition to these, the Mistral [80] category was also added since it is one of the more popular open source models. Finally, the "Others" category represents other models such as Claude, Cohere, Qwen [81] and others.

We can see GPT, Llama and Mistral represent the most used models with 20, 13 and seven implementations respectively. It is expected to see the prevalence of GPT, since it is the most popular model and one with the best results in terms of benchmarks. Llama

also has great results and potential based on the great number of studies that opted to implement this model.

In terms of the techniques used, we can see in Tables 3.3 and 3.4 the techniques used across studies related to training and prompting processes respectively.

Table 3.3. Training Techniques from Literature

Techniques	# of Studies	Citations
LLM Fine-tuning	7	[9], [64], [50], [74], [59], [60], [2]
Retrieval Augmented Generation	13	[9], [63], [64], [51], [49], [53], [66], [10], [54], [55], [37], [5], [7]
LLM Training	2	[61], [69]

While 13 articles use the RAG framework, the application of fine-tuning for LLMs is still used, specially with open-weights models like Mistral and Llama. However, for most RAG implementations, fine-tuning was not used for the final system even if it was considered and tested, as we can see the seven articles which applied such technique. Although the process of fine-tuning an LLM was not overly applied, some studies fine-tuned the embedding models or used already fine-tuned models for legal data and/or respective language in which the solution is being applied, leading to some studies considering this process one of the most defining ones in order to produce better results. We can also see two cases where LLMs were trained from scratch, which is an expensive process leading to the few examples of such technique.

Another aspect we can analyze that is not represented in the tables is the choice of the retrieval similarity measure. For the RAG implementations, as mentioned in previous sections, the cosine similarity is the similarity measure most used across studies. Since most of the approaches followed a naive or traditional application of said framework, it was expected the choice of using cosine similarity. Some studies use other techniques for retrieval, such as Elasticsearch [53] [73] or even divide the retrieval process between types of data, such as only legal cases or only legal legislation [63].

Table 3.4. Prompting Techniques

Prompting Technique	# of Studies	Citations
Zero-shot	4	[67], [56], [58], [37]
Few-shot (One, Two, Three, etc)	5	[74], [56], [58], [37], [73]
Chain of Thought	2	[67], [37]
Others	3	[67], [56], [37]

In terms of the prompting techniques presented in Table 3.4, most studies applied either Zero-shot or Few-shot, and some applied both. In addition to this, two studies experimented with the use of CoT and three others applied other types of prompting techniques. Reji *et al.* [37] use IRAL, a variation of the IRAC [38] method for analyzing legal issues, alongside other prompting techniques.

3.4.2.3. *Question-Answering*

One of the keywords most present in the literature was QA (Question-Answering), which involves having a system receive questions or queries from a given user and answer it accordingly. Normally, these types of systems are based on retrieval, which was corroborated during the research for this dissertation, where the solutions implemented the RAG framework, which is based on Retrieval. We can see in Table 3.5 a summary of the techniques used in the articles retrieved.

TABLE 3.5. Question Answering Approaches

Technique	# of Studies	Citations
Traditional	8	[5], [64], [51], [60], [61], [69], [10], [53]
LLMs for Data Augmentation	7	[9],[55], [67], [59], [49], [72], [63]

As we can see in Table 3.5, for the implementations focused on QA, eight studies approached this task traditionally, where given a dataset with questions and answers, an Information Retrieval System is implemented that based on the question made by a user, the system will answer based on the answer of the most similar question from the dataset. However, we also see another seven examples of approaches where LLMs are used to augment the dataset or even aid in creating a dataset for the specific purpose of the study, in addition to the traditional application of QA techniques.

In terms of RAG framework, we can see that a resonable number of authors apply this technique in their system in different degrees ([5], [64], [51], [10], [53], [9], [55], [49] and [63]), which will be discussed in the RAG section of this study. While the other six studies apply other retrieval techniques.

Ryu *et al.* [55], Hoang *et al.* [67] Espírito Santo *et al.* [59] use LLMs to generate questions based on legal data. This comes from the lack of datasets which contain the necessary data, in this case, both questions and/or answers, related to a data entry.

Visciarelli *et al.* [9] go further compared to the previous studies, where, with the aid of experts, a complete QA dataset is created with metadata related to the legislative information related to a specific entry in the dataset. Likewise, Yao *et al.* [50] used LLMs to not only pre-process the dataset with grammatical corrections, but also generate synthetic data in order to have a higher amount of data with higher diversity and subsequently higher quality. Hu, Luo, and Feng [63] also generate a dataset with LLMs based on the data the authors chose to use, where, after a question is made, the answer will be based not only on the question but also on responses for each retrieved article.

Mamalis *et al.* [49] augment their data with LLMs by using these models to generate sub-questions for each question of the dataset, with the purpose of when an answer needs to be generated, if a given question is rather complex, the model can answer each sub-question, leading to a more in-depth and coherent answer.

Moreira *et al.* [72], like previous solutions, also generate questions, but also generates summaries for each answer to enhance the retrieval step in their system. However, they

go a step further and use clustering techniques to create groups of Common Questions, Groups of Similar Answers and Final Questions where depending on the user's query, the LLM will analyze the cluster and find the most relevant answer for the user.

Bakir, Yildiz, and Aktas [69] focus on traditional techniques for QA also presents metrics to evaluate the quality of the answers generated by LLMs and discuss how traditional metrics and similarity measures such as ROUGE, BLEU, METEOR, Cosine Similarity, F1-Score, Precision and Jaccard are not enough to evaluate QA systems.

In terms of results, although the approaches from these studies show some promise with the application of LLMs and/or RAG, there are still challenges and limitations that directly impact the capabilities and possible formal use of these models in the field which will be explored in the Limitations section.

3.4.2.4. *Retrieval Augmented Generation*

For the RAG implementations, we can see the use of many different models for different tasks. This framework is applied due to allowing these systems to use up-to-date information without the need to re-train or fine-tune, which can be costly, and even tries to minimize the consequences of hallucinations [9]. While there is a traditional way to apply RAG, which was discussed in Chapter 2, some researchers also try to apply different techniques in this framework, be it through changing how the retrieval is done or how similarity is calculated. We can see in Table 3.6 how which studies apply the traditional framework and which apply a non-traditional approach.

Table 3.6. Retrieval Augmented Generation Approaches

Technique	# of Studies	Citations
Traditional	8	[7], [54], [37], [5], [9], [49],[10], [55]
Non-Traditional Retrieval Methods	10	[66], [63], [51], [72], [82], [83], [67], [53], [8], [65]

As we can see in Table 3.6 we have eight examples of studies that follow the traditional approach and 10 that alter how RAG is done.

Some articles follow the traditional method for RAG and address the potential of this framework and LLMs and in some cases presenting their respective limitations [7], [54], [37], [5], [9], [49],[10], [55].

Jacob *et al.* [7] use Mixtral-8x7B-Instruct-v0.1 for RAG which showed moderate results with space for improvement. Alongside this study, Visciarelli *et al.* [9] and Cherubini *et al.* [10] also show similar conclusions where Visciarelli *et al.* [9] tested Mistral-7B-v0.1, Mixtral-8x7B-Instruct-v0.1 and LLaMAntino-2-7b-hf-ITA for legal QA and, with the help of experts the models were evaluated, and the authors reached the conclusion that the technology shows promise. Cherubini *et al.* [10] describe the benefits of RAG over fine-tuning and built a prototype based on RAG and the evaluation was made through a questionnaire for experts. The authors mention how promising these implementations are in answering simple questions, but despite this, they also mention the need for experts to truly evaluate the answers. Kurniawan and Hiererra [5] explain the benefits of RAG

and implement a prototype that is tested using a Likert scale question for users and interviews with experts, in which they showed the potential of these models, although there are ethical concerns, most answers were not accurate enough in certain cases and its empathized the need for ongoing collaboration between legal experts. Mamalis *et al.* [49] used GPT3.5 and GPT4 and evaluated the models by using three sets of four questions varying in difficulty, of which the models for the first two sets produced reasonable answers, but the hardest questions had more inadequate answers according to experts.

Amri, Bani, and Bani [54] use a Llama-70B and compare this model with other open-weights models on traditional benchmarks, showing that their model displayed better results in comparison.

Reji *et al.* [37] while following the traditional approach for RAG, tested different prompt strategies like CoT and IRAL, a modified version of IRAC [38], and different generation strategies like the traditional Naive greedy decoding, Beam search, Contrastive Search, and Multinomial sampling. Of these techniques, the combination of CoT and Beam search showed the best overall results.

Ryu *et al.* [55] also follow the traditional RAG approach, however, they also use GPT4 to generate questions to possibly improve the retrieval step when the retriever tries to find the relevant documents for a given query. The implementation of RAG brought better results over the use of just LLMs however, there were still some limitations that will be mentioned further.

Louis, Dijck, and Spanakis [64] use two approaches for RAG, one where the Generator training is based on In-Context Learning and the other based on fine-tuning. The In-Context Learning approach works by giving the LLM a context, which is defined by a task description, a test question, and the top-K most pertinent legal provisions to the question as identified by the retriever. As for the results obtained in this study, the retrieval reached superior values compared to baselines, but these results leave substantial room for improvement since less than half of the relevant documents are retrieved due to the hallucinations of the model. However, the fine-tuning of the model showed a more consistent increase in performance.

Besides the traditional method used for retrieval, some studies tried to apply different approaches.

Wiratunga *et al.* [66] used Case-Based Reasoning for the retrieval step by organizing non-parametric memory so that cases are more efficiently associated with queries. This technique increased RAG performance and was shown to be better than the baseline.

Hu, Luo, and Feng [63] retrieval step is divided between cases and articles, so when a query is made, the top-K most relevant articles are retrieved and then a response interpreter generates answers based on a dataset for each article and a similarity is calculated between the answer and each article's sentence. If the similarity exceeds a threshold the answer is given and the same is applied to the case retrieval step. Non-expert users

evaluated both the article and case retrieval, which was noted the potential of these retrieval methods since some answers had a certain amount of relevant information for the questions made.

Garlyal, Hariharan, and Singh [51], for their retrieval, used a Dense Passage Retriever, a retriever based on MIPS to identify the top-K nearest indexed document. For the LLMs used Llama 2, MISTRAL and PHI2, however, the F1 measure had a value between 0.4 and 0.48 across all models, showing some promise but still relatively under-performing.

Moreira *et al.* [72] mention that the limitations of the traditional retrieval methods still depend on the manual effort of experts. This approach uses GPT-4 as a way to create clusters of different types of questions and answers in order to become more efficient and accurate. The results of this technique proved to be satisfactory, but with a Recall of 0.51, which indicates that the clustering process needs refinement.

Zhou *et al.* [82] experiment with a semantic embedding model alongside the use of comparative learning for retrieval. The model was compared with other embedding models like OpenAI's, Baidu and GLM, which achieved better results over them due to the use of comparative learning techniques.

Gao *et al.* [83] implemented a Graph Matching Network based on Graph Similarity Learning with the use of GNNs. A cross-graph matching strategy based on attention mechanisms was also applied in order to capture significant features. For the results achieved, the network was able to pass state-of-the-art models like TF-IDF and BERT.

Hoang *et al.* [67] in terms of IR use a lexical model based on TF-IDF, BM25 Okapi, and Paraformer to determine the relevance between documents and a query and also use ChatGPT to generate both questions and answer. This approach showed good results for the dataset used, although it mentioned the need to do prompt tuning, since this technique, in the perspective of the authors, increases the performance.

Nai *et al.* [53] used a fine-tuned LLM and, for the similarity measure, decided to use Elasticsearch instead of the traditional cosine similarity, and a ranked list with the relevance score of the documents. The evaluation was made by a qualitative method through experts' analysis, which mentioned the answers were relevant.

Alif Adillah *et al.* [8] differ from the previous approaches by using metrics like Recall, Precision, F-measure and Mean Reciprocal Rank to calculate the relevance of a document for a given query. This study shows promising results, but it is mentioned that there is a need to adjust their embedding model.

Zhou, Huang, and Wu [65] focus on analyzing the best methodologies for Legal Retrieval, where, with the aid of experts they tested many models in the identification of salient or relevant content. They used various models such as TF-IDF, BM25, QL, BERT-CLS and BERT-PLI and tested different combinations of them to assess their performance for retrieval. They calculated the attention score for each combination and then applied specific metrics such as Recall and Precision. While this approach did not produce

great results, GPT3.5 was tested instead of the alternatives and produced better results generally.

One aspect we can conclude from these methods is that even though there are distinct approaches to assign relevance scores or organize documentation, most of these approaches apply traditional Encoder-only models such as BERT in their solutions due not only their capability of using the complete query for their embeddings through the use of the bi-directional attention mechanism, but also due to the stability of these models. Besides recent developments with ModernBERT [84], there have not been other innovative or disruptive methodologies for this task, leading to the prevalence of using models such as BERT for retrieval. It is also important to mention that ModernBERT is very recent, and as of the time of writing this dissertation, there have not been Portuguese models based on ModernBERT.

In terms of the results achieved, we can also see the prevalent trend in these solutions being that these models show promise or potential but still have space for improvement with the support of legal experts, specially for the evaluation process, while still being mindful about the ethical implications this technology brings.

3.4.2.5. *Information Extraction*

Information Extraction (IE) is one of the tasks that, while not directly related to retrieval, is still related to the objectives of this study. This led us to analyze solutions that used LLMs as their main approach and see how it compares to other approaches.

One key factor to consider is that IE has other tasks that belong to this task, such as NER, Relationship Extraction, and Event Extraction, which some articles covered. As such, we can see the Table 3.7 with the approaches used by the studies.

Table 3.7. Information Extraction Tasks

Tasks	**# of Studies**	**Citations**
General Information Extraction	5	[85], [56], [68], [73], [71]
Relationship Extraction	3	[62] [86], [52]
Entity Extraction	6	[62] [86] [57] [52] [70], [71]
Event Extraction	1	[57]

As we can see, the studies that have a focus on Information Extraction either extract what the authors would consider relevant information, categorized by "General Information Extraction" or specific information such as relationships, entities and events.

Amorim *et al.* [85] contrary to other solutions built a Python package to extract information while not ignoring the narrative structure without the use of LLMs, instead they used GPT3 as a point of comparison to test their solution's performance. The results obtained by the authors were comparable to the state-of-the-art and GPT3, however, in future work, they intend to test LLMs.

Coelho *et al.* [71] show traditional methods such as Text Classification and NER methods for extraction of legal opinions and how LLMs compare to them. These methods include Naive-Bayes, SVM, Random Forests, Classification LSTM for Text Classification

and BERT for Portuguese NER. The LLM of choice was GPT3.5-Turbo, and in terms of results, it showed to be comparable to other models with space for improvement.

Cho *et al.* [73] built a no-code tool for information search and extraction alongside visualization of statistical information of the structured data. For information extraction, the authors built a custom system, which is a hybrid system between LLMs and open-source SLMs (Small Language Models). In addition to this, they also use LLMs for keyword and sentence extractions. The results the models obtained showed still limitations due to the hallucinations, leading to limitations in the use of the system in formal legal work.

Zambrano [56] tested the capabilities of LLMs using a zero-shot approach for IE of court rulings. They tested Llama and Mixtral with multiple prompts for each legal area. The results showed to be better than the traditional Machine Learning approaches, with Llama as the model that achieved the better results overall.

Zin *et al.* [68] used GPT3.5 to extract information from summaries of contracts, and it was compared to other more traditional models such as RoBERTa and DeBERTa, where even though GPT3.5 had better and comparable performance to the other models, it also had a limitation where the maximum number of tokens was surpassed.

For the other cases and tasks, the overall results achieved for the objective of the respective articles were either an enhancement of results compared to other methods [65], comparable at maximum [57] [62] or not capable of achieving substantial results [86], [52].

Krasadakis, Sakkopoulos, and Verykios [86] tested LLMs in many tasks of IE such as NER, EL (Entity Linking), RelEx (Relation Extraction) and Coref (Coreference Resolution), and their conclusion was that these models fall short when compared to state-of-the-art models in handling these tasks in specialized domain such as the legal domain in non-English languages. Also, they showed difficulties in terms of understanding the terminology, structure and nuances of legal text.

Litaina *et al.* [52] also showed that these models have limitations in combining information between two documents, predicting legal actions, and even identifying the number of relationships between entities, specifically in Greek. The authors also raised ethical questions, bias and non-determinism in the answers, making the solution limited in certain aspects.

One aspect to consider for this task nowadays is the prevalence of LLM based systems, which has led to a smaller number of solutions and implementations of these technologies, as we can see from the number of studies returned from the research query.

3.4.3. Evaluation

Due to the supervised approach mentioned, we can see in Table 3.8 that most studies, in this case 23, focused on following a quantitative approach as their chosen strategy for the evaluation process. Metrics such as Precision, Accuracy, Recall and F1 seem to be the most prevalent across studies, with ROUGE also being widely used. However, there are 10 studies that either apply a more qualitative evaluation or combine the use of metrics

TABLE 3.8. Evaluation Approaches across studies

Evaluation Approach	# of Studies	Citations
Quantitative	23	[64], [65], [71], [51], [72], [52],[53], [66], [62], [73], [70], [67], [68], [74], [56], [57], [58], [37], [8], [2], [61], [7], [75]
Qualitative	5	[63], [49], [50], [60], [5]
Quantitative + Qualitative	5	[9], [10], [55], [59], [69]

with a qualitative approach. In different studies, different methodologies for this type of evaluation are applied, leading to a diverse number of distinct approaches. For example, Cherubini *et al.* [10] use alongside metrics such as Mean Reciprocal Rank, a questionnaire for experts to answer questions related to the quality of the LLM-generated answers. Hu, Luo, and Feng [63] noted feedback from users when they use their respective solution, and Mamalis *et al.* [49] asked experts to develop a set of questions with increasing levels of difficulty for the systems to answer and also used the experts to evaluate the system's answers. Yao *et al.* [50] defined metrics like completeness, reliability, logical coherence, for experts to evaluate their solution on based on their respective opinions.

As we can see, there is not a set methodology to evaluate the answers of these systems, the evaluation method depends on which data is being used and the respective techniques and processes applied, but also if it is possible to access experts to evaluate the outputs of these models based on the task and objective. Most studies conform to the use of classification metrics due to the nature of their dataset and even due to the techniques like LLM-generated question-answer sets, since these metrics are more easily applied.

3.4.4. Limitations

Just like every technology, LLMs possess limitations in their applications across the many fields where such models are applied. Zhang *et al.* [6] in order to circumvent some of the existing limitations of LLMs, they present techniques such as the use of Legal Prompting by leveraging techniques like CoT and Tree-of-Thought, and the use of Legal Embeddings, as best practices. However, the problem with hallucinations still persisted, leading to the use RAG and/or fine-tuned embedding models being considered. With this knowledge, across the literature, many authors presented some of the limitations and their impact on their respective solutions and some of the techniques used to create better solutions.

Espírito Santo *et al.* [59] use different models like GPT3.5 and a fine-tuned LlaMa 2-7B for Question-Answering and evaluated the model through qualitative evaluation. They noted some limitations such as adherence to legal text, hallucinations and inappropriate text concluding that the bias, and the inaccurate and unsafe information present in outputs, prevents formal or practical use of these models.

Ryu *et al.* [55] while using LLMs to generate questions in addition to traditional RAG approach for retrieval enhancement, it is still mentioned that even though the RAG

framework improved the results over solely using LLMs, there is a need for the evaluation process to include expert's evaluation in addition to metrics.

Kaoutar *et al.* [60] addresses the potential of chatbots like ChatGPT, Gemini and fine-tuned open-source models like Mistral-7B-Instruct-v0.2 and Gemma in the legal domain. From the tests executed, the models achieved a very low accuracy, and the authors mention how these models are inaccurate and inefficient. The authors concluded by indicating that without higher computational power, bigger annotated datasets and better consideration for ethical questions these models pose, it is not possible to put these implementations in production despite their potential.

