Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5703

Turbocharging Brazilian Mergers and Acquisitions: Questions & Answers Evaluation

Francis Spiegel Rubin [**Federal University of the State of Rio de Janeiro (UNIRIO)** | ***fran.spiegel@edu.unirio.br***]
Pedro Nuno de Souza Moura [**Federal University of the State of Rio de Janeiro (UNIRIO)** | ***pedro.moura@uniriotec.br***]
Adriana Cesario de Faria Alvim [**Federal University of the State of Rio de Janeiro (UNIRIO)** | ***adriana@uniriotec.br***]

Graduate Program in Informatics (PPGI) – Federal University of the State of Rio de Janeiro (UNIRIO) – Rio de Janeiro, RJ, 22290-255, Brazil.

Received: 01 April 2025 • **Accepted:** 22 October 2025 • **Published:** 17 March 2026

Abstract. Economic power abuse is a concern in Brazil, where CADE (Administrative Council for Economic Defense) institution combats anti-competitive behaviors to ensure fair competition. Artificial intelligence (AI) can aid CADE by identifying and extracting relevant information from technical reports published in Brazilian Portuguese language, improving the detection and prevention of economic abuse. This paper presents a case study using AI to improve regulatory reviews of CADE documents via a Retrieval-Augmented Generation (RAG) pipeline architecture. Our key contribution is the creation of a specialized Questions & Answers benchmark dataset and a pipeline evaluation methodology, providing a standardized framework for Portuguese-language regulatory document analysis. A chain of thought (CoT) approach was used for problem solving. It leverages the RAG retrieval mechanism to access relevant information and incorporates the sequential reasoning of the CoT framework to generate responses that follow a logical flow of ideas, thus enhancing response accuracy. A vector database employing cosine similarity was used to retrieve the main arguments combined with metadata filters, reducing hallucinations and improving the Large Language Model (LLM) performance. RAG metrics were then combined with a robust human fact-check assessment to validate the pipeline. Our findings establish a new benchmark for Questions & Answers evaluation in Brazilian Mergers and Acquisitions, demonstrating that the proposed strategy effectively enhances the analysis of organizational merger and acquisition reports, unlocking substantial benefits for society.

Keywords: Retrieval-Augmented Generation, Questions & Answers Evaluation, Large Language Models, Natural Language Processing.

1 Introduction

Information Retrieval (IR), as an essential technique in the realm of data mining, aims to understand the organization, indexing, and retrieval of information from extensive repositories [Zhu *et al*., 2024; Baeza-Yates *et al*., 1999]. It focuses on developing methods and algorithms to efficiently search, identify and rank relevant documents based on user queries. It has been applied in various domains [Buttcher *et al*., 2016; Kolomiyets and Moens, 2011; Rajput *et al*., 2023], encompassing tasks such as search, question answering, and recommendation systems. More recently, retrieval models, through the proficient management of external databases, have been delivering accurate and timely external knowledge, playing a pivotal role in various knowledge-intensive tasks. Due to their capabilities, retrieval techniques have been seamlessly integrated into advanced generative models in the era of Artificial Intelligence (AI) [Izacard and Grave, 2021]. The fusion of retrieval models and language models, leading to the creation of Retrieval-Augmented Generation (RAG), is a significant and noteworthy advancement [Yu *et al*., 2023], which has turned up as a prominent technique in the domain of generative AI.

The RAG concept combines the use of Large Language Models (LLMs) with external knowledge retrieval to enhance the generation of responses. Instead of relying solely on the language model's internal knowledge, RAG retrieves relevant information from external sources and incorporates it as additional context. This approach enables the generation of more accurate and informed responses [Zhao *et al*., 2022; Touvron *et al*., 2023], while also helping to reduce hallucinations from the LLM. In addition, it is beneficial for knowledge-intensive tasks and effectively addresses limitations inherent in traditional LLM generation methods. By integrating a retrieval component that searches for relevant information from a large corpus and a generative component that constructs responses based on both the retrieved information and the input query, RAG techniques provide more accurate, context-sensitive and informative outputs [Ram *et al*., 2023]. This hybrid approach combines the precision of retrieval systems with the flexibility of generative models to capture semantic relationships between queries and documents, thereby improving performance in tasks such as question answering and knowledge-intensive dialogue.

However, finding the best search method for RAG is still an emerging area of research. This paper applies a pipeline architecture to enable accurate IR of contextual data from

CADE[1] (Administrative Council for Economic Defense), the Brazilian Competition Authority, by integrating intermediate answers from various sections of CADE's documents. These documents provide analyses, findings, and recommendations on aspects of market competition, such as mergers, acquisitions, antitrust practices, and other economic activities. In this context, we applied chaining steps that prompt LLM, often referred to as chain-of-thought (CoT) [Wei *et al*., 2023; Huang and Chang, 2022], particularly useful for tasks requiring multistep reasoning [Zhou *et al*., 2022]. These steps can help break down the problem systematically into a list of simpler subproblems to simplify the generation of the correct answer. Thus, our main contribution is leveraging the CoT approach by proposing a Questions & Answers (Q & A) pipeline that instructs the model to "think" step-by-step, enabling it to efficiently respond to questions related to CADE's documents published in Brazilian Portuguese. Although most of the techniques (RAG, prompt, CoT, and vector database) are based on existing literature, the novelty lies in applying these techniques to this particular domain, which, to the best of our knowledge, has not yet been explored in the literature. Additionally, we introduce a structured fact-checking Questions & Answers evaluation framework designed to guide CADE experts in assessing responses based on key criteria elements. This second contribution ensures accurate and reliable evaluations, further strengthening the trustworthiness of our approach in the context of analyzing CADE documents and addressing the specific challenges inherent to this domain.

In each step, the prompt serves as an instruction that guides the LLM to generate a suitable response to a question, based on the context provided in the previous step. We employed a vector database to store high-dimensional embeddings of the generated context and used a cosine similarity distance function to identify the most similar documents, thus reducing unintended AI-generated texts, known as "hallucinations" [Ji *et al*., 2023]. By breaking down the problem into logical steps and using metadata retrieval filters, we managed to enhance the LLM's reliability. Finally, we carried out experiments that showed that our proposal was able to achieve better results than a standalone LLM baseline (without RAG), according to the evaluation metrics adopted.