Bakir, Yildiz, and Aktas [69] focuses on using LLama2-7B and Mistral-7B for Question-Answering and how different metrics and similarity measures can be used to improve the performance for the legal domain. They mention that the traditional similarity measures do not capture factual and contextual accuracy, logical consistency and legislation relevance to the question. So, they tested the different existing metrics and used model-based strategies to calculate the congruence between the actual answer and the predicted counterpart by using the pre-trained models mentioned. The author's conclusion is that the use of model-based metrics shows to be more capable than similarity-based metrics in some cases, while affirming the need for more complex and situation-aware evaluation methods.

Cong-Lem, Soyoof, and Tsering [4] focuses on the limitations of LLMs across fields where they register the main problems of this technology. In terms of accuracy, the authors show the technical inconsistencies and inaccuracies in specific domains. For some domains, they also show limitations related to critical thinking, where models were not able to infer more complex answers or even justify them, and were not able to relate concepts in the answer generated, showing limitations in complex reasoning, and even political bias which leads to legal and ethical concerns.

As previously mentioned Zhang *et al.* [6] focus on the best practices, however, they present ethical concerns relating to bias, intellectual property and data privacy. Other studies like Louis, Dijck, and Spanakis [64] also elevate these concerns. Yao *et al.* [50] also mention problems relating to bias, but also show a problem with how answers generated are non-deterministic, leading to a high heterogeneity in the legal domain. Anh *et al.* [2] mentions some limitations for the legal domain, such as the length of the text, language ambiguity, context sensibility, hallucinations, the contextual window based on the fixed number of tokens which in the domain is a challenge, they also mention concerns relating to bias, privacy, security, intellectual property, explainability, transparency, responsible use and others. Quevedo *et al.* [3] also mention other limitations such as data quantity, lack of standard legal procedures across nations, lack of information relating to knowledge modeling, legal reasoning, interpretability, ethical concerns, document length, need for experts to determine relevance, data scarcity, lack of examples for multilingual examples, unbalanced distribution of charges, lack of human evaluation, language ambiguity and

confusion, lack of resources such as benchmarks. Cherubini *et al.* [10] comment on the non-deterministic nature of these solutions and also on other aspects such as text length, query quality and how the results tend to be of better quality when derived from a single document, when compared to when the answer is derived from various documents. Kurniawan and Hiererra [5] also mention concerns relating to data privacy and security.

As we can see, most of these studies encountered the same limitations relating not only to the ethical questions these models bring, such as bias in the training data, intellectual property concerns, data privacy and security, but also issues relating to the non-deterministic nature of these models, the hallucinations they generate, and the need for legal experts for evaluation. In addition to these, concerns related to legal data, such as the reduced data quantity, high text length, data scarcity and language ambiguity are also noted since these limitations impact the performance and even how solutions are implemented. These limitations and concerns pose challenges for this study's solution and are going to be considered for the development of the tool. Some of these challenges have workarounds, which some studies try to solve, however, it is not possible to remove or resolve these challenges completely since most of them are intrinsic to the nature of LLMs, RAG or even the legal domain.

CHAPTER 4

Ask Supreme - Design & Development

This dissertation aimed to build a prototype capable of receiving questions of legal experts related to the Portuguese Supreme Court of Justice (*Supremo Tribunal de Justiça*). The objective of the proposed system is to aid legal experts during the legal search process by implementing an LLM using the RAG framework to reduce the effects of hallucinations, allowing such experts to more easily search for information related to the documents used.

In this chapter, we will present how we followed the CRISP-DM methodology for the development of Ask Supreme, where we will present all phases of development we followed, except the Deployment phase, since our objective is to present a prototype, leaving us to end the development of the system after Evaluation.

4.1. Design and Development Methodology

Considering the data-driven nature of this system, it was opted for the development of the prototype to apply the CRISP-DM methodology. CRISP-DM or the Cross Industry Standard Process for Data Mining methodology consists of a model for data mining projects composed of six phases, these being Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [87].

Business Understanding pertains to the understanding of the business objectives, the assessment of the available resources, requirements, constraints, risks, costs and benefits, the determination of the data mining goals and production of a project plan. Here, possessing a deep level of knowledge and understanding of the domain of the project is crucial.

Data Understanding entails collecting, describing and exploring data and verifying its quality. This can also include analyzing the information the data relates to, which in this case involves having a basic understanding of how the legal data is structured and its meaning.

Data Preparation involves cleaning, formatting and filtering the data. In addition, in this section, we can create new features and/or merge already stipulated characteristics of the data in order to create new features that we can use for modeling or to have a better understanding of the data.

Modeling consists of selecting, building and assessing models to apply to our data in order to fulfill our objectives and goals. In this phase, many models are tested and their parameters altered in order to evaluate what aligns best both with our data and its underlying structure and our goals.

Deployment is related to how the model is deployed and respectively maintained and monitored. In this phase, the final reports and review are also done.

With the understanding of the phases of the mentioned methodology, we can see CRISP-DM's structure in Figure 4.1

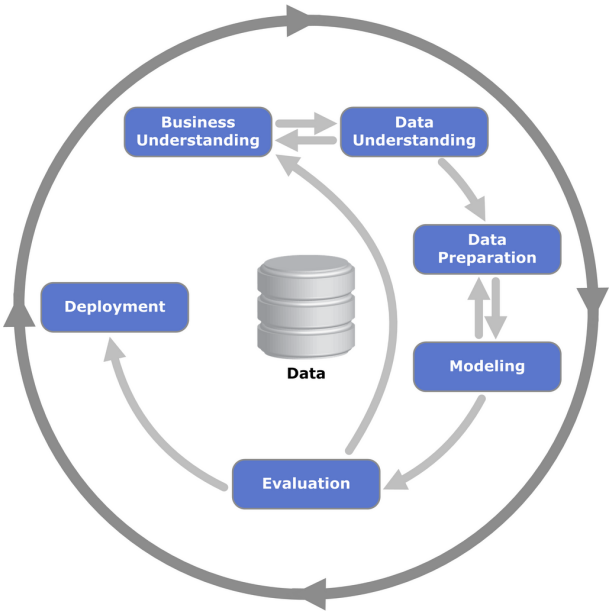

FIGURE 4.1. CRISP-DM Diagram by Kenneth Jensen

Given this methodology, for the development of the prototype, the structure was slightly modified in order to fit the development of this dissertation. The changes consisted of the removal of the Deployment phase. Deployment involves having the infrastructure needed for our system to be accessible anytime, anywhere, by anyone. For this dissertation, deploying the system is out of our scope since we intended to build a proof-of-concept system, leading to the removal of this step.

With this structure as a basis, the next sections in this chapter will be divided into the phases already mentioned, and it will delve into the details, steps and decisions made during said phases of development.

4.2. Business Understanding

Before delving into the development of our system, it is crucial to understand our objectives, the available resources, constraints, risks, and benefits. In our case, these aspects have been presented throughout this dissertation, such as our objectives, where in Chapter 1 we discuss the objective to developing a system capable of supporting legal experts in legal search. Chapter 2 and Chapter 3 describe some of the available resources at our disposal to build such a system, alongside the constraints, risks, costs and benefits as presented more in depth in Sections 3.4.2, 3.4.3 and 3.4.5 in Chapter 3.

4.3. Data Understanding

It is imperative to conduct an initial exploration of the data to assess its structure, quality, and suitability for the objectives of this dissertation. This stage, designated as Data Understanding, serves to provide a detailed overview of the datasets employed and to identify potential issues or limitations that will need to be addressed during Data Preparation.

For the development of this work, two distinct datasets were considered. The first dataset consists of Portuguese court judgments made available through an online repository from the *Conselho Superior da Magistratura*, where we extracted the data from. This dataset includes both metadata and full-text legal content [88]. The second dataset, supplied by ByTheLaw, is a private dataset composed of expert-formulated legal questions and will later be used primarily for evaluation of the system. While both datasets are relevant to the project, this section will focus predominantly on the judgments dataset, as it constitutes the foundation of the retrieval component.

4.3.1. Overview of the Dataset

The judgments dataset contains records spanning from 1931 to 2025, with each entry representing a legal decision. Each record is associated with a set of features, including identifiers such as *ecli*, *tribunal* and *relator*, categorical fields such as *tematica*, *descritores* and *decisao*, textual fields such as *sumario* and *sumarioCompleto*, and temporal attributes such as *dataAcordao*. Additional numeric features, such as *recurso*, *score* and *maxScore*, are also present but are often sparsely populated.

To better understand the meaning and significance behind the features, in Table 4.1, we can see the description of each feature and their meaning relative to a respective court judgment.

Table 4.1. Feature Descriptions

Feature Name	Description
ecli	The European Case Law Identifier (ECLI) assigned to each judgment, providing a unique and standardized reference across jurisdictions.
tribunal	The court responsible for issuing the judgment.
relator	Court reporter who authored or summarized the judgment, serving as an identifier of judicial authorship.
tematica	A set of thematic keywords assigned to the judgment by the reporter, intended to create a thematic index for retrieval
descritores	A descriptive keyword assigned to the judgment by the reporter
decisao	The final ruling or outcome of the case (e.g., dismissal, approval, rejection), representing the conclusion reached by the court.
sumario	Summary of sumarioCompleto
sumarioCompleto	Summary of the judgment containing the questions made by legal professionals and information relative to who made them and their respective interpretation of the arguments presented
dataAcordao	The date on which the judgment was issued, indicating the temporal context of the decision.
recurso	An indicator of whether the case involved an appeal.
score	A numerical score provided in the dataset for retrieval purposes.
maxScore	A numerical score provided in the dataset for retrieval purposes.

A preliminary exploration revealed that the dataset is heterogeneous in nature, combining structured attributes such as dates and identifiers with unstructured textual information. The text features, in particular, vary considerably in length, with *sumario* and *sumarioCompleto* representing different degrees of summarization. *sumarioCompleto* refers to the complete summary of the judgment containing the questions made by legal professionals and information relative to who made them and their respective interpretation of the arguments presented, and *sumario* is a manually written summary of that respective information, hence, *sumarioCompleto* tends to contain longer descriptive passages and more detailed information compared to *sumario*.

Another important factor analyzed was how textual fields such as *tematica*, *descritores*, and *decisao* were populated. These fields are open text fields that are filled by the respective court reporter. This, accompanied by the lack of a standardized and agreed vocabulary for filling these fields, leads to a high degree of heterogeneity between values in those features, which consequently can lead to two different reporters for two distinct cases while filling the respective value in one of the fields mentioned. Such might be perceived as distinct values in the dataset due to how they were registered, even if they might be perceived as equal in the eyes of experts.

4.3.2. Data Quality and Completeness

An essential component of data understanding is the assessment of missing or inconsistent values. Analysis of the dataset revealed that all records contained at least one missing value due to three features, these being *recurso*, *score* and *maxscore*. Excluding these features, we were able to identify 32.97% of all records contained at least one missing value, with the highest proportion of missing entries was *tematica*, with 32.62% of its values absent. Other features such as *descritores* and *decisao* also exhibited missing or incomplete information, though to a lesser extent, 2.14% and 4.46% respectively. To an even lesser degree, *dataAcordao* contained only 0.22% or 157 of its records missing, and both *sumario* and *sumarioCompleto* only had 0.03% or 22 records missing. In contrast, identifier features such as *ecli*, *tribunal*, and *relator* were completely populated across all records.

Beyond missing data, as already mentioned, the dataset also presented issues of inconsistency. Features *decisao*, *tematica*, and *descritores* lacked a controlled vocabulary, which resulted in semantically similar documents being assigned to distinct labels. Similarly, certain textual attributes, particularly *sumarioCompleto*, included extraneous elements such as HTML tags and metadata that do not contribute meaningfully to legal analysis.

4.3.3. Feature Analysis

With a basic understanding of the structure of the dataset and its contents, we were then able to analyze each feature individually and draw some conclusions from this analysis.

The first feature we analyzed in more detail was the *dataAcordao* feature, where we analyzed the temporal distribution of the records. As illustrated in Figure 4.2, the dataset

covers nearly a century of judicial activity. However, the density of judgments increases significantly in more recent decades, reflecting the growing availability of digitized legal resources.

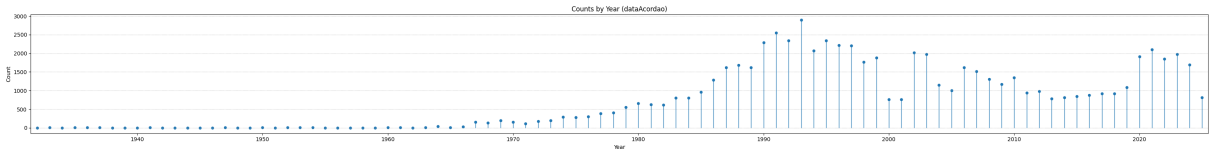

FIGURE 4.2. Distribution of Documents over Time

In order to analyze the effects of the semantic inconsistencies mentioned above, we opted to visualize the distribution of values for those respective features. We can see the respective distributions in Figures 4.3, 4.4 and 4.5.

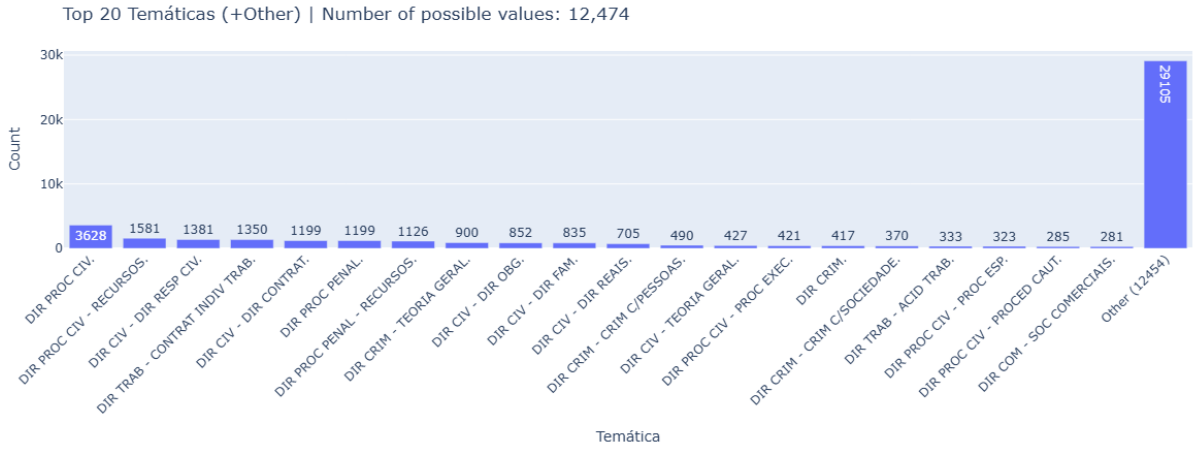

FIGURE 4.3. Top 20 values of column *tematica*

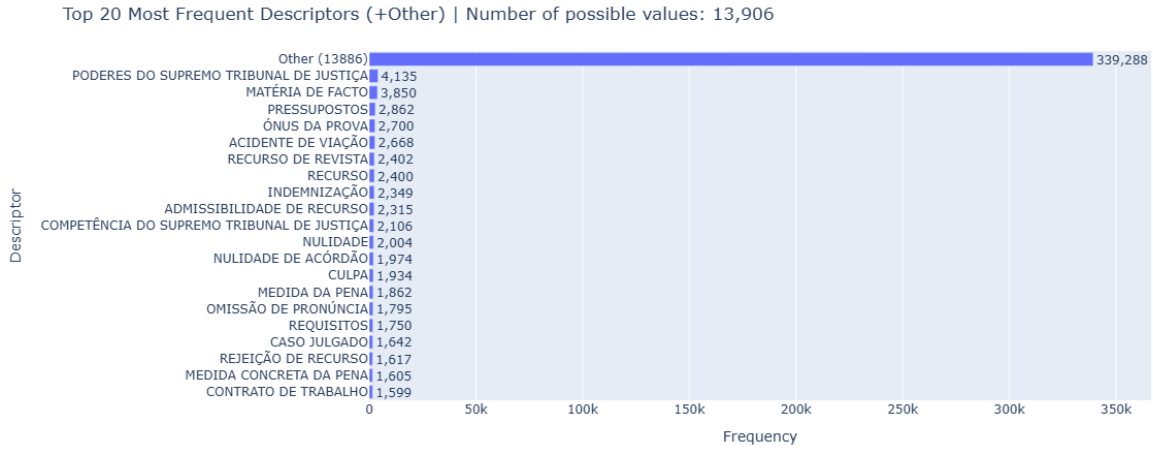

FIGURE 4.4. Top 20 values of column *descritores*

In Figure 4.5, we can see that the first column in the figure represents a collection of values contrary to the other columns which represent a single one. This was due to the high number of possible values the column contains, so in order to be able to visualize some

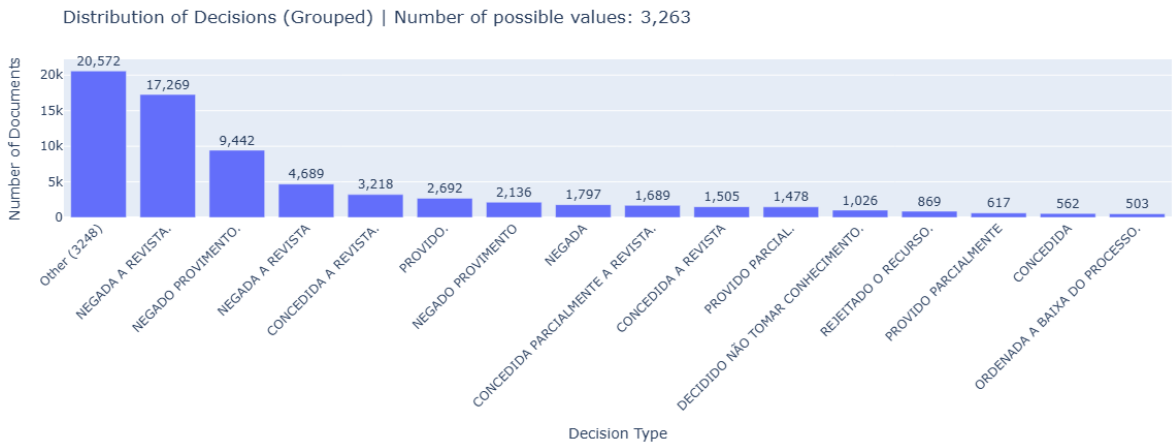

FIGURE 4.5. Distribution of values from column *decisao*

of them, all values with counts inferior to 500 were grouped. The 20527 values represented in the "Other" column belong to 3248 distinct values, so we can then conclude there is high variability of values for this column.

As we can see, these columns contain high variability in their respective values, even if there are certain values that are more represented. In some instances, particularly in features *decisao* and *descritores*, there seems to be some overlap between values where they could represent the same information, but they are recognized as different in a computer system. This might be considered a limitation of the dataset that can impact the performance of the retrieval process, since in order for the system to better retrieve the most similar documents, some degree of consistency across features is needed. However, *decisao* is a feature of great importance to the dataset since it represents the judgment's final decision in court, leading us to consider its use on the proposed system.

4.3.4. Data Selection

From the analysis above, it becomes evident that not all features contribute equally to the objectives of this dissertation. Attributes such as *decisao* and *sumarioCompleto* emerge as central to supporting the retrieval process, while features with high rates of missing values or inconsistent categorization are less reliable. Moreover, understanding the temporal scope of the dataset highlights the trade-off between completeness and efficiency, guiding the choice of a reduced yet representative time interval. With this, it was opted to use the features *ecli*, *sumarioCompleto*, *decisao* and *dataAcordao* as the foundational data for our system.

4.4. Data Preparation

With the knowledge from our data obtained from the previous section, it was possible to advance into the Data Preparation phase of the development. As previously mentioned, this step involves cleaning, formatting and filtering our data in order to prepare for the Modeling phase. For the purposes of this dissertation, two distinct datasets were used, one which contains the legal text of the Portuguese judgments, and one which contains

legal questions made by experts. The first dataset will serve as the retriever's knowledge base, and the second will be used for the evaluation of this tool. In this section, we will focus on the first dataset mentioned, and the second will be presented more deeply in Chapter 5.