The remainder of this paper is organized as follows: In Section 2, we present the background necessary to understand this work. In Section 3, we discuss related works. Section 4 presents our proposal. Section 5 describes the methodology used. Section 6 presents the experiments. Section 7 details the results obtained. In Section 8, we discuss the results and limitations. Finally, in Section 9, we conclude the article with final considerations and future work.

2 Background

The analysis of antitrust documents presents major challenges for analysts due to the complexity and variability of legal standards across jurisdictions. Golovanova *et al*. [2025] empirically demonstrates this through their comparative study of BRICS countries, showing that "countries differ in the legal standard used, even on the same conduct type" with Brazil and India applying higher standards than Russia and South Africa. This variability underscores the need for automated tools that can efficiently process technical antitrust documents and extract relevant information consistently. Therefore, a proposed retrieval approach system based on CADE documents can significantly reduce the time-consuming tasks associated with document analysis. Since CADE analysts are responsible for deriving conclusions about economic policies and government official recommendations, it is important to follow the same logical flows that lead to a decision-making policy. These logical flows include the structured sequence of analytical steps, the criteria used to evaluate information, the connections between relevant data, and the reasoning methods applied to support policy decisions and recommend government actions. A standard keyword search engine would have a limited impact on this goal because CADE documents do not always follow the same pattern for similar decisions. For example, we can have two documents with the same players and with similar merging and acquisition operations that lead to different decisions because of different clauses, so if a third operation requires a new analysis, it is important to have a thorough automated access to these past documents before making a new decision, with a summarized context consolidation (arguments used, constraints applied, etc.) and indication of approval.

2.1 CADE Documents

The Conselho Administrativo de Defesa Econômica (CADE), in English *the Administrative Council for Economic Defense*, has the authority to investigate and prosecute infractions that violate the principle of economic order in Brazil. The documents produced by CADE are diverse, written in Brazilian Portuguese, and cover various economic aspects of competition defense. The primary purpose of these reports is to provide a technical foundation for the council's decision-making process. In this research, we focus on the analysis of CADE's documents classified as Technical Reports because they are critical for providing in-depth analyses of economic competition-related cases, including mergers, acquisitions, and anticompetitive practices.

Given the official nature of these regulatory documents, our research methodology prioritizes ethical data handling by using only publicly available CADE technical reports from official government channels. These documents are intended for public review, reflecting the institution's commitment to transparent decision-making while serving as foundational instruments for regulatory efficiency in its mission to promote competition and combat anticompetitive practices in Brazil. It should be noted that CADE documents have previously had sensitive data hidden. They offer regulatory recommendations and assess economic impacts, playing a key role in ensuring transparency, accountability, and promoting fair competition. The official CADE website[2] provides detailed and official information on the types of documents and their functions by directly consulting the reports and publications available in the document section of the site. Furthermore, Brazilian legislation related to competition defense[3] also pro-

[1] https://www.gov.br/cade/en/access-to-information/about-us

[2] http://www.cade.gov.br/

[3] Law No. 12.529/2011

vides relevant information about CADE's operations and responsibilities.

A typical technical report structure typically follows a logical flow. It begins with an introduction to the involved parties and the formal aspects of the operation. Then it provides a detailed description of the operation, including its nature, financial aspects, and intentions. The document continues with an analysis of the market impact, discussing the market shares, potential overlaps, and the competitive landscape, which involves market concentration and competition. Finally, it concludes with the decision and the recommendation for approval, ensuring clarity and thoroughness, stating if the operation raises competitive concerns. Therefore, AI is believed to be able to promote automation in this context and assist in legal compliance checks and precedent analysis by systematically extracting the main arguments, players, conclusions, among others, of these documents.

2.2 Natural Language Understanding

Natural Language Processing (NLP) has been mentioned since the early days of AI as a key technology to enable machines to understand and interact with human language [Joshi, 1991; Manning, 1999]. Over the decades, advances in NLP have revolutionized various fields by automating text analysis, sentiment detection, language translation, among others. Such algorithms are applied in different tasks such as retrieving texts (IR), splitting them into parts, checking the spellings, and word-level analysis. However, when it comes to the question of interpreting sentences and extracting meaningful information, the capabilities of these algorithms are still very limited [Chowdhary, 2020]. To shift from mere NLP to what is usually called Natural Language Understanding (NLU), systems must go beyond basic text processing and incorporate deeper semantic comprehension. This involves understanding the context, intent, and nuances of human language, enabling machines to interpret meaning accurately, handle ambiguities, and generate coherent responses [Singh and Mahmood, 2021].

In this context, in 2017, the transformer model [Vaswani, 2017] laid the foundations for many advances in NLU by enabling a more effective handling of contextual information, revolutionizing the field of NLP by leveraging self-attention mechanisms to capture dependencies between words in a sentence, regardless of their distance. This approach drastically improved the performance of language models, leading to the development of state-of-the-art large language models (LLMs) such as GPT-3 (Generative Pre-trained Transformer) [Brown *et al*., 2020], which excels at tasks requiring deep contextual understanding and text generation. Using the decoder part of the Transformer architecture (or a combined encoder-decoder architecture in some cases), the model learns to predict the next k words in a sequence given the previous t words (context).

However, as digital information grows exponentially, purely generative models struggle to produce accurate and relevant responses without incorporating external knowledge. RAG [Lewis *et al*., 2020] frameworks address this challenge by retrieving pertinent information from large unstructured datasets from a given domain and using it to guide the generation process. Using transformers, RAG architectures can effectively retrieve relevant information from large unstructured datasets and generate coherent, contextually accurate responses, thus improving performance in tasks that require deep contextual understanding for NLU.

3 Related Work

In recent years, AI adoption in legal document analysis has become prevalent, automating tasks such as document classification, contract analysis, legal research, and predictive analytics. Legal professionals are leveraging AI to streamline document review processes, improve efficiency, reduce costs, and mitigate human errors. In this context, several studies report the use of AI to accelerate the speed of analysis and assist professionals in retrieving and classifying documents, which is often very useful for searching previous documents [de Sousa *et al*., 2022]. Fernandes *et al*. [2022, 2020] proposed a methodology to extract value from Brazilian Court decisions to support judges and lawyers in their decision-making. Lai *et al*. [2023] published a survey to provide an overview of AI applications in legal text processing, highlighting advances in the field. Pinto *et al*. [2020] published a study to investigate biased language detection in court decisions. Ashley [2017] explored how AI and analytics are transforming law practice, including the analysis of legal documents. Wei *et al*. [2018] conducted experiments to compare deep learning results with results obtained using a machine learning algorithm (Support Vector Machines) on legal matters. Shaheen *et al*. [2020] studied the performance of various recent transformer-based models in combination with strategies such as generative pretraining.

Additionally, the evaluation of question-answering capabilities in LLMs has emerged as a critical research area, with several important contributions that address the challenges of creating robust evaluation frameworks. For example, Mohseni *et al*. [2018] established a human-based evaluation benchmark specifically to assess explanations in machine learning systems, highlighting the essential role of human judgment in evaluating AI-generated content. Building on this foundation, the integration of human evaluation with automated metrics has been further investigated by Stassin *et al*. [2023], whose experimental study on explainability method evaluation demonstrated that combining these approaches results in more reliable assessments compared to relying on either method in isolation. In another study, Li *et al*. [2024b] introduced an LLM-based automation agent that can simulate human behaviors to automatically generate and evaluate interactions for question answering.

Despite the brief concepts and studies presented, AI-assisted legal document analysis still faces challenges such as hallucination, where LLM generates information that is plausible in the given context but incorrect or not grounded in the input data [Ji *et al*., 2023]. This issue can undermine the reliability and accuracy of AI-assisted documents. However, leveraging advanced techniques such as RAG [Fan *et al*., 2024], improving training data quality and incorporating human supervision can help mitigate these challenges, leading to more reliable and accurate AI-assisted legal processes [Lewis

et al., 2021; Li *et al*., 2024a; Shuster *et al*., 2021].

4 Our Proposal

The advent of LLMs such as GPT has caused a paradigm shift in natural language processing [Yenduri *et al*., 2024]. The integration of LLM techniques and IR tasks opens new possibilities for more accurate and contextually relevant search results, improved question-answering systems, and enhanced information extraction from large textual datasets [Wang *et al*., 2024; Dai *et al*., 2024], leveraging the capabilities of IR systems to include both human-content and LLM-generated texts. Although LLMs can produce rich responses, they can introduce irrelevant or false information, also known as hallucination, in natural language generation [Ji *et al*., 2023; Lewis *et al*., 2021]. To address this issue and mitigate the outcome of hallucination-generated texts, we apply a RAG pipeline, capable of answering complex questions about the retrieved documents from CADE. It combines retrieval-based and generation-based models to improve natural language understanding and generation tasks.

Our approach employs chained LLM techniques applied to a comprehensive *corpus* of documents obtained from CADE's website, enriched with domain-specific metadata filters. These documents include conclusions indicating whether the analyzed subject was recommended for approval, along with information about the applicants. In the antitrust context, applicants are the parties involved in the transaction or operation under review, such as merging companies, entities forming joint ventures, or organizations seeking clearance. They provide supporting documentation to demonstrate compliance with competition laws and justify the operation's impact on market dynamics.

We propose an AI-driven system to assess the capability of LLMs in retrieving and addressing domain-specific queries related to CADE. The following queries have been thoughtfully designed as the initial steps to extract and structure data from CADE's legal opinions. The objective is to systematically organize the information, ensuring the identification of key elements in the analysis. This structured data will serve as the basis for creating a comprehensive summary (global summary) of the legal opinions, offering a clear and concise overview of the decisions, involved parties, contractual provisions and supporting arguments. By prioritizing these core questions, the system aims to streamline data extraction and enhance the efficiency of summarizing complex regulatory documents.

1. Was the operation approved (YES/NO)?
2. What are the names of the applicants?
3. Is there a non-competition clause?
4. What is the conclusion or recommendation, along with its supporting arguments?

Using AI-assisted legal document analysis offers several advantages over standard keyword searches, particularly in terms of efficiency, accuracy, and insight generation [Torre *et al*., 2020; Sil *et al*., 2019; Hilabadu and Zaytsev, 2024; Nay *et al*., 2024; Chakrabarti *et al*., 2018], as follows:

- Improved Efficiency: AI systems process data much faster than humans, reducing the time required to complete complex tasks, such as reviewing large volumes of documents.
- Enhanced Accuracy: By minimizing human error and understanding context, AI systems can deliver more precise results and reduce oversight risks.
- Cost Savings: Automating repetitive tasks with AI can reduce long-term operational costs by reducing the need for manual labor.
- Actionable Insights: AI can analyze data and provide recommendations, summaries, or alerts, enabling better decision making.
- Adaptability and Learning: Many AI systems improve over time through machine learning, refining their output based on new data and feedback.
- Consistency: AI improves uniformity in processes and results, avoiding variations caused by human interpretation or fatigue.
- Risk Mitigation: By identifying patterns, anomalies, or potential risks, AI improves compliance with standards or regulations.

The proposed AI-based system is designed primarily for analysts and consultants at the CADE antitrust institution, but its utility extends beyond this audience. It is equally valuable for academics, researchers, and other professionals studying competition-related topics. By enabling the systematic extraction and structuring of key information from CADE legal opinions, the system facilitates deeper insights into antitrust decisions and market dynamics. This makes it a powerful tool not only for regulatory analysis but also for academic research and policy development in the domain of competition. These professionals currently rely on manual document review to extract arguments that support opinions and decisions. Thus, the core objective of the system is to automate this process, making it more efficient and less time-consuming.