4.4.1. Missing Data and Feature Removal

Firstly, empty columns were removed, meaning features which did not contain values were removed from the dataset, since there is no data, information or knowledge we can take advantage of from using these features. The features were, *score*, *maxScore* and *recurso*. In addition to these features, identifier columns such as *tribunal* and *relator*. Such columns were removed due to their nature as identification columns, meaning they unequivocally identify the registry, in this case, who recorded the data and where. For the objective of this dissertation, these features are not needed, since not only is the column *tribunal* filled with only one value, *Supremo Tribunal de Justiça* in this case, but also information about who wrote the information does not aid in legal search according to the definition presented in Chapter 1.

After removing these columns, we analyzed the number of records that contained missing values, which comprised 32.967% of the available data. Such a percentage is due to the amount of missing values of feature *tematica*, which has 32.62% missing values as we have seen in the previous section. While other features also contain missing data, their respective quantity is considerably low to the point that those registries can be removed without worrying about the impact such process will have on the system's performance.

It is also important to mention how in conversation with one expert, when they do legal search, they do not commonly use the details such as *tematica* or *descritores*, but they find the decision to be crucial when researching. According to this and considering these features as open text fields, it was opted to remove them since there is not a delimited vocabulary of possible categories for their respective values, since they are open text fields. Such can lead to similar documents having distinct categories in either or both features, which can lead to situations where similar cases are distant in the vector space, significantly impacting accuracy. Although the feature *decisao* also has the same nature as the mentioned columns, this feature possesses great importance for legal experts when trying to research similar cases.

4.4.2. Filtering

As presented in the previous section, our dataset contains data from 1931 to 2025 (as seen in Figure 4.2). With the aid of a legal expert, it was discussed which time interval would better fit for the system, since when doing legal search, specially related to these types of documents, it is not needed to use all historical documents. In discussion with experts, it was defined that an interval between 15 and 20 years would be the best fit for the system. With this knowledge, the data was then filtered according to the 20 year interval, so our

data includes documents from 2005 to 2025, considerably reducing the amount of data, which can be a benefit since it reduces processing time.

4.4.3. Cleaning

Given the chosen features to keep for modeling and the respective time interval, it is important to analyze if the data, specially textual data, contains characters or sets of characters that do not contain understandable information. These characters can be special characters and/or numbers spread across the text, making it less legible for humans and less "understandable" for computer systems. After analyzing the columns *decisao* and *sumarioCompleto*, there were some aspects which we could clean in order to make the text more usable for modeling. For the feature *decisao*, it was noted that every registry was upper-cased and some of them included a "." at the end, which in certain cases made so that the same text was considered a different category due to the presence of the period, so all registries were changed to lower-case and the periods removed. For *sumarioCompleto* the presence of HTML delimiters and metadata, which for the purposes of the system are not needed, were then removed, leading to each registry containing only the text associated with its respective document. The feature *dataAcordao*, which corresponds to the date of emission of the document, was modified in terms of its type, making it so the system could identify that said column contains information related to the date of the document instead solely text.

4.4.4. Chunking

When dealing with LLMs, one step many tend to include in the pre-processing phase is chunking. Chunking or text splitting consists of dividing large pieces of text into smaller pieces. This process is considered crucial for the development of RAG because when we are dealing with large documents or even large datasets, chunking allows a more efficient retrieval while also allowing us to retrieve more specific information, enabling faster and more accurate results [89]. The process of breaking down large pieces of text can be done in multiple distinct ways, such as by characters, tokens, paragraphs, pages or other elements of a given document structure. However, it is also possible to do semantic or contextual chunking, where chunking is done based on meaning or context, respectively, possibly with the aid of an LLM. Another aspect to take into consideration is the limitation of the token size LLMs usually employ, meaning, for most LLMs after a certain number of tokens is reached, the model can not receive more data even if there is more information, so this process is crucial to preserve the whole context of our data while mitigating truncation of text by the LLM.

As mentioned in Section 4.3.1, our *sumarioCompleto* feature contains information related to the questions and interpretations made by experts throughout the case. In this dataset, the questions or interpretations posed are presented in the form of a list, with each row being assigned a Roman numeral indicating when they were declared in court. Given this structure, it was opted to chunk the data based on Roman numerals

of each registry, meaning each chunk would represent a single question or interpretation for a given judgment with its respective Roman numeral. This allows us to have a more concise piece of text while also being specific enough to benefit from chunking. Figure 4.6 illustrates the resulting distribution of chunk sizes. In total, 99,372 chunks were obtained, with an average length of 84.85 tokens. The distribution of tokens per chunk shows that most text segments remain relatively short. This property will later be advantageous when aligning the chunks with expert-formulated queries.

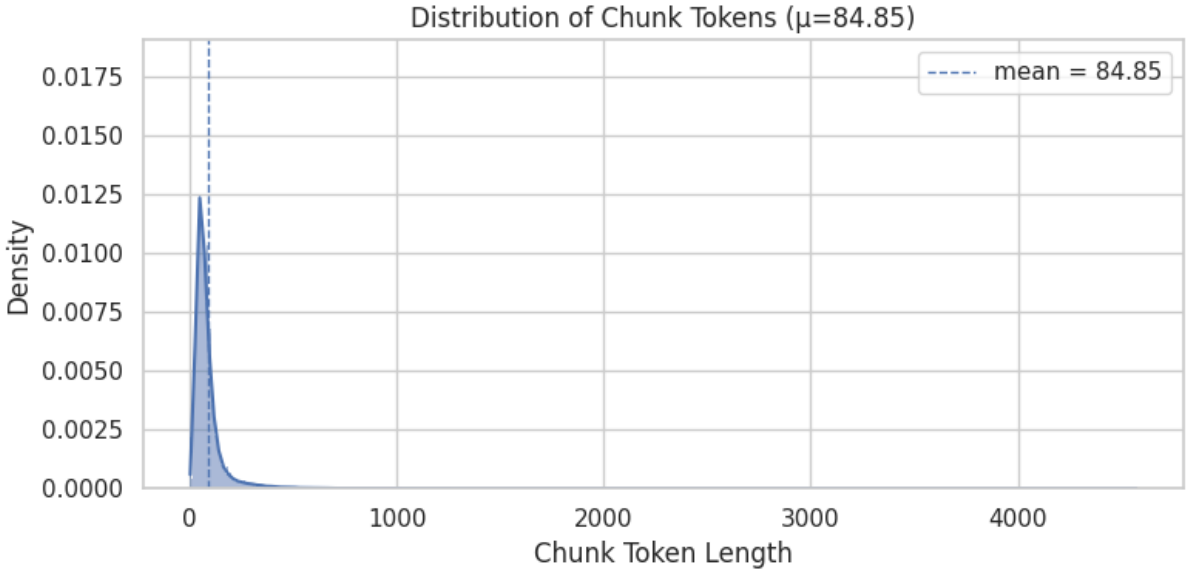

FIGURE 4.6. Distribution of Chunk Token Length

4.5. Modeling

With the Data Preparation phase completed, the finalized dataset is now ready to be utilized in the Modeling phase of the process. In this section, we will focus on presenting the objectives of this stage, the models selected and used, the retrieval parameters, and the orchestration framework of the overall pipeline.

The goal of the modeling phase was to construct a system capable of answering legal-domain questions accurately and coherently by leveraging both the generative capabilities of an LLM and the contextual knowledge of a legal judgments database. By building this system, it will be possible to answer one of the proposed Research Questions, mainly "How can we extract relevant information for legal experts from Portuguese legal documents?".

This section will present the multiple components developed for the Ask Supreme system alongside the respective details, techniques, prompts and models employed.

4.5.1. System Overview

Figure 4.7 presents the core components of the system, and this section will delve into the components shown.

The integration of these components was implemented in Python, enabling the development of a functional pipeline capable of handling expert-formulated queries. During

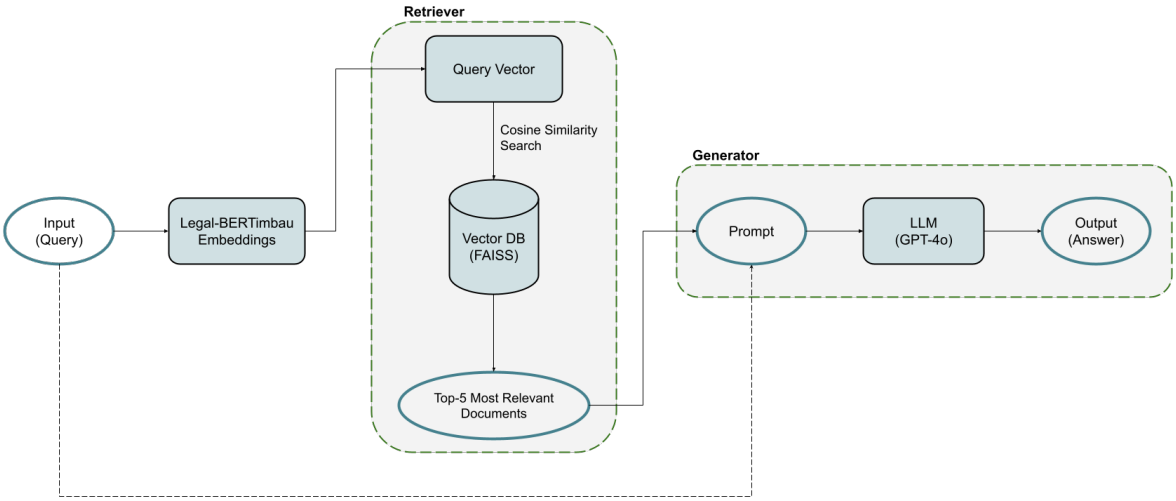

FIGURE 4.7. Ask Supreme Pipeline

inference, the system first retrieves semantically similar chunks from the knowledge base using cosine similarity. These retrieved segments are then combined with the user query and passed as context to the generative model, which produces the final answer.

4.5.2. Embeddings

With our data prepared to start the modeling, before delving into the components of our system, we need a way of creating and storing the embeddings of our data. This step involves encoding the processed data into vector representations that populate the system's knowledge base, storing them in a vector database so our RAG system is capable of retrieving information relating to a user's query.

First, in order to create our embeddings, we need to choose an embedding model to use. Traditionally, there are three main approaches when using Transformers. The first is using a base or agnostic model, such as BERT and its variations, and applying it directly to our data. The second is using already fine-tuned models that might have been trained on similar contexts, or trained on the same language or even on our data. The third is to fine-tune or train our own model with the resources available to us. In our case, since our data relates to the *Supremo Tribunal de Justiça* of Portugal, our data is related to the legal domain and in Portugal's Portuguese. There is an existing model already fine-tuned for our context, the Legal-BERTimbau [90]. Such model was chosen given how well the model meets the linguistic and domain specificities of this dissertation. Due to the limited amount of data, we chose to use local embeddings, meaning, embeddings which are saved locally on a machine in order to more efficiently process the retrieval step.

Having the embedding model ready to use, now we need a way to store the data so that our RAG system can access it and retrieve the documents needed when prompted. For our solution, since we are using a limited amount of data, we opted to use FAISS [91] given its efficient handling of local embeddings and seamless compatibility with the

selected framework. This decision was reinforced by the practical constraints encountered during experimentation, where alternative databases, such as Pinecone [92], necessitated greater resource expenditure. For this dissertation, we opted to assign metadata to the embedded chunks. The metadata assigned were the data related to the *ecli*, *dataAcordao* and *decisao*. By embedding this information, we support queries based on different facets of the documents, which can be useful for experts. With this, we were then capable of creating 99,372 text embeddings alongside the respective metadata related to the identifier of the document, the date it was written and the respective final decision.

4.5.3. Orchestration

To integrate the generative model, the embeddings, and the retrieval mechanism into a unified system, the LlamaIndex [93] framework was selected. LlamaIndex offers substantial capabilities for the development and validation of RAG systems, facilitating modular construction and experimentation with various configurations, making it a suitable choice for this dissertation. It is also important to mention its direct compatibility with FAISS, allowing us to more easily use the database and efficiently process the retrieval step.

4.5.4. Retrieval

The core component of the RAG framework is the retriever, which is responsible for identifying the most relevant documents to support generation. We followed a similarity-based retrieval approach, where among the various distance metrics applicable to high-dimensional embeddings, **cosine similarity** was adopted due to its widespread use and proven effectiveness in high-dimensional vector spaces.

In addition to the similarity function, several retrieval parameters were defined. Specifically, for each query, the **top-5** most similar chunks were selected. Retrieved segments were concatenated and passed to the generative model as context, ensuring both coverage of relevant material and adherence to token limits.

An important aspect to mention is how the metadata used in our embeddings impacts the retrieval process. By using metadata with our embeddings, the retrieval process can better filter our data by the respective metadata, and if in our dataset there are similar chunks but with different metadata, our system will better recognize those differences and retrieve the information most related to the query. An example of this is, if we asked the system for documents from 2020 to 2025, when retrieving, given we have the date as metadata, the model would be able to recognize that filter and the documents retrieved would only be from the interval mentioned. With this, the usage of metadata gives us better performance for the retrieval process and more specific filtering and selection of relevant information.

4.5.5. Generative Model

The second component of the RAG architecture is the generative model. This model is responsible for producing natural language answers conditioned on both the user's query and the retrieved contextual documents. Guided by the analysis of related work and

the prevalence of OpenAI models in the literature, the **GPT-4o** model was selected. This choice was motivated by its state-of-the-art performance in text generation and multilingual capabilities, which align with the requirements of processing Portuguese legal texts.

To guide the model's behavior and ensure legally grounded responses, a structured system prompt was designed and employed. The following prompt was used [1]:

> Usa o contexto seguinte para responder à questão posta. O contexto refere-se a subconjuntos de acordãos portugueses de Portugal do Supremo Tribunal de Justiça. As respostas devem ser respondidas em português de Portugal. Caso o contexto não seja suficiente para responder à questão responde que não é possível responder à questão dada com a informação disponível.
>
> Contexto: [context]
>
> Questão: [query]
>
> Resposta:

For this component, our embeddings' metadata aids in mitigating hallucinations by granting the generative model more context besides solely the text retrieved. This is possible because before the LLM generates text, it will analyze the context's text and metadata and then generate the data. For example, if we specify in our query that we want to analyze the decisions of similar cases to a specific one we described, after the documents are retrieved, the LLM will be able to see the metadata related to the decisions assigned to the documents and use that information in the generated answer. This detailed context reduces the chance of our model hallucinating, giving us more reliable answers.

4.6. Evaluation

While traditionally in this section we would present how we evaluated our system, and the respective results achieved and discussion, our research methodology is the Design Science Research Methodology, and this methodology also has an evaluation phase. With this, we opted to present these details in Chapter 5. This is possible to do due to the iterative nature of both methodologies, meaning we are able to develop our system iteratively within our design and development methodology (CRISP-DM) while also doing the same within our macro research methodology (Design Science Research Methodology).

[1] Example of prompt usage in Appendix A

CHAPTER 5

Evaluation

With the RAG pipeline built, we can proceed with the evaluation step. This phase consists of developing a set of experiments with the pipeline through an iterative process where we assess the model's performance based on metrics and then proceed to make adjustments to our pipeline accordingly in order to achieve reasonable results while fulfilling our objectives. In addition to this, we will present and discuss the results achieved during the evaluation of the proposed RAG system and determine the extent to which it meets our objectives defined in the earlier stages of this dissertation and if it answers our research questions "How can we extract relevant information for legal experts from Portuguese legal documents?" and "How can we validate and evaluate a question-answering system for the legal domain without a dataset of questions and answers annotated by experts?".

The chapter is structured as follows. First, the evaluation methodology will be presented, where we delve into how the process was conducted and also present some of the core concepts and techniques applied. Second, we will go into detail on how developed the question generation method and show examples of generated questions. Third, we will explain the iterations or experiments done in order to improve our system's performance. Finally, we will discuss our results and their significance to this study.

5.1. Evaluation Methodology

Before we proceed to explain the evaluation process, it is important to share how we assessed the system's performance. As mentioned in Chapter 2, traditionally, experts make use of quantitative and/or qualitative measures to assess the tool's effectiveness, accuracy, and overall performance in relation to its intended purpose. While for this specific use case, using qualitative measures would bring us more in-depth knowledge about our results, due to time constraints, for this dissertation, we opted to use a quantitative measure in order to evaluate our results and determine some of its possible limitations.

One recent contribution developed for the process of evaluation of RAG systems is the RAGAS [94] framework. RAGAS allows us to evaluate our system by relying on automated metrics rather than annotated datasets, making it a potential alternative for evaluating our prototype's performance. The way this framework works is by first, evaluating the retrieved context for a given query, measuring metrics such as relevance, precision, and recall of the retrieved documents with the aid of an LLM. Next, it assesses the generated responses, analyzing their faithfulness to the retrieved context, semantic alignment with the query, and overall answer quality, also with the aid of an LLM. Finally, it combines these metrics to provide a holistic assessment of the system, capturing how well the RAG

model retrieves information and generates accurate, context-aware answers. However, this framework possesses some limitations, specially for the legal domain, since it uses LLMs to analyze if an answer answers a question and if it is related to the context. This is a problem, since it makes the evaluation process too opaque and assumes a generalist LLM is capable of evaluating the task correctly, which for the legal domain and in our case, where we can not fine-tune our LLM due to a lack of resources, we believe it is not appropriate.

Considering this, we then propose a new evaluation methodology when we do not have access to annotated datasets inspired by RAGAS. This methodology is capable of being applied even when not having access to the ground-truth, since we will generate examples of questions based on original texts through Bloom's Taxonomy. Like RAGAS, this methodology will analyze the relation between a question and an answer, and the relation between an answer and the retrieved context. Instead of using an LLM to measure the relation between pairs, we will use BERTScore.

The BERTScore metric [46], as presented in Chapter 2, is used for Transformers-based models, such as BERT, and computes cosine similarity between candidate and reference sentences. The reference sentences represent our standard text, which the candidate, our output text, will compare against by aligning the tokens between sentences and computing the cosine similarity between them, producing a score. The scores are then aggregated to produce Precision, Recall and F1 measures.

In our case, the F1 measure will be the primary analyzed metric since it represents the harmonic mean between Precision and Recall. By leveraging the contextual embedding from BERT and matching them with a similarity measure, we are capable of evaluating how semantically close our text is between different components, in this case, how semantically close our answers are to our questions, and how close our answers are to the retrieved context. This metric does not assess if the information generated is correct or applicable to the respective questions. In order to be able to discern if a given answer is appropriate to a question and if the context is useful to answer the question, we would need to apply a different evaluation methodology based on qualitative metrics, be it through reporting user tests, expert evaluation or using a questionnaire to assess experts' opinions on the generated results. However, as mentioned, due to time constraints, it was not possible to incorporate such techniques for this dissertation.

With this metric and our dataset, we then apply BERTScore between (i) the system's output (answers) and our questions; and (ii) the system's output and the retrieved context. By using such metric, we are able to analyze the semantic alignment of our answers and determine if the model is capable of generating answers closely aligned with our questions and the retrieved context, which in turn, reflects the system's ability to generate legally grounded responses and helps identify if hallucinations were mitigated.

5.2. Question Generation Methodology

Given this new evaluation framework, as mentioned, we need to generate examples of questions based on original texts. To do this, we developed a method of question generation based on Bloom's Taxonomy [95] in order to generate specific types of questions we could use to evaluate our system and to specialize it in. This section will first present the core concepts of Bloom's Taxonomy. Second, demonstrate the question classification process based on the taxonomy. Finally, show how the questions were generated and some examples we generated.

5.2.1. Bloom's Taxonomy

As mentioned in Chapter 4, we have a second dataset containing legal questions made by experts, which we could use for this phase, however, a great majority of the questions made are not related to our legal topic of choice, the *Supremo Tribunal de Justiça* judgments, and there is a vast diversity of topics and types of questions in the dataset. Due to these constraints, it was opted to first categorize the existing questions and use an LLM to generate questions of a set number of categories in order to have a way to create questions in a topic and for our system to answer them.

While there is not a standard agreed method for question classification, specially for our context, during research we found Bloom's Taxonomy [95]. This taxonomy is a recognized hierarchical framework used by educators to classify and structure educational objectives according to their complexity and specificity. There are two versions of this taxonomy, the first one was created in 1956, and the other is a revised version created in 2001, which is the version we will apply for the purpose of this dissertation. This revised taxonomy has altered the categories from the previous version and now includes action verbs associated with each of the six aspects of cognition, which we took advantage of when classifying the questions.

Bloom's Taxonomy encompasses three primary domains: cognitive, affective, and psychomotor. For the purposes of our dissertation, we will focus on the cognitive domain, which contains six categories and a knowledge dimension with four other categories. The core six categories are as follows:

- Remember: Be able to recall information such as dates, events, places, ideas, definitions, formulas, and theories;
- Understand: Be able to grasp the meaning of the information, express it in own words, and/or cite examples;
- Apply: Be able to apply knowledge or skills to new situations. Use information and knowledge to solve a problem, answer a question, or perform another task;
- Analyze: Be able to break down knowledge into parts and show and explain the relationships among the parts;
- Evaluate: Be able to judge or assess the value of material and methods for a given purpose;

- Create: Be able to pull together parts of knowledge to form a new whole and build relationships for new situations.

As for the knowledge dimension and the four distinct types of knowledge which the above categories can apply. They are as follows:

- Factual Knowledge: Pertains to knowledge which involves understanding terminology and details;
- Conceptual Knowledge: Involves understanding the relationships among pieces of concepts or theories;
- Procedural Knowledge: Consists of understanding processes and methods of theories and problems;
- Metacognitive Knowledge: Relates to learning strategies and processes.

In Figure 5.1 we can see examples of the action verbs associated with each category, and in Figure 5.2 we can see the relation between categories and knowledge dimensions:

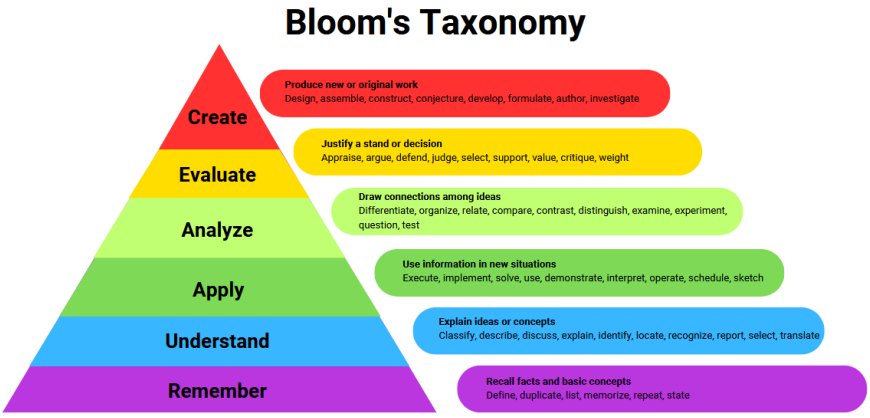

FIGURE 5.1. Diagram of Bloom's Taxonomy with Action Verbs [96]

5.2.2. Bloom Classification

By using Bloom's taxonomy as a basis for our question classification process, we are able to use keywords or key expressions in order to assign one of the six categories to each question in the dataset. For this process, two different approaches were used: one rule-based approach and an LLM-based approach.

The rule-based approach was made by building a function that, based on a set of Portuguese keywords and semantic expressions assigned to each Bloom category, the system looks to each question and counts how many expressions from the different categories are in said question, where the category which has the most representation present in the question will be the one assigned to the question. In the case where two or more categories are equally represented in a question, we use the category hierarchy to determine which category is assigned, and in the case where no category is represented, the question is

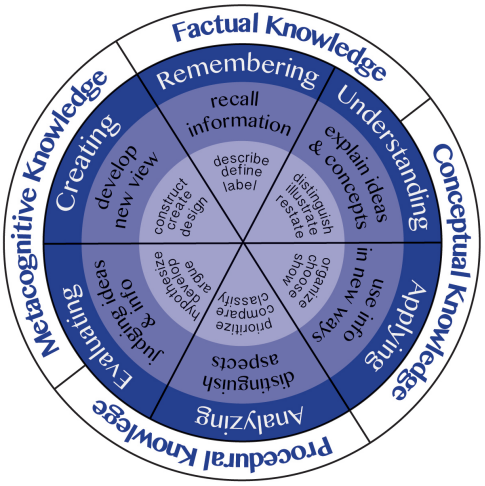

FIGURE 5.2. Diagram of Bloom's Taxonomy with Knowledge Dimensions [97]

TABLE 5.1. Keywords and Expressions for Bloom Classification

Bloom Category	Keyword and Expressions
Remember	qual; quando; quem; quais; onde; quanto; (que /no) (artigo / artigos); que lei / que acórdão; resumidamente; sabes; o que é; o que são
Understand	o que significa; qual o significado; interpretação de; como funciona; para que serve; fala-me; descreve; explica; define; porquê; para que; como
Apply	se eu; em caso de; ocorrer; supõe que; é aplicável; quid juris; situação concreta; aplica
Analyze	diferença entre; compara; analisa; distinção; contrasta; análise; diferencia; relaciona
Evaluate	é legal; é justo; opinião; deveria; achas que; concorda; acha que; opiniões
Create	resume; faz um resumo; reformula; inventa; elabora; cria; sumariza; formula; desenvolve

classified as "Uncategorized". We can see the list of keywords used for the classification process in Table 5.1.

In Figure 5.3 we can analyze an example of how sentences are classified through the rule-based method.

For the LLM-based approach, GPT-4o was used to classify each question based on its understanding of Bloom's taxonomy. By analyzing its Chain-of Thought, it was possible to analyze how the classification process was made, and it followed a similar approach to the rule-based method used, however, it contained a different set of keywords and expressions. The prompt used for the classification process was the following:

> I want to classify legal Portugal's portuguese questions made by users according to the 2001 revised version of Bloom's Taxonomy. I will give

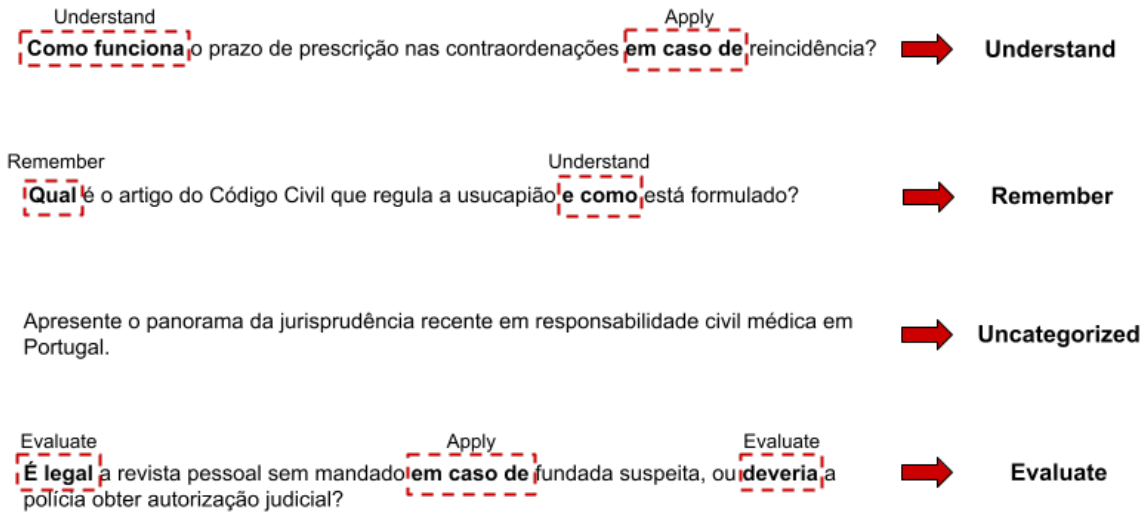

FIGURE 5.3. Rule-based Bloom Classification Example

context for the taxonomy, the data that contains the questions and the objective you will have to do.

Bloom Context: " A group of educational researchers and cognitive psychologists developed the new and revised Bloom's Taxonomy framework in 2001 to be more action-oriented. This way, students work their way through a series of verbs to meet learning objectives. Below are descriptions of each of the levels in revised Bloom's Taxonomy:

- Remember: Be able to recall information such as dates, events, places, ideas, definitions, formulas, and theories. Action verbs examples: list, define, identify, recall, name, state, describe

- Understand: Be able to grasp the meaning of the information, express it in own words, and/or cite examples. Action verbs examples: explain, summarize, interpret, paraphrase, illustrate, classify

- Apply: Be able to apply knowledge or skills to new situations. Use information and knowledge to solve a problem, answer a question, or perform another task. Action verbs examples: apply, use, demonstrate, implement, execute, solve

- Analyze: Be able to break down knowledge into parts and show and explain the relationships among the parts. Action verbs examples: analyze, compare, contrast, examine, differentiate, categorize

- Evaluate: Be able to judge or assess the value of material and methods for a given purpose. Action verbs examples: evaluate, assess, critique, justify, argue, defend, rate

- Create: Be able to pull together parts of knowledge to form a new whole and build relationships for new situations. Action verbs examples: create, design, invent, develop, construct, compose, propose

The hierarchy goes as the following (bottom to top): Remember > Understand > Apply > Analyze > Evaluate > Create

Knowledge Dimension: Bloom's revised taxonomy (by Anderson and Krathwohl, 2001) introduces the Knowledge Dimension, which complements the cognitive domain by clarifying what learners are expected to know.

While the cognitive domain describes how learners use and process knowledge, the Knowledge Dimension classifies the type of knowledge students engage with:

- Factual Knowledge: Knowledge of basic elements, terminology, and details essential to a discipline (e.g. historical dates, mathematical symbols, scientific terms)

- Conceptual Knowledge: Understanding relationships among concepts, principles, and theories (e.g. grasping the principles behind economic theories like supply and demand).

- Procedural Knowledge: Knowing how to perform tasks and processes (e.g. conducting laboratory experiments or following a mathematical algorithm).

- Metacognitive Knowledge: Awareness and understanding of one's own thought processes (e.g. reflecting on study habits to improve personal learning strategies)."

Data Context: The data is a csv file with 1240 registries with a column named "content". Each registry contains Portugal's Portuguese question(s) made to a Chatbot where the subject matter of said question(s) is related to legal matters.

Objective: Create 2 columns in the data called "bloom category" and "knowledge dim" where for the first column the value for each question will be one of Bloom's Category and the second will be the respective dimension. In case a given question/text does not contain enough information to be classified, or the text does not have a question, categorize it as "Uncategorized" for either column; if the text involves a resolving an exercise or a very specific request that ca not be categorized in one of the given categories, classify it also as "Uncategorized".

We found the model was able to better fulfill the prompt's request by using an English prompt instead of Portuguese, and while the model does possess information relating to Bloom's Taxonomy we found it better to use a given interpretation in order to avoid possible hallucinations and misinterpretations outside of the preconceived one we possess.

In terms of the results of the classification between both approaches, we can analyze Figure 5.4.

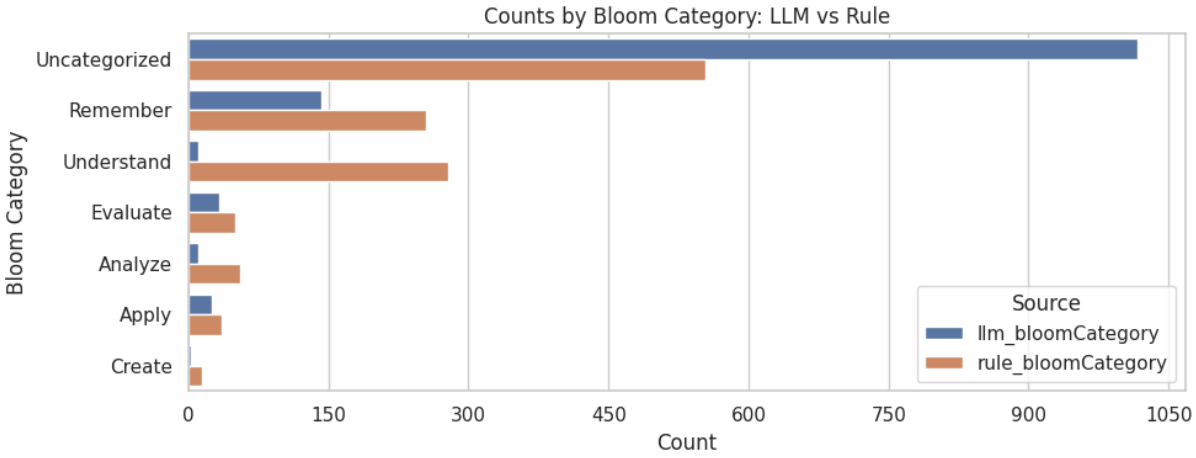

FIGURE 5.4. Comparison between Rule-based and LLM-based Classification

As we can see in Figure 5.4, the LLM process tends to classify more questions as "Uncategorized". The reason behind such phenomenon, after analyzing the Chain-of-Thought of the LLM, was that the model used fewer keywords than the ones used by the rule-based method, leading to the LLM classifying more questions as "Uncategorized".

By analyzing Figure 5.4 and examples from both methods, we are able to understand which type of questions we want to focus on for our model. Taking into consideration the objectives of this work and the types of questions most present in the dataset, it was opted to use questions relative to the Remember and Understand categories since those are the most well represented types of questions in the dataset. Given this information, we can then proceed with the generation of questions.

5.2.3. Question Set Definition

As mentioned previously, the question dataset did not possess questions relative to matters of the *Supremo Tribunal de Justiça*. It is imperative to have questions related to our topic in order for our system to be able to answer them with accuracy and reliability. Due to this, it was determined that it would be necessary for experts to either create a set of questions we would use or generate them. The chosen strategy was the latter, since due to constraints, it was not possible to have expert-made questions.

With this information, and with the aid of GPT-4o, we proceeded to generate 100 questions based not only on the data relative to the *Supremo Tribunal de Justiça* but also according to Bloom's categories, in this case, Remember and Understand. The prompt given to the system was the following:

> Based on these examples I want you to act as a Legal Expert (Lawyer/Judge) and generate questions based on the annexed data file. I'll give the context for the data and I'll give your objective alongside some restrictions. Be mindful the text from the data and the text you'll generate will be

in Portugal's portuguese. The questions to be generated must belong to the Remember and Understand dimensions of the Bloom taxonomy.

Data context: 97300 registries that are segments of judgments ("acórdãos") of the "Supremo Tribunal de Justiça". The data contains 4 columns: ecli, chunk, date and decision. ecli refers to the document identifier which the segment belongs to; chunk contains the text of said segment; date refers to when the document was emitted; and decision refers to the final decision given by the Supremo Tribunal de Justiça.

Objective: First generate 20 examples of questions. The questions MUST be able to be answered by looking only at the dataset given, avoid mentioning questions to other legal documents or legislation not mentioned in the data. Questions must be general but specific to the dataset, meaning, avoid general questions such as "O que diz o artigo nº X?", "Quais as decisões mais comuns do STJ?", "O que significa X decisão?", "O que o STJ quer dizer com decisão Y?", "Em que datas foram proferidos acórdãos com decisões de absolvição?", "Quais são os diferentes tipos de decisão final registados nos acórdãos do STJ?", "Em que segmentos são referidos os fundamentos para a prescrição da pena?", assume the legal expert you'll act as already has a basic understanding of Portugal's legislation and legal system.

In addition also avoid questions that can only be answered by a single instance or document of the dataset such as "Qual crime foi julgado com base no art. 24º neste acórdão do STJ?", avoid mentioning specifically a document or segment. The questions you'll generate must be in the middle, so not general enough or refer to the dataset or even analyze it, and also not specific enough so that only a single segment or document can answer. In addition also avoid mentioning either the segments or even the dataset so when generating questions avoid saying "Em que segmentos...", "Segundo o conjunto de dados / dataset / segmentos...". Take advantage of the other columns, keep the structure and the main focus of the questions still about the text but add aspects to the decision or even the date.

Use this template to generate the questions: Question Template: " X. (X = Number of the question) Questão: (question) Bloom Cognitive Dimension: (Remember or Understand) Knowledge Dimension: (Respective Knowledge Dimension "

Given the prompt shown and some adjustments to the results given, we were able to obtain the questions we would use for our evaluation process. Here are some examples of the generated questions:

1) Que norma exige unanimidade dos condóminos para alterar o uso de uma fração?

2) Como o tribunal pondera a proporcionalidade da reação do arguido face à ofensa sofrida?

3) Porque a improcedência de pedido principal pode tornar irrelevante o conhecimento de pedidos acessórios?

4) Qual é a consequência legal da ausência de "declaração formal" na aceitação de cláusulas contratuais?

5) De que forma o STJ trata a conexão entre erro-vício e erro-obstáculo em litígios contratuais?

6) Como o princípio da boa-fé interfere na execução de cláusulas contratuais omissas?

5.3. Experiments

With access to a reasonable amount of questions which contain content relevant to our focus and objective, we then proceed with the application of the RAG pipeline to our evaluation dataset.

In this section, we will delve into the parameters and additional techniques employed for our system, and the iterative experiments made to assess the model's performance.

5.3.1. Parameters and Techniques

A technique we applied for the evaluation process pertains to a crucial consideration to keep in mind when developing a solution based on LLMs, which is their non-deterministic (e.g. the output of a model is unpredictable for any given input) nature. Although LLMs contain a parameter called temperature, which controls how non-deterministic the model is when answering prompts, we still decided to implement a technique in order to account for the multiple ways a model can answer a given question. For this, even while our system's temperature was set at 0.1, in total, three answers were generated for each question to account for the non-deterministic nature of LLMs when generating answers. By applying this, our evaluation process will account for the different possible answers a model can give, even if the changes between answers are relatively small.

In terms of parameters of BERTScore, since our text is in Portugal's Portuguese, the parameter *lang* was set to "*pt*". Additionally, since the text was not processed, certain words were not removed, such as prepositions, which by themselves do not hold relevant information (we consider such words stopwords). Given this, it was opted to keep the parameter *idf* set to *True*, which allows the system to take into account the relative importance of words when calculating the score.

5.3.2. Iterations

As already mentioned, the scoring algorithm was inspired by RAGAs [94] and was applied to two groups of candidate–reference pairs: (i) questions and answers, and (ii) answers and context. With this, three different experimental iterations were carried out, namely:

1) Application of base BERTScore without additional processing;
2) Use of BERTScore with token-based segmentation for answers and context, along with a comparison against mismatched question–answer and answer-context pairs;
3) Application of Legal-BERTimbau instead of BERT-multilingual in BERTScore, followed by a comparison of both models.

For the first one, BERTScore was applied directly without any type of processing, meaning the token length was not taken into consideration when applying the metric and the contexts were joined together as a single unit of text. For the score between questions and answers, it was assigned to the question the average F1 Score between each result, so between each question and its respective answers.

The second iteration took into consideration the token length between questions, answers and contexts and processing of the text was applied. In Figure 5.5 we can analyze the difference between token lengths for the questions, answers and contexts and note how, on average, answers contain more tokens than questions and contexts. While there is not a great difference between answers and contexts in terms of token length, we can see how questions have considerably fewer tokens than answers and contexts.