5 Methodology

The proposed methodology aims to introduce an approach to expedite the analysis of CADE documents related to the outcomes of mergers and business transactions between organizations, ultimately generating a positive social impact. Our methodology consists of three phases.

The first phase focuses on context retrieval and structured reasoning, guiding the model in breaking down the complex problem of analyzing mergers and business transactions into manageable and logical intermediate steps by retrieving relevant information from technical reports, defining discrete sub-tasks (e.g., assessing regulatory compliance or economic impact), and logically sequencing these steps to ensure coherent and context-informed reasoning throughout the process.

The second focuses on the evaluation of our pipeline, which is automatically performed by another LLM using an open-source knowledge graph-based framework designed to measure the effectiveness of retrieval-augmented systems. This approach ensures a structured and efficient assessment of the pipeline's ability to extract and organize data from CADE

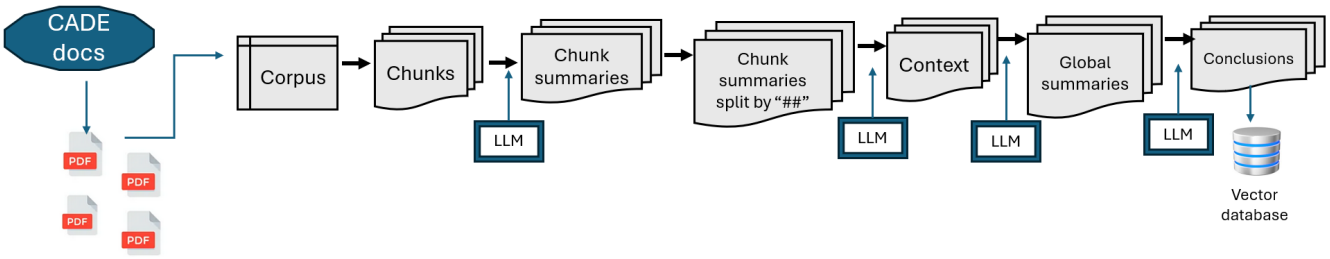

Figure 1. First phase schema pipeline for processing CADE documents in PDF format. Text is extracted into a corpus, split into x chunks, and summarized at multiple levels using an LLM—first into concise chunk summaries (`ccs`), followed by chunk summary blocks (`sb`). These arguments serve as input to the LLM, to generate the summary generated contexts (`sgc`), followed by global summaries (`gs`). Finally, the central key elements are distilled into the conclusion (`c`). These conclusions are stored in a vectorial database for efficient retrieval and analysis. Each entry is represented as a vector in a multidimensional space with global summary embeddings, accompanied by metadata including approval status (YES|NO), applicants, the non-compete clause, and the conclusion or recommendation. The LLM is pivotal at every summarization step.

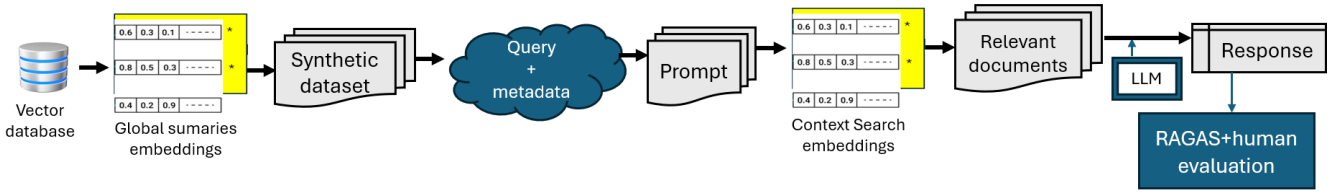

Figure 2. Second phase schema depicts our RAG+LLM search and retrieval process: global summary embeddings from our vectorial database are used to create a synthetic dataset. Combined with a query and metadata, a prompt is generated, leading to a context search for relevant CADE documents. These documents are processed by an LLM to generate a response, which is evaluated using RAGAS framework and human validation assessment for better quality assurance.

legal opinions, leveraging the advanced capabilities of LLMs for accurate and transparent evaluation.

Since the goal of our research is to provide valuable information for human users, relying solely on LLM-based evaluation can make it difficult to fully demonstrate the strengths and limitations of the proposed solutions. To address this, the third phase focuses on validating the LLM evaluation metrics through human fact-checking assessments. Finally, we benchmark our RAG+LLM results against a standalone LLM baseline for comparison.

Our methodology unfolds in a sequence of progressive steps, by applying a combined series of retrieval (RAG) and reasoning technique (CoT). Figure 1 illustrates the flow of our proposal for the dense retrieval mechanism, which contains a first phase, responsible for preprocessing our input documents into summarized and conclusion snippets, containing the answers to our domain-specific questions: items 1 to 4 from Section 4. This is practical as the input corpus is small, with fewer than 900 documents. Figure 2 illustrates the main activities of our second phase, responsible for the search and retrieval flow and the evaluation of our RAG+LLM pipeline.

5.1 First Phase

The first phase of our methodology builds a knowledge base that can extract the main arguments, operations, and conclusions from the collected documents in a way that can be later consumed and evaluated by an IR strategy. It involves the following steps:

- Text preprocessing: The input consists of n documents, where each document (e.g., a PDF) contains its own corpus, consisting of the text within the document. A Recursive Character Text Splitting technique[4] is applied to break down the corpus of each document into smaller chunks, resulting in x chunks across all documents. This technique prioritizes preserving logical structures by keeping paragraphs together (using new double lines), followed by sentences (new single line), and words (space), ensuring meaningful segmentation and avoiding unnecessary splits caused by empty strings.
- Model configuration: In a nutshell, LLMs generate human-like text by processing sequences of tokens and using nucleus sampling [Holtzman *et al.*, 2020] to select the next token from a set of probable options, ensuring that the text remains coherent and relevant.
  For processing these chunks, we employ an LLM with hyperparameters optimized for generating coherent and accurate summaries of regulatory text. The specific model selection, generation hyperparameters (temperature, top_k, top_p), and configuration details are provided in Section 6.1.
- Chunk summarization: Summarizing each chunk ensures that the core information is retained without exceeding token constraints. As input, it receives x chunks, applies the model to each chunk, and returns x coherent and concise chunk summaries (`ccs`). The following RAG prompt was applied: "You are an assistant tasked with summarizing tables and texts. Provide a concise summary of the table or text:{chunk}".
- Context generation: Chunk summary segments can be combined to form a coherent understanding of the entire document. As input, it receives x `ccs`'s, consolidate these x `ccs`'s into n summary blocks (`sb`) split by a delimiter ("##"), applies the model to each `sb` to generate y main arguments and operations mentioned in the document, and returns n summary generated contexts (`sgc`). The following RAG prompt was applied: "Answer the

[4] https://python.langchain.com/api_reference/text_splitters/index.html

question based only on the following context, which may include text and tables: {sb}. Question: 'Describe the text. Identify, if possible, the arguments and the main information of the operation."'

- Global summary generation: Chunk summaries might overlap or repeat similar points, as different sections of a document can include related ideas. A global summary consolidates and eliminates redundant information, providing a more streamlined and coherent output, ensuring that the overarching themes and key points are captured effectively. As input, it receives n `sgc`'s, applies the model to each `sgc`, and returns n global summaries (`gs`) to address: Whether the operation caused harm to competition {a}; Whether there were non-competition clauses {b}; The recommendation or indication of the conclusion and its arguments {c}. The following RAG prompt was applied: "You will receive a list of summaries. Consider the context: {sgc}. Combine the list of summaries into a global summary of the document and report whether: {a}; {b}; {c}."
- Conclusion generation: This step distills the most important takeaways from the global summary, providing a clear and concise recap of the central key elements of each document. As input, it receives n `sgc`'s and n `gs`'s, applies the model to the combination of these inputs, and returns a conclusion (`c`) generated by the following RAG prompt: "You are an expert in technical reports and should analyze the conclusion. Consider the context: {`sgc`} Consider the summary: {`gs`} Consider the source: {`filename`}. Return YES, if the response concludes or recommends that the operation was approved. Otherwise, return NO. Indicate the Applicants | Non-compete Clause | Conclusion or Recommendation."
- Knowledge base generation: An information repository, powered by a vector database, can lead to significant advancements in enabling semantic searches by transforming how information is retrieved and understood [Ghali *et al*., 2025]. Unlike traditional keyword-based search, semantic search relies on the meaning and context of queries, leveraging embeddings to find relevant results even when exact matches are absent, ensuring that retrieved content aligns with the intent behind the question or input. Furthermore, querying a vector database for relevant embeddings is far more computationally efficient and cost-effective than feeding an entire dataset to an LLM [Han *et al*., 2023], serving as the backbone for storing, searching, and retrieving the most contextually relevant pieces of information from CADE domain. As input, it receives n global summaries (`gs`) and n conclusions (`c`), and as output, the data are saved in a vector database where each entry is represented as a vector in a multidimensional space containing the global summaries embeddings, and as metadata, whether the operation was approved or not approved (YES|NO), the applicants, the non-compete clauses and the conclusion or recommendation.

  Additionally, the proposed pipeline also lays the foundation for a potential AI-driven system tailored to the antitrust Brazilian authority.

5.2 Second Phase

To evaluate the proposed pipeline, ideally, we need samples of question-context-answer triples that are annotated with human-produced ground-truth judgments. The ground truth corresponds to the verified correct answers to the questions. We could then assess to what extent LLM quality metrics agree with human assessments. At the time of this research, we were unaware of any publicly available CADE datasets that could be used for this purpose, which motivated us to create our own dataset, referenced as CADE_EVAL.

Traditional methods for manually creating question-answer samples are often time-consuming, so we developed our own dataset using synthetic test data generated from CADE documents. This dataset includes carefully crafted questions with diverse characteristics derived from the CADE domain, designed to ensure a more robust evaluation process. However, generating various types of queries from the provided set of CADE documents presents the challenge of selecting a subset of these documents to guide the LLM in producing the desired queries. To address this, we adopt a knowledge graph-based approach implemented through the open-source RAGAS framework [Es *et al*., 2023] to generate synthetic test data[5] for evaluation. Inspired by Evol-Instruct [Xu *et al*., 2023], a method that leverages LLMs rather than humans to automatically generate large volumes of open-domain instructions, RAGAS uses a structured evolutionary approach to generate questions. This entails creating questions with diverse characteristics, such as reasoning, conditioning, and multi-context, derived from the provided documents.

Initially, a collection of documents is needed to generate synthetic question/context/ground_truth samples. In our research, each document is defined by the global summary (`gs`) and contains a metadata dictionary (the approval status, the applicants, the non-compete clauses and the conclusion or recommendation). The approval status metadata field defined the two sets of scenarios for our synthetic data generation: approved (YES) and non-approved (NO) documents. Based on the defined set of documents, the RAGAS framework creates a knowledge graph using the collection we provide and applies various transformations to enrich it with additional information, to generate the test samples for our defined scenarios. The synthetic results are then exported into a table.

Then, we can focus on answering each generated question from our synthetic dataset. Each question of CADE_EVAL and its metadata information is converted into embeddings so that its similarity to other vector embeddings can be measured. This search and fetch process aims to retrieve contextually relevant data based on similarity search and is empowered by Sentence Transformers[6]. The retrieved response is determined based on the proximity (cosine similarity) between queries and document embeddings, which are stored in the vector database, then consolidated to serve as a context for generating an appropriate answer, which is also stored in the CADE_EVAL dataset. To handle increasing computational demands and mitigate potential bottlenecks, the ChromaDB vector database uses an efficient graph-based indexing strat-

[5] https://docs.ragas.io/en/latest/getstarted/rag_testset_generation/
[6] https://www.sbert.net

egy[7] to search for a given vector embedding.

With the expected questions, contexts, responses, and ground truth at hand, we can run our evaluation pipeline. RAGAS also provides a suite of metrics to evaluate the quality of our question-context-answer samples without relying on manual human annotations. In our research, we focus on the following metrics from the RAGAS framework, where, for all of them, the higher the value the better it is. Further details about the RAGAS metrics can be obtained from the author's publication [Es *et al*., 2023].