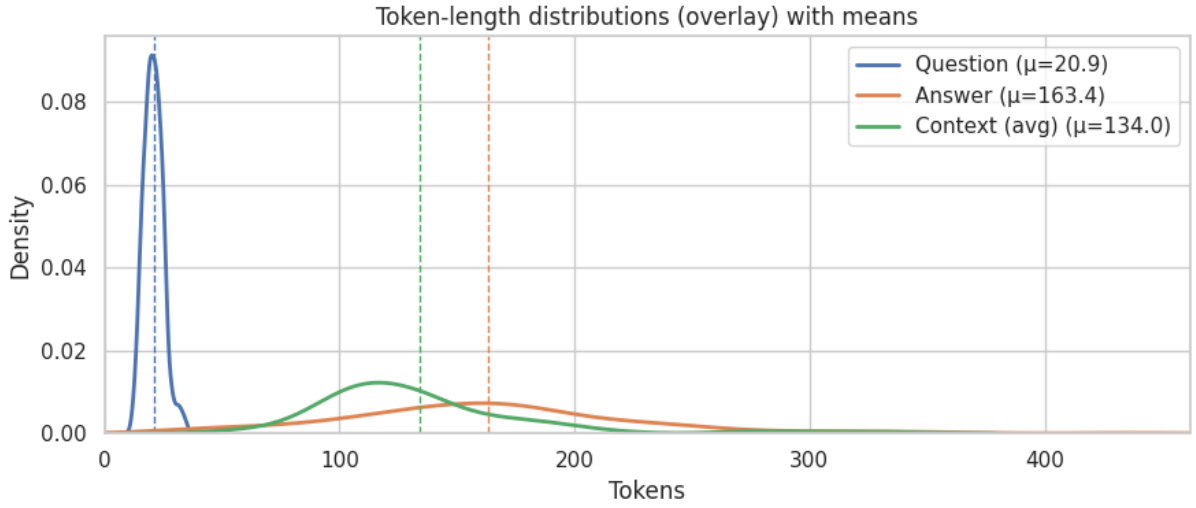

FIGURE 5.5. Token Length Comparison between Questions, Answers and Contexts

As part of this iteration, we tested how the scoring algorithm would behave with mismatched pairs of questions and answers, meaning we built a function which selected a random question and a random answer that is not paired with their respective question. This allowed us to understand how well the scoring method fits our system and its respective limitations.

The third iteration centered in using a distinct model for the scoring instead of the base *bert-base-multilingual-cased* model, in this case, Legal-BERTimbau. This iteration focused on analyzing the differences between the models and the impact a different model has on our results. Similar to the previous iterations, the new score was applied to the question-answers pairs, answer-contexts pairs and mismatched question-answer pairs.

With this information, it was decided to, when calculating the BERTScore between questions and answers, each answer was segmented into sets with token length similar to the question, where we added a overlap of three more tokens in between segments. The BERTScore was then calculated for each segment, with the mean or maximum value being the one assigned for the pair. Since we possess three answers for each question, again, the average of the BERTScore of each pair was calculated and assigned to that question.

As for the scores between answers and contexts, in this iteration, instead of joining the contexts retrieved into a single text, the BERTScore was calculated for each context retrieved with the answers, and the average was assigned to that answer. Segmentation was not applied in this situation due to the token length between these pairs being similar.

5.4. Discussion

In this section, we present the results obtained across the experiments and draw the corresponding conclusions. The subsequent sections are divided between the distinct iterations of development, where it will BE presented not only the respective visualizations which represent and compare our results, but also the respective interpretations of said results.

For context, the data used across experiments was composed of 100 questions, 300 answers and 300 sets of context with an average of five chunks of text in each set of context, making it 300 pairs in total. The reason behind having 300 answers, as mentioned in Section 5.3.2, is that for each question, three answers were generated and consequently, three sets of context were retrieved for each answer. For the first iteration, we used the complete dataset with every question, answer and context. In the second iteration, when creating a mismatched set, we used 100 questions, answers and sets of context. Finally, for the third iteration, we applied the data used in previous iterations.

5.4.1. BERTScore without Processing

The first iteration served as a baseline for evaluating the system's performance and for understanding how subsequent experiments affected it. Figure 5.6 shows the distributions of BERTScore values achieved by the system for each set.

The average scores are relatively high, with 0.722 for the Question-Answer set and 0.662 for the Answer-Context set. These values suggest that the system is capable of achieving semantic alignment across different text pairs. However, it remains necessary to determine whether, for a given question, an unrelated answer yields sufficiently distinct results to allow a clear assessment of performance, and the same for a given answer and an unrelated context.

Another aspect to take into consideration is segmentation in both sets. As illustrated in Figure 5.5, questions and answers differ substantially in their average token length, which may impact our system's performance since the difference in token lengths between candidates and reference pairs impacts BERTScore scoring. In contrast, answers and context are more balanced in this regard. In this first iteration, no segmentation was

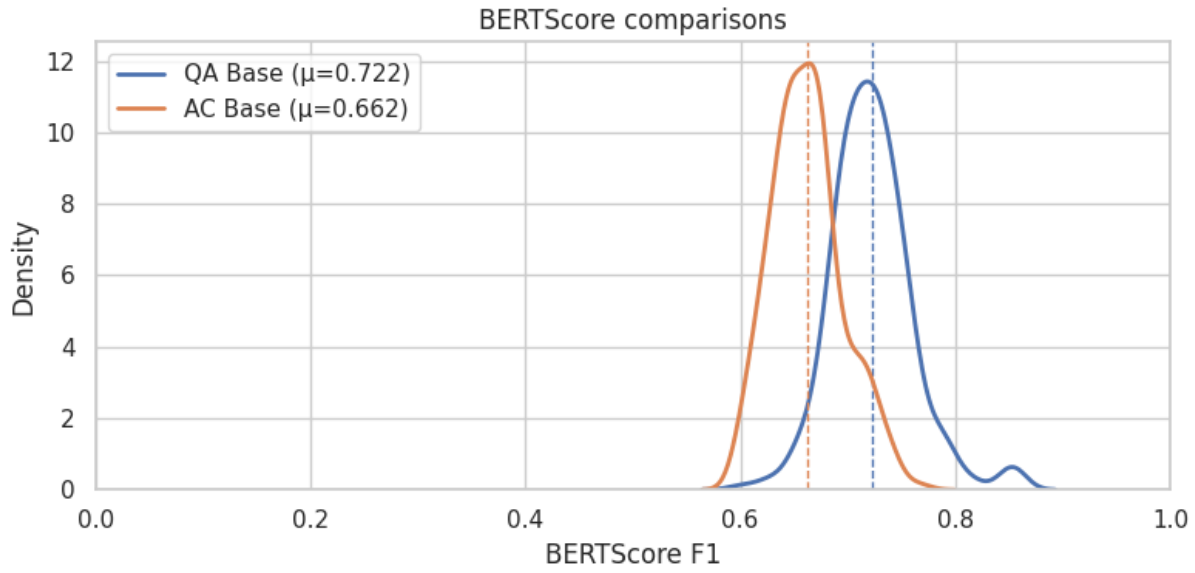

FIGURE 5.6. Distribution of BERTScore values between Question-Answer (QA) and Answer-Context (AC) sets

applied, meaning the context was represented by concatenating the top-5 documents retrieved for each answer, leading to a considerable increase in token length for the context, which might have negatively influenced our results.

5.4.2. BERTScore with Mismatched sets and Segmentation

Building on the baseline analysis, the second iteration aimed to determine whether the model could effectively discriminate between matched and mismatched sets of questions and answers, and whether segmentation had a measurable impact on results. Figures 5.7 and 5.8 illustrate the effect of segmentation on QA and AC, respectively. Figure 5.9 shows the distribution comparison between matched and mismatched Question–Answer (QA) sets, while Figure 5.10 presents the same comparison for the Answer–Context (AC) sets.

From Figures 5.7, 5.8, 5.9, and 5.10, several insights can be drawn. Firstly, considering the effects of segmentation, Figures 5.7 and 5.8 show that segmentation improved the average score by 0.054 for the AC set, but decreased the average score by 0.031 for the QA set. This indicates segmentation is beneficial for the AC set but not for the QA set. It is important to note that the scores from the segmentation of answers for the QA set were averaged across segments rather than taking the maximum. This adjustment was made because, in some cases, the maximum produced abnormally high values, even reaching 1.0, which indicated potential overfitting.

Turning to mismatched sets, from Figures 5.9 and 5.10, we can observe how the distributions for both sets are relatively close, especially in the AC set. This indicates the model is not effectively discriminating between related and unrelated pairs. Such discriminative capability is crucial since if the model cannot distinguish between matched

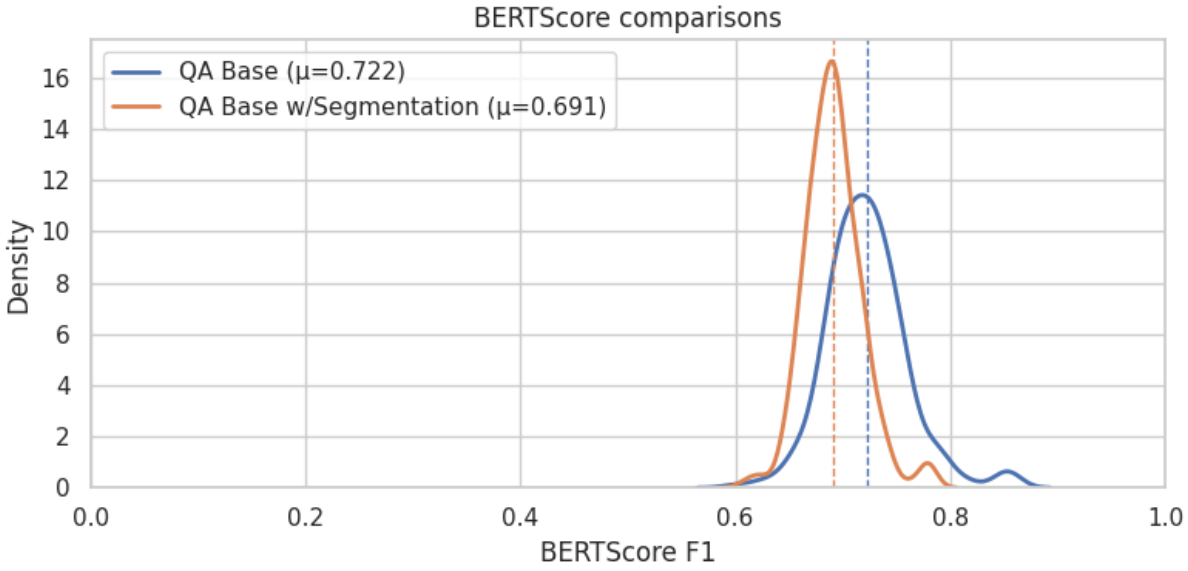

FIGURE 5.7. BERTScore Distribution Comparison between with and without Segmentation for QA set

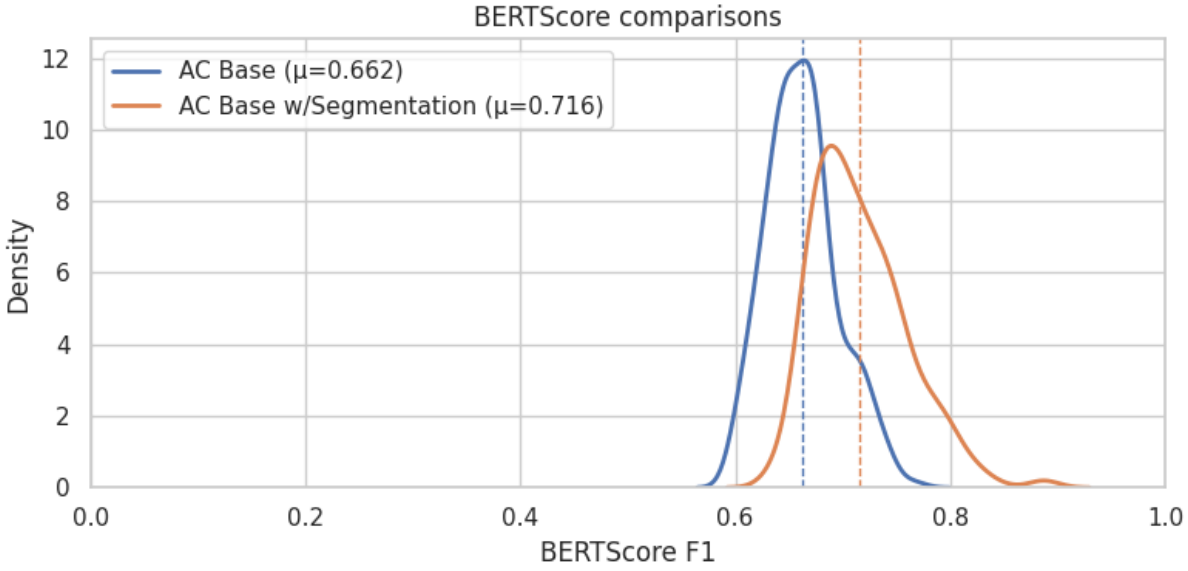

FIGURE 5.8. BERTScore Distribution Comparison between with and without Segmentation for AC set

and mismatched sets, then even when semantically similar answers are produced, it cannot be relied upon to generate the type of legally grounded answers that experts would pose.

Given this analysis, one possible explanation for these results lies in the underlying model used by BERTScore. With the parameter *lang="pt"*, BERTScore defaults to the application of the *bert-base-multilingual-cased* model, which is not fine-tuned for Portuguese legal text. Consequently, the embeddings may not capture sufficient semantic distinctions in this domain since there is a great number of legal terms used across questions, answers and context. This motivated the third iteration, in which we decided to

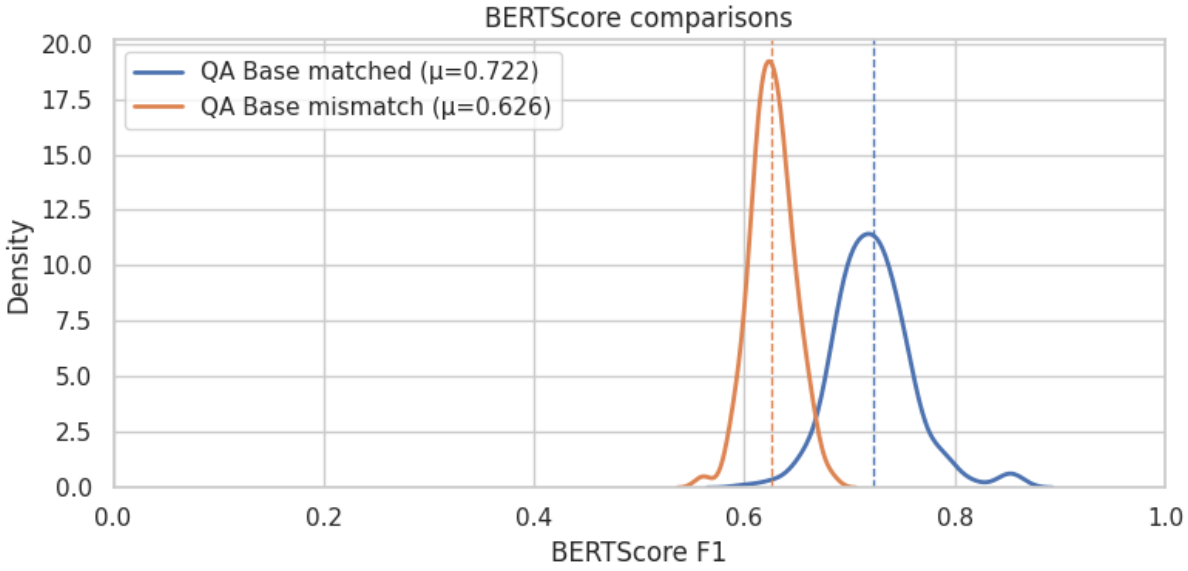

FIGURE 5.9. BERTScore Distribution Comparison between Matched and Mismatched QA sets

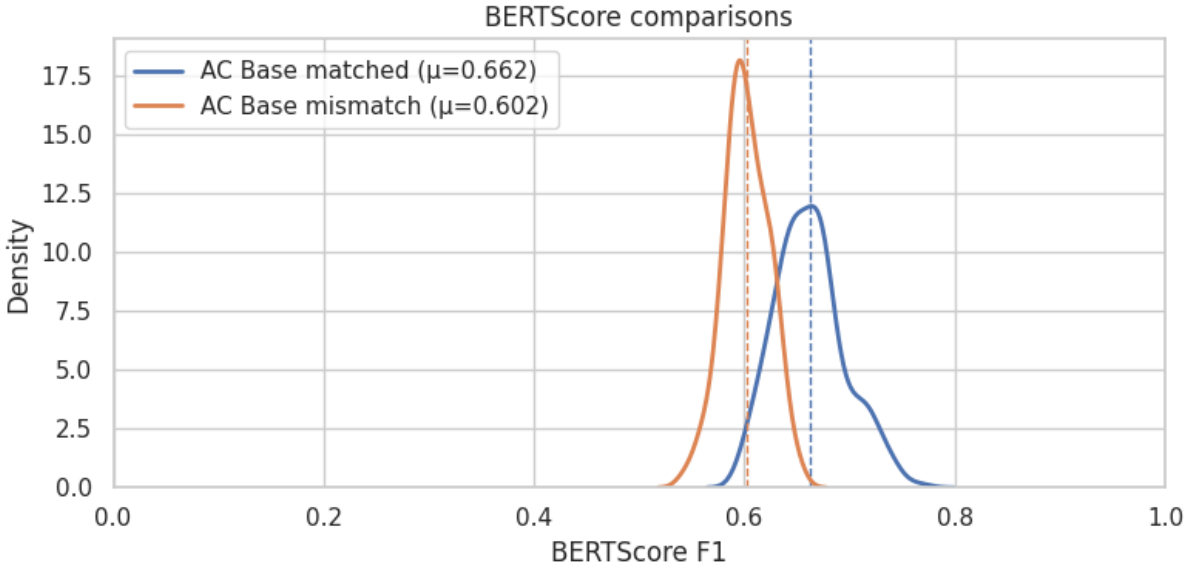

FIGURE 5.10. BERTScore Distribution Comparison between Matched and Mismatched AC sets

apply Legal-BERTimbau instead, with the expectation that it would discriminate more effectively between the distributions.

5.4.3. BERTScore with LegalBERTimbau

Based on the previous analysis, the third and final iteration applied Legal-BERTimbau instead of `bert-base-multilingual-cased` to assess whether the model could better discriminate between matched and mismatched sets, and whether overall performance would improve. We titled this new score, BERTimbauScore, since it uses Legal-BERTimbau instead of the base BERT model. Through Figures 5.11 and 5.12, we can analyze the impact

of Legal-BERTimbau made between matched and mismatched sets for the QA and AC sets, respectively, while Figures 5.13 and 5.14 compare Legal-BERTimbau's performance with the base model between sets.