- **Faithfulness** (m_1): This metric refers to the idea that the answer should be grounded in the given context. The answer is scaled to the range $(0, 1)$. The context refers to the information or source against which the claims in the answer are being evaluated. Thus, each claim is defined as a distinct, self-contained statement or piece of information within the given answer that can be individually evaluated for its truthfulness or faithfulness to the context. The authors that proposed the RAGAS framework define a generated answer as faithful if all the claims within the answer can be inferred from the provided context.
  To assess this, a set of claims is first extracted from the generated answer. Each claim is then cross-checked against the given context to determine whether it can be logically inferred. The faithfulness score is formulated in Equation 1, given by the ratio of the number of faithful claims in the generated answer that can be inferred from the given context ($ClaimsInf$) to the total number of claims in the generated answer ($Claims$):

$$m_1 = \frac{\text{ClaimsInf}}{\text{Claims}} \tag{1}$$

  As an example, given the following question and context:

  - Question: "How many Nobel Prizes did Marie Curie win, and where and when was she born?"
  - Context: "Marie Curie won two Nobel Prizes: one in Physics (1903) and another in Chemistry (1911) and she was born in Warsaw in 1867."

  Let's consider two possible answers:

  - High-fidelity answer: "Marie Curie won two Nobel Prizes: Physics and Chemistry and was born in Warsaw in 1867."
  - Low-fidelity answer: "Marie Curie won two Nobel Prizes in Physics and was born in in 1867."

  We can calculate the faithfulness metric for the low-fidelity answer as follows:

  - Step 1: Breaking the generated answer into individual claims: Claim 1: "Marie Curie won two Nobel Prizes in Physics." (Incorrect); Claim 2: "Marie Curie was born in Warsaw in 1867." (correct).
  - Step 2: For each claim, verify if it can be inferred from the given context: Claim 1: No; Claim 2: Yes.
  - Step 3: Using Equation 1 to calculate $m1$ for the low-fidelity answer: $m1 = \frac{1}{2} = 0.5$.

- **Answer Relevance** (m_2): This metric concerns the idea that the generated answer should address the actual prompt that was provided. To calculate this score, the LLM is prompted to generate an appropriate question for the generated answer multiple times. The mean cosine similarity between these generated questions and the original question is measured. Using Equation 2, Answer Relevance can be inferred from the embeddings of the generated questions (E_{g_i}) and the embeddings of the original question (E_o), where N is the number of generated questions and i refers to the question number:

$$m_2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \cos(E_{g_i}, E_o) \tag{2}$$

- **Context Precision** (m_3): This metric measures the ability to retrieve the necessary information needed to answer the question. It ranges from 0 to 1 and represents the relative importance of contexts, indicating that the ground truth should be present within the initial sets of contexts. Using Equation 3, Context Precision can be calculated as the mean of the $Precision@k$ for each chunk in the context. $TopK$ represents the total number of relevant items in the top K results. $Precision@k$ is the ratio of the number of relevant chunks at rank k to the total number of chunks at rank k. The rank indicates the ordinal position of a chunk among the retrieved results, starting from the most relevant chunk (rank 1) and continuing down. K is the total number of chunks in retrieved contexts and $v_k \in 0, 1$ is the relevance indicator at rank k:

$$m_3 = \frac{\sum_{k=1}^{K}(Precision_k \cdot v_k)}{TopK} \tag{3}$$

- **Context Recall** (m_4): This metric measures the ability of the retriever to gather all the necessary information needed to answer the question, such as for classification tasks. Higher recall means fewer relevant documents were left out. Using Equation 4, Context Recall can be computed using the prompt (question), reference (ground-truth) and the retrieved contexts. The reference is broken down into claims. The process of analyzing each claim involves breaking the ground truth into individual claims and comparing them against the retrieved contexts to determine whether the information exists, either explicitly, paraphrased, or supported. Claims are labeled as relevant if found in the retrieved context, or as not relevant if absent from the retrieved context. This comparison is repeated for all claims and the results are aggregated to evaluate the retrieval system's completeness. $ClaimsGT$ refers to the ground-truth claims that can be attributed to the context, and $NClaimsGT$ refers to the number of ground-truth claims.

$$m_4 = \frac{ClaimsGT}{NClaimsGT} \tag{4}$$

Our work is motivated by the need to evaluate these metrics within our RAG pipeline in CADE's domain, establishing a foundation for evaluating baseline RAGAS metrics tailored to technical reports from Brazilian government institutions.

[7] https://cookbook.chromadb.dev/core/concepts/

To address this challenge, synthetic questions were generated and evaluated across two distinct scenarios: approved and non-approved document reports, and the obtained results were saved for analysis.

5.3 Third Phase

To evaluate the alignment between RAGAS predictions and human judgments, as well as to validate our evaluation metrics, we propose a fact-checking assessment framework performed by human experts in the CADE domain.

The criteria for a human fact-checking assessment are critical to ensure a systematic and reliable approach to verifying our RAG pipeline. To propose a framework for assessing trustworthiness in RAG systems, key criteria should include source credibility, bias-free, objectivity, timeliness, clarity, and cross-verification, among others Zhou *et al*. [2024]. The credibility of sources is paramount, which requires the use of reliable and authoritative references. Maintaining bias-free objectivity is essential to ensure neutrality throughout the process. Timeliness is another critical factor, as answers must be evaluated using current and relevant information, taking into account any subsequent developments. The clarity of the answer is also important, as specific and unambiguous statements are easier to fact-check than vague or generalized ones. Lastly, cross-verification with multiple independent sources ensures consistency and reduces the risk of misinformation.

5.3.1 Fact-Checking Framework

To ensure accurate and reliable assessments, we propose a structured fact-checking framework aimed at guiding CADE experts in evaluating responses based on key criteria elements. This framework is designed to systematically verify the accuracy, relevance, completeness, and consistency of information. The components of the framework are as follows:

- Answer Identification: Clearly and transparently identify the response being assessed. Example: "The acquisition of EMI Music Publishing by Sony Corporation of America was approved by CADE without restrictions." This statement explicitly identifies that CADE approved the acquisition and specifies that no restrictions were imposed.
- Context Gathering: Investigate CADE's role and responsibilities in the approval process for mergers and acquisitions. This step is crucial for understanding the key factors and considerations that guide CADE's decisions, such as market concentration, competition risks, and potential benefits to the public interest. By thoroughly analyzing these aspects, we can identify and define the specific evaluation criteria needed to assess the relevance and completeness of the retrieved information.
- Evidence Collection: Gather supporting materials, such as official documents or opinions from CADE, that directly address the response being evaluated. Example: Supplement these findings with news reports or press releases from involved parties, such as Sony or EMI Music Publishing, to provide a broader perspective on the approval process.
- Key Fact Verification: Verify the accuracy of the key assertions within the response. Example: Determine whether CADE approved the acquisition and, if so, whether any restrictions or conditions were imposed.
- Answer Evaluation: Assess the answer based on the evidence collected. If CADE's official report confirms the approval but mentions conditions, the answer is considered partially true. If no restrictions were imposed, the answer is considered fully true. If the acquisition was not approved, the answer is considered false.
- Conclusion and Documentation: Consolidate the findings into a final judgment. Ensure that the decision is well documented, including citations for all evidence and supporting materials used in the evaluation.

With this framework, we propose that researchers specializing in the field of competition, who are familiar with the analysis of CADE documents, evaluate key criteria elements (`kce`) for each response to ensure a thorough assessment. These criteria include:

- Accuracy (kce_1): Does the answer align with the verified ground truth?
- Relevance (kce_2): Does the answer directly address the question being asked?
- Completeness (kce_3): Does the answer provide all critical information required?
- Consistency (kce_4): Is the answer logically coherent and aligned with the informed context?

This structured approach ensures that the quality and reliability of the answers are rigorously assessed, providing CADE experts with a robust tool for their evaluations.

Figure 3 outlines our human fact-check assessment and how the key criteria elements are interconnected and applied at different stages of the question-answering process to ensure that the generated answer aligns with the input question and context. The accuracy ($kce1$) checks if the final answer matches the reference or ground truth, validating the overall reliability and quality of the system. Relevance ($kce2$) ensures that the retrieved documents and the final answer directly address the question. The completeness ($kce3$) verifies if the context includes all critical information from the retrieved documents. Consistency ($kce4$) ensures that the generated answer is grounded in and aligned with the provided context.

The collected findings were then classified with a numeric score to measure the concluding judgment of each criterion. Then, for a balanced and accurate holistic assessment, we propose a combined final score ($C_{\texttt{final(q)}}$) for each question q that integrates a human judgment ($S_{\texttt{human(q)}}$) with our RAGAS evaluation metrics ($M_{\texttt{ragas(q)}}$). The formulation is presented in Equation 5:

$$C_{\texttt{final(q)}} = w_1 \cdot S_{\texttt{human(q)}} + r \cdot w_2 \cdot M_{\texttt{ragas(q)}} \quad (5)$$

where:

$$S_{\texttt{human(q)}} = \frac{1}{n}\sum_{i=1}^{n} kce_{\texttt{i(q)}} \quad \text{and} \quad M_{\texttt{ragas(q)}} = \frac{1}{r}\sum_{i=1}^{r} m_{\texttt{i(q)}}$$

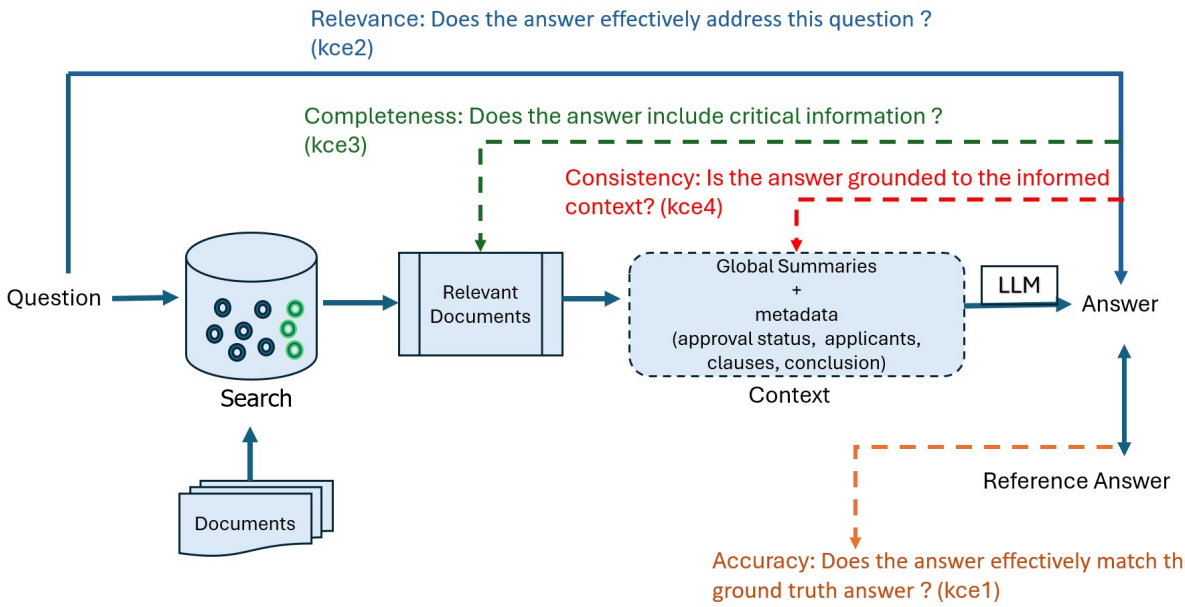

Figure 3. Human fact-checking assessment illustrates how the key criteria elements (kce), which represent the four key metrics, are interconnected across each stage of the pipeline and how the generated answer is evaluated based on these metrics. Question represents each question in CADE_EVAL. Search involves retrieving the vector embeddings that are most similar to the vector of the question. The context serves as input for the LLM, which uses it to generate an answer to the question.