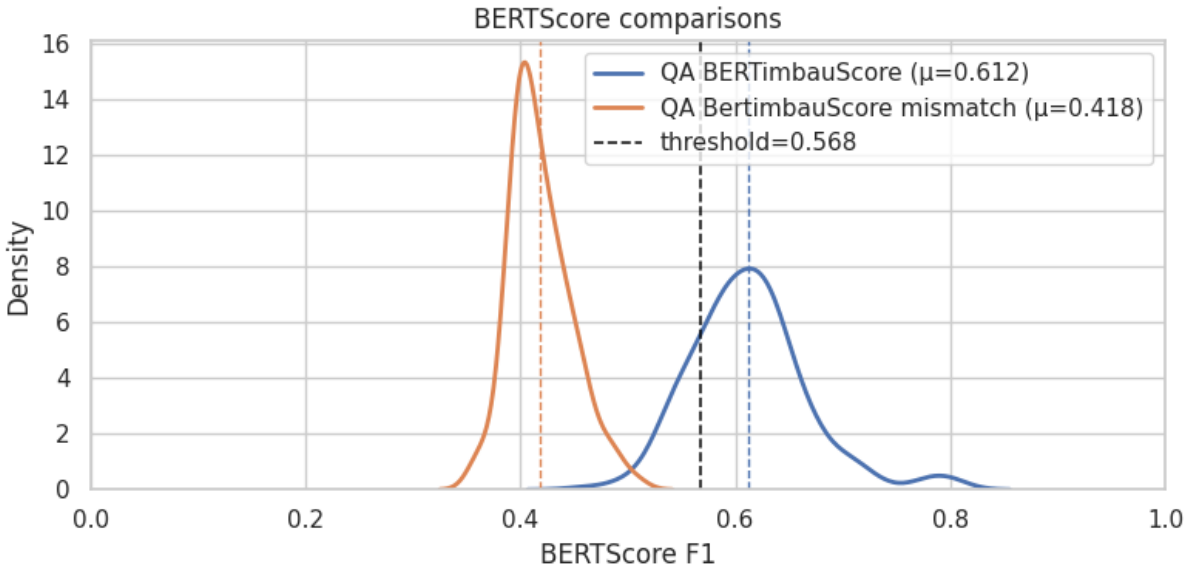

FIGURE 5.11. BERTScore Distribution Comparison between Matched and Mismatched QA sets for Legal-BERTimbau

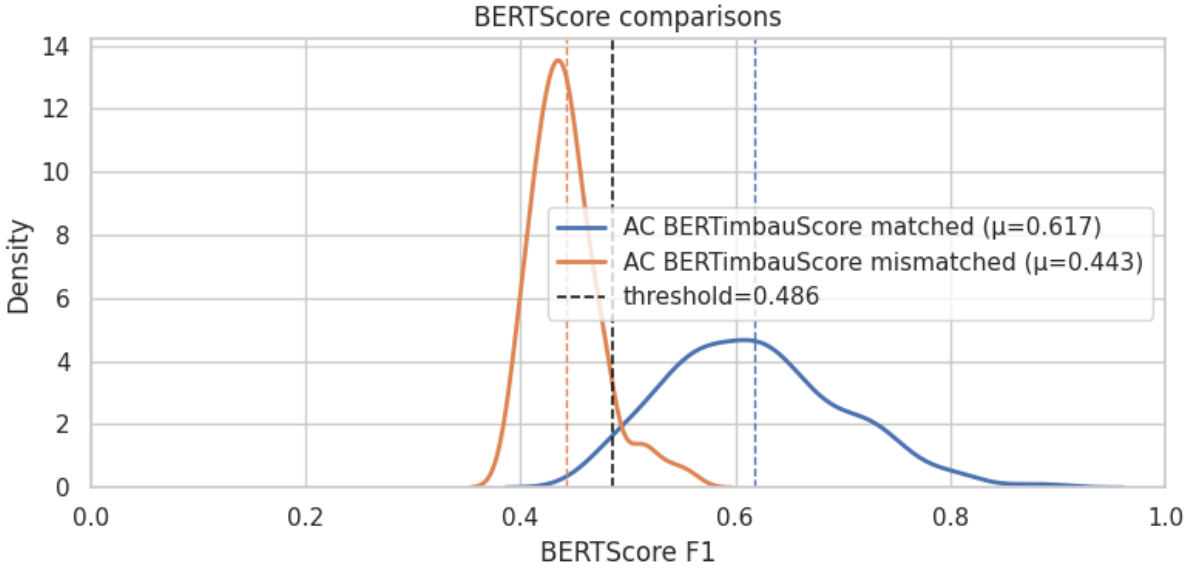

FIGURE 5.12. BERTScore Distribution Comparison between Matched and Mismatched AC sets for Legal-BERTimbau

Figures 5.11 and 5.12 show evidently that the new model discriminates, on average, more effectively between matched and mismatched sets, particularly for the QA set. However, for the AC set, even though the distributions are more distant than with the base model, the overlap between them remains relatively high.

Considering this, we analyzed the system's answers that explicitly indicated insufficient context such as "*...não é mencionado diretamente nos excertos fornecidos, pelo que não é possível...*" or "*Não é possível fornecer...com base nos acórdãos fornecidos, pois...*". We identified four such cases, with mean BERTScores of 0.486 for AC and 0.568 for QA. Although this sample's size is small, these values provide a useful reference point for interpreting the lower end of the distributions. Figures 5.11 and 5.12 show the reference lines marking the aforementioned means, alongside the distributions of scores.

Based on Figures 5.11 and 5.12 with the reference lines we can see how lower-scoring answers tend to cluster closer to the mismatched distribution, indicating Legal-BERTimbau is better at separating relevant from irrelevant cases. In Figure 5.11, the mismatched QA distribution lies mostly below the reference line, while in Figure 5.12 the AC mismatched distribution shows partial overlap above the line, though most values remain below it.

Figures 5.13 and 5.14 provide a direct comparison between Legal-BERTimbau and the base model. By analyzing these figures, we can observe that while the base model achieves higher absolute scores in both QA and AC compared to Legal-BERTimbau, this does not necessarily reflect better performance, since, as shown in the second iteration, it fails to discriminate effectively between matched and mismatched pairs.

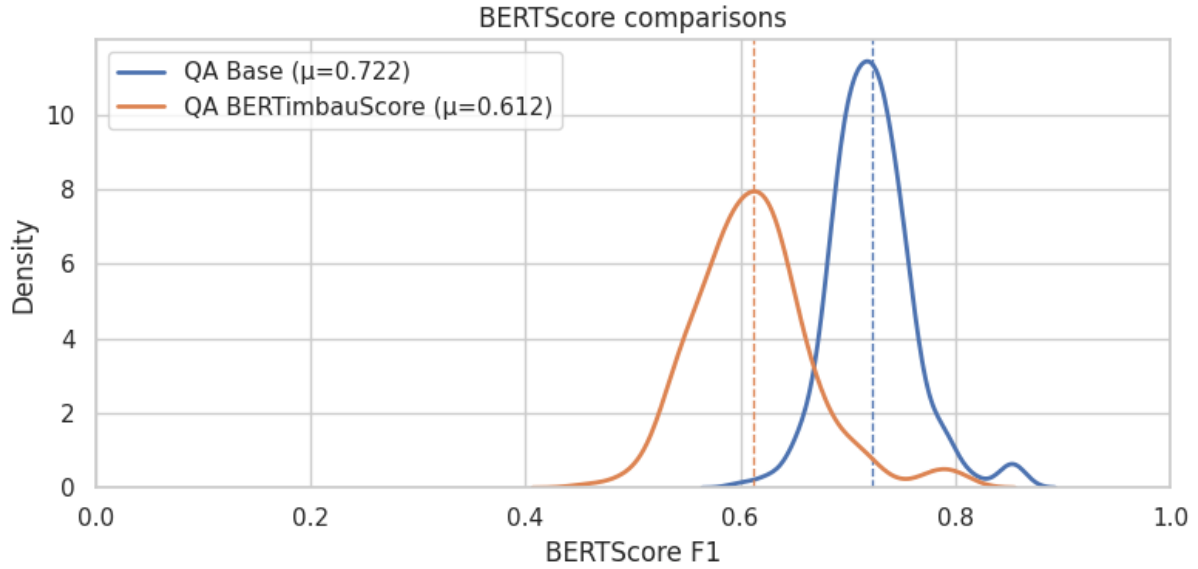

FIGURE 5.13. BERTScore Distribution Comparison between BERT-base and Legal-BERTimbau for QA set

In summary, despite producing slightly lower overall scores, Legal-BERTimbau demonstrated superior discriminative ability, especially for the QA set. While some overlap remains in the AC case, the model still distinguishes matched from mismatched pairs more reliably than the base model. One possible reason for this phenomenon comes from how legal text is complex, where arguments stated might have distinct formulations for the same meaning. Consequently, Legal-BERTimbau, a model fine-tuned on this data, is able

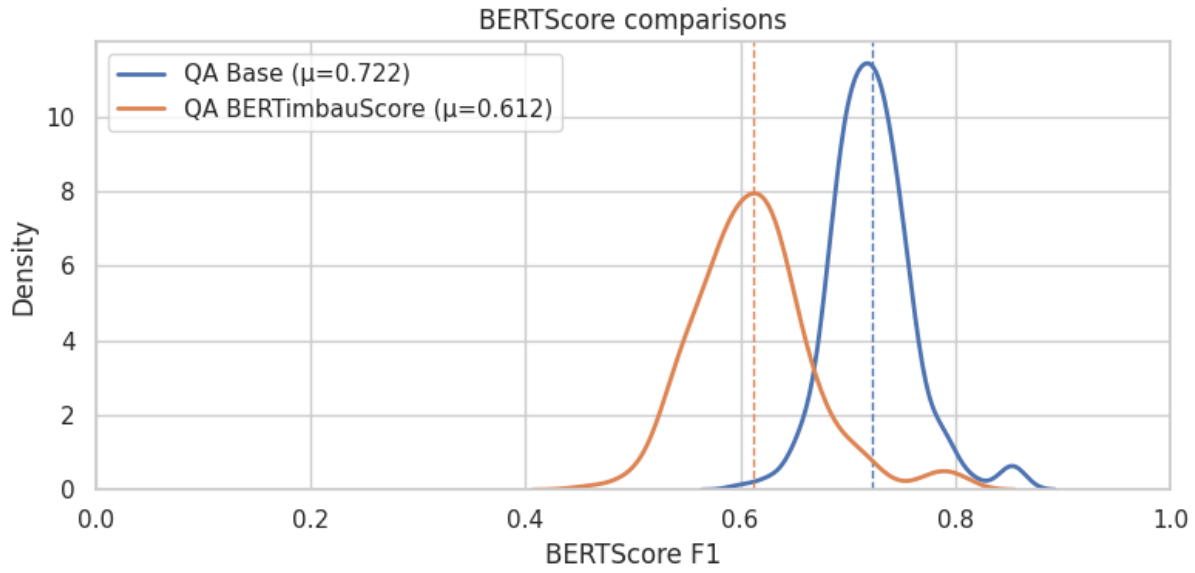

FIGURE 5.14. BERTScore Distribution Comparison between BERT-base and Legal-BERTimbau for AC set

to more effectively embed information given such complexity, making it the best option for this process of evaluation within the tests made.

While this evaluation process does not allow us to assess the quality of the answers in terms of accuracy of information, it still allows us to differentiate between possible unreliable and reliable answers by analyzing how semantically similar the answers are to their questions and context. As we have seen throughout this chapter, Ask Supreme with the use of Legal-BERTimbau possesses capabilities of distinguishing between semantically similar and distinct pairs of questions and answers, and answers and context. This demonstrates the potential that RAG systems have in the legal domain for Portuguese legal documents and that, with support from experts, we could build a system that would effectively aid them in legal search.

CHAPTER 6

Conclusion

Having developed, implemented, and evaluated our proposed system, this chapter presents the final conclusions of this dissertation. This chapter aims to revisit the main objectives defined at the beginning of this work and to reflect on how they were addressed throughout the different stages of development and evaluation.

In addition to summarizing the main findings and contributions of this research, this chapter also discusses its limitations, which help to frame the scope of the results obtained and highlight the aspects that could not be fully explored. Finally, directions for future work are outlined, pointing to possible extensions and improvements that may further enhance the system or broaden its applicability.

6.1. Summary of Contributions

Throughout this dissertation, the central objective was to investigate how NLP-based technologies such as LLMs could aid in the legal search process by using this technology to extract information and provide an analysis to legal experts. From this objective, we composed two research questions related to how we can extract relevant information from legal documents, how LLMs can aid in this extraction and respective analysis, and how we can evaluate these systems without annotated datasets. Guided by our objectives and research questions, the work resulted in a set of contributions that are summarized in this section.

In order to answer our research questions and fulfill our objectives, multiple methodological strategies were used and combined. The Design Science Research methodology provided the overarching structure, guiding the research through iterative cycles of design, development, and evaluation. Within this framework, an adapted version of CRISP-DM was employed during the Design and Development step, enabling an incremental and modular construction of the system while ensuring that each stage remained closely tied to the data, our expertise, and our objectives. At the technical level, the system was implemented as a traditional RAG pipeline, allowing the integration of legal document retrieval with LLM-based answer generation. This combination of methodologies made it possible not only to structure the research process effectively but also to gradually refine the system in a way that reflects both technical and domain-specific requirements.

Besides the development of our tool, this dissertation also possesses other contributions that extend beyond the development of Ask Supreme. The tool itself was built as a system capable of answering legal questions related to a specific topic, namely the *Supremo Tribunal de Justiça* judgments, demonstrating the applicability of RAG to Portuguese

legal texts while also aiding legal experts in legal search. Second, we conducted an updated and systematic literature review on LLM-based approaches in the legal domain, which helps situate this work within the current state of research. Third, we introduced a novel approach to question classification and filtering, which leverages Bloom's Taxonomy as a means of analyzing the cognitive level of questions. This method not only can be used to evaluate similar systems without the use of annotated dataset, it also informs how data can be curated and filtered, and also provides insights into how Question-Answering systems might be specialized for different categories of questions instead of being developed as a generalist system. Together, these contributions provide both a practical artifact and theoretical insights that can support future work at the intersection of law and artificial intelligence.

6.2. Conclusions

In Chapter 5, we presented a series of visualizations of the system's performance alongside their interpretation. With this analysis we could verify the degree of semantic alignment achieved across the different text sets. Through our iterative process, we could identify how the use of a fine-tuned model substantially improved the system's ability to discriminate between matched and mismatched information, even though the overall distribution of scores remains modest.

With regard to this dissertation's objectives, just as proposed, we successfully built a tool capable of extracting and analyzing information while generating insights based on user input. By applying an RAG methodology, the tool was able to, based on a given question, retrieve the related context and extract the relevant information from the retrieved documents. This enables the system to generate answers grounded in contextual evidence, thereby reducing hallucinations and subsequent unreliability.

Regarding the first research question, "How can LLMs aid in the extraction and analysis of legal information from Portuguese legal documents", even though the development of the prototype and the systematic literature review highlighted the capabilities of LLMs and their respective limitations, the system itself and the evaluation methodology employed, do not allow us to directly evaluate the accuracy and relevance of the answers generated. Firstly, based on our results, although LLMs show potential in assisting legal experts in legal search through the use of the RAG framework, the limitations of these models must be acknowledged, along with the ethical issues raised by their use and development, which remain critical considerations across the domain. Secondly, even though the evaluation methodology based on BERTScore and RAGAS does not allow us to assess the accuracy or relevance of the generated answers, this methodology serves as an indicator of the similarity between question, answer and context, providing a way to assess the presence of hallucinations, specially in situations where we are not able to apply a qualitative evaluation methodology. These observations are consistent with the findings presented in the systematic literature review, where domain specific fine-tuning

improves performance, but challenges remain in ensuring reliability, relevance of generated information, and mitigating the ethical concerns these models bring.

Based on our results, these models possess the capacity to assist legal experts in legal search by retrieving relevant documentation and providing contextualized analysis. Although LLMs show such capabilities, their limitations must be acknowledged, along with the ethical issues raised by their use and development, which remain critical considerations across the domain. These observations are consistent with the findings presented in the systematic literature review, where domain specific fine-tuning improves performance, but challenges remain in ensuring reliability and mitigating ethical concerns.

LLMs show such capabilities, their limitations must be acknowledged, along with the ethical issues raised by their use and development, which remain critical considerations across the domain. These observations are consistent with the findings presented in the systematic literature review, where domain specific fine-tuning improves performance, but challenges remain in ensuring reliability and mitigating ethical concerns. In relation to our results, even though the approaches employed do not assess the accuracy and quality

Considering the methodologies employed and results achieved, we can see how for this research questions we were not able to directly identify relevant information for experts without a qualitative evaluation methodology. However, through the BERTScore approach, we have an indicator

Turning to the second research question, "How can we validate and evaluate a question-answering system for the legal domain without a dataset of questions and answers annotated by experts?", through the development of the question generation method based on Bloom's Taxonomy alongside the usage of BERTScore we were able to circumvent one of the challenges of developing a legal NLP system, this being the evaluation of the systems without either a annotated dataset or with the help of experts. With Bloom's Taxonomy, we were capable of generating specific types of questions related to our legal topic of choice while also controlling the level of detail of the respective questions. For the usage of BERTScore with a RAGAS-based approach, as mentioned in Chapter 5, this metric allows us to analyze the semantic similarity between our pairs of questions and answers, and answers and context in order to determine if our system is capable of producing answers related to our questions and context, indicating the usability of RAG systems for the legal domain. Even though we were able to address the challenge of evaluating our system without annotated datasets or the aid of experts, such would greatly benefit our system and our respective results and allow us to evaluate the accuracy, applicability and reliability of our system as an aid to the process of legal search.

The results achieved and presented in Chapter 5, demonstrate how the use of a fine-tuned model greatly improves the discrimination capabilities between matched and mismatched sets. Nonetheless, while promising, these results reveal considerable scope for improvement, particularly in data processing, modeling, and validation.

In summary, the objectives set at the outset of this dissertation were achieved, and the research questions were adequately addressed. Despite the constraints and limitations encountered during this process, with the system's prototype, we were able to analyze the potential of LLMs and RAG-based systems to support legal experts in legal search for Portuguese legal documents. At the same time, some limitations were also noted, and experts from both the technological and legal domains must remain attentive to issues of accuracy, reliability, and ethics. In addition to this, the integration of Bloom's Taxonomy and the development of a RAGAS-based approach with BERTScore for evaluation into our approach represents a valuable and novel contribution, as it provides not only a way to validate and evaluate similar systems without an annotated dataset, but also provides a means of classifying questions and opens opportunities for focusing Question-Answering-based systems. This method still offers a promising avenue for future search and development, which could help build more specialized systems and guide analysis of the information being fed to LLMs.

6.3. Limitations

As discussed throughout this dissertation, a number of limitations and constraints were encountered during the development of this study, which impacted the final artifact and respective results achieved. These limitations primarily concern two core aspects of the system, its degree of specialization and the process adopted for validation.

6.3.1. Data

Regarding our data, as presented in Chapter 4, from the repository used to extract our data, we only had access to the summaries of the available judgments. By only having access to the summaries instead of the full-text, we end up possibly losing portions of information that could benefit our systems, even if it would translate in more time being spent in processing the text.

6.3.2. Specialization

With respect to specialization, the system favors answering questions specifically related to the *Supremo Tribunal de Justiça* and only in Portuguese. Moreover, most of the supported questions fall into the "Remember" and "Understand" categories of Bloom's Taxonomy. These restrictions reflect the choice of building a specialized proof-of-concept prototype, rather than a general-purpose solution, which would have introduced considerable complexity. Although this narrow scope limits the generalization of the results, it allowed for greater control over the development and validation process while still enabling the system capability to meet the objectives defined for this dissertation. In this regard, the prototype serves as an initial demonstration of the applicability of RAG systems in the Portuguese legal domain.

6.3.3. Validation

In our validation phase, we relied on the use of LLM generated questions and the BERTScore metric. While these choices facilitated experimentation, they also introduced limitations. First, the exclusive use of generated questions meant that accuracy may have been compromised, as these questions do not necessarily mirror the phrasing or complexity of expert-level questions. Second, the reliance on quantitative metrics restricted the analysis to semantic alignment. Although BERTScore provides an indicator of possible hallucinations, it does not capture the full accuracy, completeness, or usefulness of the generated answers by itself. Third, the validation dataset itself was limited in size, containing only 100 initial question–answer pairs. A larger and more diverse dataset would have improved the reliability of the evaluation and improved the system's retrieval capabilities. These limitations were partly the result of time and computational resource constraints, which also prevented us from conducting a complementary qualitative evaluation with legal experts. Considering this, it is important to note that, as presented in Chapter 3 and in Appendix B, a great number of studies applied a similar approach to ours when generating Question-Answering datasets and using quantitative metrics while recognizing its limitations.

Despite these constraints, the prototype nonetheless provided meaningful insights into the design and operation of RAG systems for legal applications. These limitations, in turn, highlight important directions for improvement, many of which are addressed in the following section on Future Work.

6.4. Future Work

While the system and techniques developed in this study demonstrate the potential of the RAG framework for assisting experts and novel ways to validate and evaluate these systems without annotated datasets, several opportunities still exist for further improvement and expansion of this work. The following sections outline the potential avenues for future research and development, highlighting directions that could enhance the system's performance and reliability, and address some of the limitations identified throughout this dissertation.

6.4.1. Feature Processing

In our Portuguese judgments dataset, we had access to features such as *decisao*, *tematica* and *descritores*. These features as described in Chapter 4, possess high variability of values leading us to ignore these features in our system since some values present in the features overlap where some of these values could represent the same information but they are recognized as different in our system, which can impact our results since the model might not effectively and accurately identify the similarities and patterns present in the data. However, for *decisao* feature, with the aid of legal experts, we could identify some of these overlapping values or even better categorize them into fewer categories, allowing us to have more consistency and more evident patterns which our system can more easily

and effectively identify. As for the other features, we would need to possibly manually filter and categorize the values for each feature, with the possibility of needing a legal expert to better categorize the values.

In addition to this, exploring IE techniques and their applicability and performance in our system was not possible since not only most of our dataset was already structured, it would also require more legal expert knowledge in order to better model our data and extract meaningful insights.

The downside of these processes is that they require the orientation of legal experts and manual work in order to not only identify all possible values each feature can have and categorize them accordingly, but also to support the IE process.

6.4.2. RAG Experimentation

While we applied a traditional RAG method to build the system's prototype, other techniques could have been added to our approach, for the proof-of-concept such techniques were not applied such as Hybrid Retrieval, Reranking (e.g. addition of a reranking layer to our retrieval step allowing this process to filter more relevant information from the documents), and experimenting with other similarity measures. By applying these techniques, we would possibly see an improvement in our system across the metrics and validation methods being employed.

As mentioned in Chapter 4, for the Ask Supreme system, we used GPT-4o and Legal-BERTimbau as the LLM and embedding model of choice, respectively. Another experiment which could be done is the testing of other different models to the ones employed, which most certainly could impact performance and the generation of answers.

6.4.3. Validation Process

As described in Chapter 4, for validation of our system, we used an LLM generated dataset of questions and answers and applied the BERTScore metric between different pairs of text in order to judge the semantic alignment of our system. However, in order to create a reliable and accurate system, instead of generating a dataset and using a quantitative measure to assess performance, we would need the aid of legal experts to better validate our system through a qualitative approach.

This process would require a dataset composed of expert-level questions and answers related to our topic of choice and a team of experts to aid us in validating our system's answers and their respective accuracy and reliability, either through a questionnaire or interviews, where we would personally analyze how the system is being used by experts and document their experience and respective opinions of the system. In addition to this, as mentioned in Section 6.3, we would require a great number of examples from experts in order to have a diverse sample of examples that our system could take advantage of.

As expected, applying this approach would require a considerable amount of time and resources in order to be carried out, in addition to requiring an available team of experts. However, with these resources, this would greatly improve the system overall, not only

leading to a more accurate and reliable system, but also a system designed alongside experts, catering to their needs and requirements.

6.5. Final Remarks

This dissertation demonstrates the feasibility of building a RAG-based system for Portuguese legal text and how it can be used to aid experts in legal search. With our proof-of-concept, we are able to see the potential of LLM-based technology and how it can be implemented in the legal domain, while still being mindful of its limitations and concerns it brings forth.

We recognize the limitations of this study, such as the use of a quantitative-based approach for validation instead of a qualitative-based approach, and the use of LLM generated datasets for such process.

In summary, we believe that this study was able to achieve its objective while also making a significant contribution to the legal and technological field by providing novel approaches to question classification and filtering and providing evidence of the potential and limitations of LLMs and RAG-based systems in legal search.

[This page is intentionally left blank.]

References

[1] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large language models in law: A survey," *AI Open*, vol. 5, pp. 181–196, 2024, ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2024.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651024000172.

[2] D. H. Anh, D.-T. Do, V. Tran, and N. L. Minh, "The impact of large language modeling on natural language processing in legal texts: A comprehensive survey," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–7. DOI: 10.1109/KSE59128.2023.10299488.

[3] E. Quevedo, T. Cerny, A. Rodriguez, P. Rivas, J. Yero, K. Sooksatra, A. Zhakubayev, and D. Taibi, "Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications," *IEEE Access*, vol. 12, pp. 145 286–145 317, 2024. DOI: 10.1109/ACCESS.2023.3333946.

[4] N. Cong-Lem, A. Soyoof, and D. Tsering, "A systematic review of the limitations and associated opportunities of chatgpt," *International Journal of Human-Computer Interaction*, 2024, Cited by: 14; All Open Access, Hybrid Gold Open Access. DOI: 10.1080/10447318.2024.2344142. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192538276&doi=10.1080%2f10447318.2024.2344142&partnerID=40&md5=a5e55f7e3003866de8b9da5efec461ea.

[5] D. Kurniawan and S. E. Hiererra, "Ai legal companion: Enhancing access to justice and legal literacy for the public," in *2024 International Conference on ICT for Smart Society (ICISS)*, 2024, pp. 1–6. DOI: 10.1109/ICISS62896.2024.10751371.

[6] D. Zhang, D. Trautmann, A. Petrova, and F. Schilder, "Unleashing the power of large language models for legal applications," Cited by: 6, 2023, pp. 5257–5258. DOI: 10.1145/3583780.3615993. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178115238&doi=10.1145%2f3583780.3615993&partnerID=40&md5=3765deb1fca2b68f505e75aae3704a23.

[7] S. Jacob, P. M. Jacob, J. Cheriyan, and S. Ismail, "Legal assistance redefined: Transforming legal access with ai-powered legallink," in *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, 2024, pp. 1–6. DOI: 10.1109/SPICES62143.2024.10779909.

[8] A. Alif Adillah, I. Afrizal, A. Yusrotis Zakiyyah, and Meiliana, "Retrieval-augmented generation for indonesia non-convention vessel standards exploration," in *2024 International Conference on Information Technology and Computing (ICITCOM)*, 2024, pp. 178–182. DOI: 10.1109/ICITCOM62788.2024.10762180.

[9] M. Visciarelli, G. Guidi, L. Morselli, D. Brandoni, G. Fiameni, L. Monti, S. Bianchini, and C. Tommasi, "Savia: Artificial intelligence in support of the lawmaking process," Cited by: 0, vol. 3762, 2024, pp. 436–440. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205585505&partnerID=40&md5=a08b95d679e777674556004bf4c7052c.

[10] M. Cherubini, F. Romano, A. Bolioli, L. De Mattei, and M. Sangermano, "Improving the accessibility of eu laws: The chat-eur-lex project," Cited by: 0, vol. 3762, 2024, pp. 6–11. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205594793&partnerID=40&md5=fcec37c934f20ae05bd38acdbde4cbfe.

[11] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, pp. 45–77, Jan. 2007.

[12] C. M. Greco and A. Tagarelli, "Bringing order into the realm of transformer-based language models for artificial intelligence and law," *Artificial Intelligence and Law*, vol. 32, no. 4, pp. 863–1010, Dec. 2024, ISSN: 1572-8382. DOI: 10.1007/s10506-023-09374-7. [Online]. Available: https://doi.org/10.1007/s10506-023-09374-7.

[13] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring llms applications in law: A literature review on current legal nlp approaches," *IEEE Access*, vol. 13, pp. 18 253–18 276, 2025. DOI: 10.1109/ACCESS.2025.3533217.

[14] D. Z. Korman, E. Mack, J. Jett, and A. H. Renear, "Defining textual entailment," *Journal of the Association for Information Science and Technology*, vol. 69, pp. 763–772, 2018.

[15] Zeyuanhu, *Recognizing Textual Entailment Using Deep Learning Models in NLP*, [Online; accessed 15-April-2025], 2020. [Online]. Available: https://medium.com/@zeyuanhu/recognizing-contextual-entailment-using-nneural-network-in-nlp-ea9c5f1a216a.

[16] M. S. Jacob Murel, *What is information retrieval?* [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.ibm.com/think/topics/information-retrieval.

[17] T. Mucci, *What is question answering?* [Online; accessed 15-April-2025], 2025. [Online]. Available: https://www.ibm.com/think/topics/question-answering.

[18] C. S. Ivan Belcic, *What is information extraction?* [Online; accessed 15-April-2025], 2025. [Online]. Available: https://www.ibm.com/think/topics/information-extraction.

[19] *Cuad — contract understanding atticus dataset*, [Online; accessed 15-April-2025], 2025. [Online]. Available: https://www.atticusprojectai.org/cuad.

[20] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma, "Lecard: A legal case retrieval dataset for chinese law system," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21, Virtual Event, Canada: Association for Computing Machinery, 2021,

pp. 2342–2348, ISBN: 9781450380379. DOI: 10.1145/3404835.3463250. [Online]. Available: https://doi.org/10.1145/3404835.3463250.

[21] A. Louis, G. van Dijck, and G. Spanakis, "Interpretable long-form legal question answering with retrieval-augmented large language models," *CoRR*, vol. abs/2309.17050, 2023. arXiv: 2309.17050. [Online]. Available: https://arxiv.org/abs/2309.17050.

[22] *Coliee — competition on legal information extraction/entailment*, [Online; accessed 15-April-2025], 2025. [Online]. Available: https://coliee.org/overview.

[23] N. Hassan, *Embedding Models Explained: A Guide to NLP's Core Technology*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://medium.com/@nay1228/embedding-models-a-comprehensive-guide-for-beginners-to-experts-0cfc11d449f1.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1301.3781.

[26] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162/.

[27] *Huggingface*, [Online; accessed 24-September-2025]. [Online]. Available: https://huggingface.co/.

[28] J. Ferrer, *How Transformers Work: A Detailed Exploration of Transformer Architecture*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.datacamp.com/tutorial/how-transformers-work?utm_adgroupid=165153430042&utm_device=c&utm_matchtype=&utm_network=g&utm_adpostion=&utm_targetid=dsa-2218886984380&utm_loc_interest_ms=&utm_loc_physical_ms=1011729&dc_referrer=https%3A%2F%2Fwww.google.com%2F.

[29] *Transformer architectures*, [Online; accessed 12-May-2025], 2025. [Online]. Available: https://huggingface.co/learn/llm-course/chapter1/6.

[30] R. Singh, *Types of Transformer Model*, [Online; accessed 12-May-2025], 2024. [Online]. Available: https://medium.com/@RobuRishabh/types-of-transformer-model-1b52381fa719.

[31] S. Banerjee, A. Agarwal, and S. Singla, *Llms will always hallucinate, and we need to live with this*, 2024. arXiv: 2409.05746 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2409.05746.

[32] *Prompting tasks — hugging face transformers documentation*, [Online; accessed 15-April-2025], 2025. [Online]. Available: https://huggingface.co/docs/transformers/main/en/tasks/prompting.

[33] S. Ahmed, *Few-Shot and Zero-Shot Learning: Teaching AI with Minimal Data*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://medium.com/@sahin.samia/few-shot-and-zero-shot-learning-teaching-ai-with-minimal-data-801603ed40f8.

[34] *Few-shot prompting*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.promptingguide.ai/techniques/fewshot.

[35] *Zero-shot prompting*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.promptingguide.ai/techniques/zeroshot.

[36] *Chain-of-thought prompting*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.promptingguide.ai/techniques/cot.

[37] S. A. Reji, R. Sheik, S. A, A. R. M, and S. Jaya Nirmala, "Enhancing llm performance on legal textual entailment with few-shot cot-based rag," in *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, 2024, pp. 1–6. DOI: 10.1109/SPICES62143.2024.10779705.

[38] *Irac method*, [Online; accessed 26-September-2025]. [Online]. Available: https://www.iracmethod.com/irac-methodology.

[39] A. Jain, *A Comprehensive Guide to Performance Metrics in Machine Learning*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://medium.com/@abhishekjainindore24/a-comprehensive-guide-to-performance-metrics-in-machine-learning-4ae5bd8208ce.

[40] C. S. Rina Diane Caballar, *What are LLM benchmarks?* [Online; accessed 15-April-2025], 2024. [Online]. Available: https://www.ibm.com/think/topics/llm-benchmarks.

[41] N. Muennighoff, *MTEB: Massive Text Embedding Benchmark*, [Online; accessed 15-April-2025], 2022. [Online]. Available: https://huggingface.co/blog/mteb.

[42] N. Avsar, *Model Evaluation Metrics in Machine Learning — Classification and Regression Analysis*, [Online; accessed 15-April-2025], 2023. [Online]. Available: https://medium.com/@neslihannavsar/model-evaluation-metrics-in-machine-learning-classification-and-regression-analysis-aedf99d4fa8a.

[43] T. V. Quentin Herreros and T. Papaoikonomou, *RAG evaluation metrics: A journey through metrics*, [Online; accessed 15-April-2025], 2023. [Online]. Available: https://www.elastic.co/search-labs/blog/evaluating-rag-metrics.

[44] *Understanding bleu and rouge score for nlp evaluation*, [Online; accessed 08-December-2025]. [Online]. Available: https://www.geeksforgeeks.org/nlp/understanding-bleu-and-rouge-score-for-nlp-evaluation/.

[45] *Meteor*, [Online; accessed 08-December-2025]. [Online]. Available: https://docs.kolena.com/metrics/meteor/.

[46] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[47] S.-H. Tsang, *Brief review — bertscore: Evaluating text generation with bert*, 2024. [Online]. Available: https://sh-tsang.medium.com/brief-review-bertscore-evaluating-text-generation-with-bert-0bc5fc889d7b.

[48] Y. van Bruchem, *Qualitative evaluation of LLM responses*, [Online; accessed 15-April-2025], 2024. [Online]. Available: https://medium.com/ordina-data/qualitative-evaluation-of-llm-responses-3d6717d73a30.

[49] M. E. Mamalis, E. Kalampokis, F. Fitsilis, G. Theodorakopoulos, and K. Tarabanis, "A large language model agent based legal assistant for governance applications," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14841 LNCS, pp. 286–301, 2024, Cited by: 0. DOI: 10.1007/978-3-031-70274-7_18. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85202607024&doi=10.1007%2f978-3-031-70274-7_18&partnerID=40&md5=3f7afbb49068ef42d2bf229112cd6964.

[50] S. Yao, Q. Ke, Q. Wang, K. Li, and J. Hu, "Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities," in *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, ser. RAIIE '24, Singapore, Singapore: Association for Computing Machinery, 2024, pp. 108–112, ISBN: 9798400718311. DOI: 10.1145/3689299.3689319. [Online]. Available: https://doi.org/10.1145/3689299.3689319.

[51] J. S. Garlyal, B. Hariharan, and A. K. Singh, "An analysis on integrating advanced conversational ai in legal summarization and information retrieval," Cited by: 0, 2024, pp. 43–46. DOI: 10.1109/ICICI62254.2024.00016. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205958338&doi=10.1109%2fICICI62254.2024.00016&partnerID=40&md5=ba3473589d0d452a23657fee1cd714b8

[52] T. Litaina, A. Soularidis, G. Bouchouras, K. Kotis, and E. Kavakli, "Towards llm-based semantic analysis of historical legal documents," Cited by: 0, vol. 3724, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199145227&partnerID=40&md5=acdd1841ec0ce2d0a9fb419fc5ab4abf.

[53] R. Nai, E. Sulis, I. Fatima, and R. Meo, "Large language models and recommendation systems: A proof-of-concept study on public procurements," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14763 LNCS, pp. 280–290, 2024, Cited by: 3. DOI: 10.1007/978-3-031-70242-6_27. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205480566&doi=10.1007%2f978-3-031-70242-6_27&partnerID=40&md5=0b7d5361fa84b56d54af20b077b95c6f.

[54] S. Amri, S. Bani, and R. Bani, "Moroccan legal assistant enhanced by retrieval-augmented generation technology," Cited by: 0, 2024. DOI: 10.1145/3659677.

3659737. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85202570190&doi=10.1145%2f3659677.3659737&partnerID=40&md5=4722bea7ab42451a8a32c1cd2e8d436a.

[55] C. Ryu, S. Lee, S. Pang, C. Choi, H. Choi, M. Min, and J.-Y. Sohn, "Retrieval-based evaluation for llms: A case study in korean legal qa," Cited by: 5, 2023, pp. 132–137. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85185008870&partnerID=40&md5=52e1578316e95acc68d79fcdfb27f825.

[56] G. Zambrano, "Case law as data : Prompt engineering strategies for case outcome extraction with large language models in a zero-shot setting," *Law, Technology and Humans*, vol. 6, no. 3, pp. 80–101, Nov. 2024. DOI: 10.5204/lthj.3623. [Online]. Available: https://lthj.qut.edu.au/article/view/3623.

[57] V. R. D. Gabriela Argüelles Terrón Patricia Martín Chozas, "Event extraction and semantic representation from spanish workers' statute using large language models," pp. 329–334, 2023. DOI: 10.3233/FAIA230983.

[58] H. K. M, J. P, and A. K. M, "Large language models for indian legal text summarisation," in *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2024, pp. 1–5. DOI: 10.1109/CONECCT62155.2024.10677065.

[59] F. O. do Espírito Santo, S. Marques Peres, G. de Sousa Gramacho, A. Alves Franco Brandão, and F. G. Cozman, "Legal document-based, domain-driven qa system: Llms in perspective," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–9. DOI: 10.1109/IJCNN60899.2024.10650895.

[60] M. Kaoutar, B. J. Chaima, B. Omar, and B. Outmane, "Unlocking the potential of large language models in legal discourse: Challenges, solutions, and future directions," in *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2024, pp. 1–7. DOI: 10.1109/ICDS62089.2024.10756345.

[61] D. Chauhan, M. Singh, A. Sharma, H. Narang, S. Vats, and V. Sharma, "Development of a legal chatbot for comprehensive user support," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 2024, pp. 1–4. DOI: 10.1109/APCIT62007.2024.10673669.

[62] S. Li and L. Yi, "A few-shot entity relation extraction method in the legal domain based on large language models," Cited by: 1, 2024, pp. 580–586. DOI: 10.1145/3675417.3675513. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200355659&doi=10.1145%2f3675417.3675513&partnerID=40&md5=aa5939496179c02685b9525339c3be52.

[63] Y. Hu, K. Luo, and Y. Feng, "Ella: Empowering llms for interpretable, accurate and informative legal advice," Cited by: 0, vol. 3, 2024, pp. 374–387. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203817416&partnerID=40&md5=8eb4c898ac979016370c332b10bf124d.

[64] A. Louis, G. van Dijck, and G. Spanakis, "Interpretable long-form legal question answering with retrieval-augmented large language models," 20, Cited by: 10; All Open Access, Gold Open Access, Green Open Access, vol. 38, 2024, pp. 22 266–22 275. DOI: 10.1609/aaai.v38i20.30232. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189622802&doi=10.1609%2faaai.v38i20.30232&partnerID=40&md5=f35e03c11ff2071c6d50c516aac8beee.

[65] Y. Zhou, H. Huang, and Z. Wu, "Boosting legal case retrieval by query content selection with large language models," Cited by: 3; All Open Access, Green Open Access, 2023, pp. 176–184. DOI: 10.1145/3624918.3625328. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85180130645&doi=10.1145%2f3624918.3625328&partnerID=40&md5=ee18d9508f863ff4e58e5a12dc64653e.

[66] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, and B. Fleisch, "Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14775 LNAI, pp. 445–460, 2024, Cited by: 5. DOI: 10.1007/978-3-031-63646-2_29. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85198479753&doi=10.1007%2f978-3-031-63646-2_29&partnerID=40&md5=610cf4d4735b6fe4db74ab37c897d871.

[67] L. Hoang, T. Bui, C. Nguyen, and L.-M. Nguyen, "Aiepu at alqac 2023: Deep learning methods for legal information retrieval and question answering," Cited by: 2, 2023. DOI: 10.1109/KSE59128.2023.10299426. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178501733&doi=10.1109%2fKSE59128.2023.10299426&partnerID=40&md5=6b4c79559cc59db24afebc2c9894afca.

[68] M. M. Zin, H. T. Nguyen, K. Satoh, S. Sugawara, and F. Nishino, "Information extraction from lengthy legal contracts: Leveraging query-based summarization and gpt-3.5," *Frontiers in Artificial Intelligence and Applications*, vol. 379, pp. 177–186, 2023, Cited by: 0; All Open Access, Hybrid Gold Open Access. DOI: 10.3233/FAIA230963. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181168881&doi=10.3233%2fFAIA230963&partnerID=40&md5=ed19868e3c4cd060d3e3e54acf08266c.

[69] D. Bakir, B. Yildiz, and M. S. Aktas, "Developing and evaluating a model-based metric for legal question answering systems," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 2745–2754. DOI: 10.1109/BigData59044.2023.10386689.

[70] R. O. Nunes, A. S. Spritzer, C. M. D. S. Freitas, and D. G. Balreira, "Out of sesame street: A study of portuguese legal named entity recognition through in-context learning," Cited by: 1; All Open Access, Hybrid Gold Open Access, vol. 1, 2024, pp. 477–489. DOI: 10.5220/0012624700003690. [Online]. Available: https://www.

scopus.com/inward/record.uri?eid=2-s2.0-85193953512&doi=10.5220%2f0012624700003690&partnerID=40&md5=bddfabd813511b22e8b8b5cb942517ee.

[71] G. M. Coelho, A. Celecia, J. de Sousa, M. Lemos, M. J. Lima, A. Mangeth, I. Frajhof, and M. Casanova, "Information extraction in the legal domain: Traditional supervised learning vs. chatgpt," Cited by: 1; All Open Access, Hybrid Gold Open Access, vol. 1, 2024, pp. 579–586. DOI: 10.5220/0012499800003690. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193908123&doi=10.5220%2f0012499800003690&partnerID=40&md5=3b9fb0ae8ec25b6a669d9fb070999037.

[72] J. Moreira, A. Da Silva, E. De Moura, and L. Marinho, "A study on unsupervised question and answer generation for legal information retrieval and precedents understanding," Cited by: 0, 2024, pp. 2865–2869. DOI: 10.1145/3626772.3661354. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200599752&doi=10.1145%2f3626772.3661354&partnerID=40&md5=c4efefab3bd24ff52ceb48f4c136310c.

[73] K. Cho, S. Han, Y. R. Choi, and W. Hwang, "Nestle: A no-code tool for statistical analysis of legal corpus," Cited by: 0, 2024, pp. 52–61. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85188747970&partnerID=40&md5=e8aa29933048c82e9dea8918085048fa.

[74] S. Adhikary, P. Sen, D. Roy, and K. Ghosh, "A case study for automated attribute extraction from legal documents using large language models," *Artificial Intelligence and Law*, Nov. 2024, ISSN: 1572-8382. DOI: 10.1007/s10506-024-09425-7. [Online]. Available: https://doi.org/10.1007/s10506-024-09425-7.

[75] A. S, A. Saxena, J. Mahajan, L. Panikulangara, S. Kulkarni, and D. S. Bang, "Legalmind system and the llm-based legal judgment query system," in *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, 2024, pp. 1–5. DOI: 10.1109/TQCEBT59414.2024.10545179.

[76] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.09288.

[77] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, *Finetuned language models are zero-shot learners*, 2022. arXiv: 2109.01652 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2109.01652.

[78] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *Llama: Open and efficient foundation language models*, 2023. arXiv: 2302.13971 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2302.13971.

[79] G. Team, R. Anil, S. Borgeaud, *et al.*, *Gemini: A family of highly capable multimodal models*, 2025. arXiv: 2312.11805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2312.11805.

[80] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[81] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, *Qwen technical report*, 2023. arXiv: 2309.16609 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.16609.

[82] Y. Zhou, X. Yan, H. Huang, H. Yan, and M. Chen, "Legal text retrieval with contrastive representation learning and evolutionary data augmentation," Cited by: 0, 2024. DOI: 10.1109/CEC60901.2024.10612052. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201733917&doi=10.1109%2fCEC60901.2024.10612052&partnerID=40&md5=0eccb8933ae94706a2a745ced2c4ea98.

[83] S. Gao, Y. Li, F. Ge, M. Lin, H. Yu, S. Wang, and Z. Miao, "Match and retrieval: Legal similar case retrieval via graph matching network," Cited by: 1, 2023, pp. 227–234. DOI: 10.1109/ICDMW60847.2023.00035. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186144432&doi=10.1109%2fICDMW60847.2023.00035&partnerID=40&md5=21d05c5abb2666ee3062780a32485116.

[84] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*, 2024. arXiv: 2412.13663 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.13663.

[85] E. Amorim, R. Campos, A. Jorge, P. Mota, and R. Almeida, "Text2story: A python toolkit to extract and visualize story components of narrative text," Cited by: 0,

2024, pp. 15 761–15 772. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195938411&partnerID=40&md5=835365614b9921f32c8e7c8977c650

[86] P. Krasadakis, E. Sakkopoulos, and V. S. Verykios, "A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages," *Electronics (Switzerland)*, vol. 13, no. 3, 2024, Cited by: 6; All Open Access, Gold Open Access. DOI: 10.3390/electronics13030648. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85184675668&doi=10.3390%2felectronics13030648&partnerID=40&md5=460aee45ced79d3f01af2ed8c7cb50aa.

[87] S. C, "The crisp-dm model: The new blueprint for data mining," *Journal of Data Warehousing*, 2000.

[88] *Conselho superior da magistratura*, [Online; accessed 16-September-2025]. [Online]. Available: https://jurisprudencia.csm.org.pt/.

[89] H. A. Hasan, *The best text chunking method?* [Online; accessed 15-September-2025], 2024. [Online]. Available: https://hasanaboulhasan.medium.com/the-best-text-chunking-method-f5faeb243d80.

[90] R. Melo, P. A. Santos, and J. Dias, "A semantic search system for the supremo tribunal de justiça," in *Progress in Artificial Intelligence*, N. Moniz, Z. Vale, J. Cascalho, C. Silva, and R. Sebastião, Eds., Cham: Springer Nature Switzerland, 2023, pp. 142–154.

[91] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024. arXiv: 2401.08281 [cs.LG].

[92] *Pinecone*, [Online; accessed 26-September-2025]. [Online]. Available: https://www.pinecone.io/.

[93] J. Liu, *LlamaIndex*, Nov. 2022. DOI: 10.5281/zenodo.1234. [Online]. Available: https://github.com/jerryjliu/llama_index.

[94] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, *Ragas: Automated evaluation of retrieval augmented generation*, 2025. arXiv: 2309.15217 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.15217.

[95] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212–218, 2002. DOI: 10.1207/s15430421tip4104_2.

[96] *Bloom's taxonomy categories and action verbs*, [Online; accessed 08-December-2025]. [Online]. Available: https://www.umkc.edu/provost/docs/academics/blooms-taxonomy.pdf.

[97] *Bloom's taxonomy revised diagram*, [Online; accessed 08-December-2025]. [Online]. Available: https://cte.utah.edu/instructor-education/pdfs/bloom-handout.pdf.

APPENDIX A

Appendix: Example of Prompt usage

Usa o contexto seguinte para responder à questão posta. O contexto refere-se a subconjuntos de acordãos portugueses de Portugal do Supremo Tribunal de Justiça. As respostas devem ser respondidas em português de Portugal. Caso o contexto não seja suficiente para responder à questão responde que não é possível responder à questão dada com a informação disponível.

Contexto: ['I - I. Não pode adquirir-se a propriedade de parte física de fracção autónoma de prédio constituído em propriedade horizontal antes que haja alteração do título constitutivo que autonomize essa parte física da fracção da outra em que estava inserida. II. O Tribunal não pode alterar o título constitutivo da propriedade horizontal em violação das normas legais em vigor, designadamente, sem a aprovação de todos os condóminos e junção de documento emanado da Câmara Municipal comprovativo que a alteração está de acordo com as leis e regulamentos em vigor na autarquia, porque não pode impor a terceiros nem aos Condóminos uma decisão que a todos atinge, quando os condóminos e o Município não são sequer partes na acção. III. O Tribunal só pode declarar adquiridas por usucapião fracções autónomas completas (a menos que se trate de aquisição em compropriedade), sob pena de fraude à lei. IV. Actua com abuso de direito o construtor vendedor que depois de ter declarado, através de documento particular, doar a um condómino uma garagem e arrecadação e durante mais de 15 anos ter agido como verdadeira doação válida se tratasse, vem invocar a inobservância da forma legal, ao fim desses anos todos, para obter a declaração judicial de nulidade da doação. V. Do abuso de direito podem decorrer vários efeitos jurídicos: pode dar lugar à obrigação de indemnizar, à nulidade nos termos gerais do art. 294.º; à legitimidade da oposição; ao alongamento de um prazo de prescrição ou de caducidade, etc.', 'IV - A fim de afastar a presunção de comunhão das partes comuns que estejam afectadas ao uso exclusivo de um dos condóminos, basta uma afectação material, uma destinação objectiva, mas já existente à data da criação do condomínio, embora não se exija que ela conste do respectivo título constitutivo da propriedade horizontal, como acontece, por exemplo, quando só se pode ter acesso ou comunicação a uma parte do edifício, como seja, um terraço, através

da fracção autónoma de um condómino.', 'IV - A amplitude normativa conferida pelo legislador ao regulamento do condomínio que faça parte do título constitutivo da propriedade horizontal e a sua força vinculativa é muito maior do que a que é conferida a regulamento que seja aprovado pelos condóminos em momento posterior à constituição da propriedade horizontal. Para além de outras diferenças, importa realçar que o regulamento integrante do título pode regular e disciplinar não só a utilização das partes comuns, mas também o uso e fruição das fracções autónomas.', 'II - Não é possível através da interpretação dos negócios de constituição de propriedade horizontal e de compra e venda celebrados entre a 1.ª Ré e os condóminos Autores, concluir-se pela integração de "Clube de Lazer", como parte comum, no Edifício constituído em propriedade horizontal, se tal Clube está instalado em terreno que não pertence ao prédio a que a escritura pública se refere e onde foi construído tal Edifício, isto independentemente do sentido de vontade real (ou normativa) das partes que outorgaram naqueles negócios formais (art. 238.º do CC).', 'VI - O negócio realizado pelas partes consistiu na cessão de quotas de uma sociedade. Como contrato translativo de titulares, tem o negócio, como elementos essenciais, o reconhecimento/especificação das quotas a ceder e o preço da cedência (para além, obviamente, da identificação dos sujeitos intervenientes). Assim sendo, a estipulação quanto ao preço de alienação das quotas constitui uma cláusula essencial, devendo ser inserida na escritura pública. Não constando desta, é tal estipulação nula (art. 220.º do CC).']

Questão: Que tipo de vício pode tornar nula uma cláusula que altera a finalidade de uma fração autónoma num prédio em propriedade horizontal?

Resposta:

APPENDIX B

Appendix: List of Documents from Literature

Table B.1: List of Documents from Literature

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Visciarelli *et al.* [9]	2024	PDFs of the regional laws of Emilia-Romagna	LLaMA-2, Mistral	LLM Fine-tuning, Retrieval, RAG	Legal Search	Metric (Perplexity) + Experts
Amorim *et al.* [85]	2024	ACE 2005 dataset and the Lusa News dataset	GPT3	Information Extraction, Knowledge Graph	Not applicable	Metrics (Precision, Recall, and F1)
Hu, Luo, and Feng [63]	2024	LeCaRD	LLaMA	Legal Article and Case Retrieval, Embedding Model Fine-tuning, Cosine Similarity, RAG	Legal Search	Human Evaluation (User Tests)
Louis, Dijck, and Spanakis [64]	2024	LLeQA	LLaMA, Vicuna-1.3, WizardLM-1.0, TULU, Guanaco, gpt-3.5-turbo	ICL, LLM Fine-tuning, RAG	Legal Search	Metrics (METEOR, F1)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Zhou, Huang, and Wu [65]	2023	LeCaRD	GPT-3.5-Turbo	Retrieval, Information Extraction	Legal Search	Metrics (Recall, Precision, P@5(%), P@10(%), MAP(%), NDCG@10(%), NDCG@20(%), NDCG@30(%))
Coelho *et al.* [71]	2024	959 manually annotated legal opinions (in Brazilian Portuguese) enacted by lower court judges in the State Court of Rio de Janeiro in the context of consumer complaints involving electric power companies	GPT-3.5-Turbo	NER, Text Classification, Information Extraction, BERT Fine-tuning	Legal Search	Metrics (Precision, Recall, F1-score, Accuracy, RMSE)
Garlyal, Hariharan, and Singh [51]	2024	Indian Penal Code	LLAMA 2, MISTRAL and PHI2	Retrieval (MIPS), RAG	Legal Search	Metrics (ROUGE-N and ROUGE-L)
Moreira *et al.* [72]	2024	352 LLM-Generated Question-Answer pairs based on 50 Brazilian legal documents	GPT4 Turbo	BERT Fine-tuning, UMAP for Dimensionality Reduction, DBSCAN for Clustering of questions and answers	Legal Search	Metrics (Precision, Recall, F1 Score, Divergence (DIV) to measure the disagreement between two detected precedents of a legal matter)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Litaina *et al.* [52]	2024	17 handwritten Greek Contracts of the XIX century	GPT-3.5, Gemini	Information Extraction (NER, Relationship Identification)	Legal Document Review	Metrics (Precision, Recall) calculated from the comparison with the answers of a legal expert
Mamalis *et al.* [49]	2024	428 extracted documents of EUR-Lex which included U treaties, legislative acts, international agreements, and preparatory documents and european case-law	GPT3.5, GPT4	Question Generation, RAG	Legal Search	Qualitative Analysis by experts (Expert defined set of questions (12) related to GDPR divided into groups based on the degree of difficulty of the answers and analyzed the answers of the LLM)
Nai *et al.* [53]	2024	3 Tables with information relative to Procurement, Contract Authority, Economic Operator respectively extracted from ANAC	GPT models, Cohere	Information Retrieval, Elasticsearch, RAG	Legal Search	Metric (NDCG)
Wiratunga *et al.* [66]	2024	Australian Open Legal QA (ALQA)	Mistral-7B-open	Information Retrieval, CBR, Similarity-based KNN for Retrieval, RAG	Legal Search	Metrics (F1, Cosine, ANOVA)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Li and Yi [62]	2024	manually created a set of 100 training samples for legal entity relation extraction tasks as a seed pool. From this pool, we randomly selected six manually crafted samples and an additional two samples generated by the model in the previous step, CAIL2022, CCKS2021	Baichuan2-13B-Chat	LLM Fine-tuning, Few-shot prompting, In-Context Learning, Information Extraction	Legal Search	Metrics (Precision, Recall, F1)
Cherubini *et al.* [10]	2024	European legal acts taken from EUR-Lex repository	gpt-4	Information Retrieval, KNN, RAG	Legal Search	Metrics (Mean Reciprocal Rank based on an old dataset using expert annotators) + Human Evaluation (Experts score answers)
Amri, Bani, and Bani [54]	2024	Moroccan labor law data, including articles from the Moroccan Labor Code and the Dahirs (royal decrees)	LLaMa-70B	Information Retrieval, RAG	Legal Search	Not specified

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Yao *et al.* [50]	2024	BELLE, MOSS, Scrapped Chinese Website Information, Lawyer Llama, DISC-Lawllm, Alpaca	ChatGPT, Qwen-14B-Chat	CoT prompting, LLM Fine-tuning, Recommendation System, Information Retrieval, Document Re-ranking, Kmeans, KCenter, Reinforcement Learning	Legal Search	Human Evaluation (Experts used metrics like Completeness, Reliability, Logical Coherence, User friendliness, Legal ability, Intention Understanding, Harmlessness to evaluate the answers of the LLM)
Cho *et al.* [73]	2024	4 human-labeled, and 192 LLM-labeled examples	GPT3.5-Turbo, GPT4	Information Extraction, Retrieval, Elasticsearch, Few-shot prompting	Legal Search	Metric (F1)
Nunes *et al.* [70]	2024	2 Brazilian legal datasets for NER	Sabia (LLaMa 7B fine-tuned)	Information Extraction (NER), In-Context Learning, Cosine Similarity	Legal Search	Metrics (micro F1, precision, recall)
Ryu *et al.* [55]	2023	Korea Legal Aid Corporation and Korea Legislation Research Institute data and LLM-Generated Questions for the data	ChatEval and FairEval (GPT3.5, GPT4)	Information Retrieval, Question Generation, RAG	Legal Search	Metrics (Pearson, Spearman, and Kendall correlation coefficients) + Human Evaluation (human grading evaluated by the lawyers)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Hoang *et al.* [67]	2023	ALQAC 2022/2023, Zalo, LLM-Generated Questions	ChatGPT, Flan-T5-XXL, Flan-Alpaca-XXL, VietAI/envit5-translation	Lexicon, Zero-shot prompting, CoT Prompting, JSON Prompt Formatting, Translation	Legal Search	Metrics (Precision Recall F2-macro, Accuracy)
Zin *et al.* [68]	2023	CUAD (QA dataset with 510 contracts)	GPT3.5	Summarization, Retrieval, Reranking, Embedding model fine-tuning, Information Extraction	Legal Search	Metrics (F1, Exact Match)
Adhikary *et al.* [74]	2024	Gold standard 200 legal documents annotated by law students, collection of 6854 legal case documents from the Supreme Court of India from 1958 to 2016	GPT2, GPT3.5-Turbo, GPT4, LLaMa-2	LLM Fine-tuning, Few-shot prompting, Weak Supervision, Entity Extraction, Judgment Prediction	Legal Search	Metrics (Accuracy, Precision, Recall, F1-score)
Zambrano [56]	2024	400 court decisions	LLaMA3, Mixtral 8x7B	Zero-shot prompting, Information Extraction, JSON prompting	Legal Search	Metrics (Accuracy, precision, recall and F1 score)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Gabriela Argüelles Terrón [57]	2023	Spanish Worker's Statute	GPT-3.5	Zero-shot prompting, 3-shot prompting, 5-shot prompting, Event Extraction, In-Context learning	Legal Search	Metrics (Precision, Recall, F1)
M, P, and M [58]	2024	7130 documents	GPT4, LLaMa-2, BART, T5, Pegasus, (Legal Pegasus, Legal BERT, Encoder-Decoder	Summarization, Text Rank, Lex Rank, LSA, BERT and KL Summ, Zero-shot prompting, One-shot prompting,	Legal Search	Metrics (ROUGE - 1, ROUGE - 2, ROUGE - L, BERT, SCORE)
Reji *et al.* [37]	2024	COLIEE	Mistral 7B Instruct v0.2	Cosine Similarity, Few-shot Prompting, Zero-shot prompting, IRAL prompting, CoT prompting, Naive greedy decoding Generation, Beam search Generation, Contrastive Search Generation, Multinomial sampling Generation, RAG	Legal Search	Metrics (Accuracy, precision, recall, F1 score)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Espírito Santo *et al.* [59]	2024	172,408 Scrapped Brazilian legislation and LLM-Generated QA dataset for Fine-tuning	LLaMa-2-7B, GPT3.5, Sabia, Maritalk	LLM Fine-tuning	Legal Search	Metrics (BLEU, ROUGE-N, BERTSCORE, MOVER-SCORE) + Qualitative analysis made by the researchers
Kaoutar *et al.* [60]	2024	Documents of Canadian law	GPT3.5, Gemini, Mistral-7B-Instruct-v0.2, Gemma-7B, Falcon 180B, Llama2 70B	LLM Fine-tuning, Semantic Chunking	Not applicable	Based on a set of 40 questions provided by a tax expert
Alif Adillah *et al.* [8]	2024	Indonesia Non-Convention Vessel Standards (NCVS) books	GPT3.5-Turbo	RAG	Legal Search	Metrics-based Similarity (precision@k, recall@k, F1, and Mean Reciprocal Rank)
Anh *et al.* [2]	2023	COLIEE 2022	Flan-t5-xxl, Flan-Ul2, T0pp, Bloomz-7b1, Mt0-xxl	Casual LM Generation, Masked LM Generation, LLM Fine-tuning	Not applicable	Metrics (Accuracy)

Continued from previous page

Authors	Publication Year	Dataset	LLM	Techniques	Task	Evaluation Method
Chauhan *et al.* [61]	2024	2.5GB of Indian Legislation with articles and case law	Not specified	LLM Training, Cosine Similarity, Intent and Entity Recognition, Information Retrieval	Legal Search	Metrics (Accuracy)
Kurniawan and Hierera [5]	2024	Public Corpus of Indonesian Legislation	ChatGPT	RAG	Legal Search	Human Evaluation based on a Likert scale and Interviews with Experts
Bakir, Yildiz, and Aktas [69]	2023	SQUAD	LLama2-7B	LLM and Embedding Training, Similarity Measure Testing	Legal Search	Metrics (ROUGE, BLEU, METEOR, Cosine Similarity, F1, Precision, Jaccard, RMSE, MAE) and Human Evaluation
Jacob *et al.* [7]	2024	Curated legal documents, case laws, and relevant legal literature	Mistral-8x7B-Instruct-v0.1	RAG	Legal Search	Metrics (BLEU, ROUGE, and METEOR)
S *et al.* [75]	2024	Not specified	LLaMA-2, FLAN-T5-Base, Claude2	Summarization	Legal Search	Metrics (ROUGE 1, ROUGE 2, ROUGE L, BERT (Precision), BERT (Recall), BERT (F1))