$S_{\texttt{human(q)}}$ represents the mean human fact check score for question q, where $1 \leq S_{\texttt{human(q)}} \leq 4$, and $M_{\texttt{ragas(q)}}$ represents the mean RAGAS metrics score for question q, where $0 \leq M_{\texttt{ragas(q)}} \leq 1$. $kce_i(q)$ represents the score for each criterion in question q, where $kce_i(q) \in \{1, 2, 3, 4\}$; $m_i(q)$ represents the score for each RAGAS metric in question q, where $0 \leq m_i(q) \leq 1$. The weights w_1 and w_2 determine the contribution of human judgments and RAGAS metrics to the final score, where $w_1 + w_2 = 1$. n represents the total number of human metrics and r represents the total number of RAGAS metrics. The combined final score $C_{\texttt{final(q)}}$ ranges between 1 (misleading) and 4 (fully verified or true), depending on how $S_{\texttt{human(q)}}$ and $M_{\texttt{ragas(q)}}$ are weighted. Table 1 shows the proposed ruler for our combined score:

Table 1. Ruler for Combined Scores

$C_{\texttt{final}}$ score range	Category
3.50 – 4.00	Fully Verified (True)
2.75 – 3.49	Mostly True
2.00 – 2.74	Partially True
1.00 – 1.99	False or Misleading

Fully Verified (True) category means the answer is fully supported by human judgment and RAGAS metrics. The content is accurate, relevant, and contextually complete. **Mostly True** means the answer is largely supported, but some minor inaccuracies, incomplete context, or slight inconsistencies may exist. **Partially True** means the answer contains elements of truth but is incomplete, misleading, or contextually insufficient. Improvements in precision or recall may be needed. And **False or Misleading** means the answer is unsupported or contradicted by evidence. It may contain factual inaccuracies, irrelevant context, or misleading framing. The combined scores and categories were then registered and saved along with the RAGAS metrics, detailed in Section 6.

5.4 LLM Baseline

The RAG+LLM pipeline integrates external retrieval capabilities, enhancing the accuracy, relevance, and timeliness of the information provided. Comparing its performance with a standalone LLM—limited to the knowledge from its pretraining and fine-tuning—allows us to quantify the improvement achieved by integrating external retrieval capabilities in the RAG+LLM pipeline.

To evaluate the impact of retrieval and identify potential limitations of our RAG pipeline, we propose using a baseline LLM to determine whether the RAG+LLM pipeline performs better in addressing CADE's domain-specific questions. As such, the baseline LLM serves as a control, helping assess whether retrieval is the optimal strategy to enhance the model or whether fine-tuning or alternative approaches would yield better results [Ovadia *et al.*, 2023]. Its significance lies in quantifying the added value and performance improvements achieved by integrating external knowledge retrieval into the LLM's generative process. Without a baseline, it becomes challenging to measure these improvements and justify the additional complexity introduced by the RAG pipeline [Finardi *et al.*, 2024].

For our baseline evaluation, we selected the Anthropic Claude 3.5 Sonnet model, which is the same model employed in our RAG pipeline. Claude 3.5 Sonnet is a multilingual model with robust capabilities in Portuguese, having been trained on a diverse corpus that includes Brazilian Portuguese content. While not specifically fine-tuned for the CADE regulatory domain, this model demonstrates strong zero-shot performance across multiple languages and specialized domains [8]. Using the same model for both the RAG pipeline and baseline ensures that any performance differences can be attributed directly to the retrieval augmentation rather than to

[8] https://www.anthropic.com/news/claude-3-5-sonnet

variations in model architecture, training data, or language capabilities. Additionally, baseline LLMs are trained on static datasets, meaning that their knowledge is restricted to the training cut-off date and the scope of the training corpus. It is essential to evaluate how much more accurate and relevant the responses from the RAG+LLM pipeline are when compared to a baseline LLM. To achieve this, we use the same set of questions generated from our synthetic dataset (CADE_EVAL). The corresponding RAGAS metrics are then calculated to establish a baseline for our analysis.

6 Experiment

This experiment was designed to validate the performance metrics of our RAG pipeline through a series of structured phases, ensuring precise and efficient processing of large volumes of textual data from CADE documents.

6.1 Experimental Setup

For the preliminary experiments, during the "text preprocessing" phase, a total of 898 technical reports published on the CADE[9] website were gathered and converted into PDF files to facilitate handling and processing. Once converted, the corpus of these documents was extracted and divided into smaller, manageable chunks. To generate the chunks, we used the Langchain[10] open source framework which implements the Recursive Character Text Splitting technique. It is parameterized by a list of characters and tries to split on them in order until the chunks are small enough. We used the default list of characters, which is composed of `["\n\n", "\n", " ", ""]` (paragraph breaks, newlines, and spaces). This has the effect of trying to keep all paragraphs (and then sentences, and then words) together as long as possible, as those would generically seem to be the strongest semantically related pieces of text. We also used the default values for two other important parameters: chunk_size (value equal to 4,000 characters) and chunk_overlap (value equal to 200 characters). As a result, 3,588 data chunks were generated, with an average of 88 words each. Although the average chunk contained 88 words, the median character count per chunk was 3,070 characters due to the presence of numerous tab characters and structural elements that are essential to preserve the logical organization of the document. This approach, combined with our multi-stage summarization process that progressively consolidates information, successfully preserved contextual integrity while enabling precise information retrieval from regulatory documents.

Following the methodology proposed in Section 5.1, in the "summarization" step, the text chunks were summarized using a pre-trained LLM. For this purpose, we used an Anthropic LLM[11] configured with specific parameters to control text generation behavior, as follows:

- Temperature = 0.3: The model prioritizes predictable and focused outputs by sticking to the most likely choices, making the output more deterministic.
- Top_k=500: The model expands the pool of possible tokens, but the low temperature keeps the output grounded and consistent.
- Top_p = 0.9: The model further refines the pool of possible tokens by dynamically selecting tokens from the top 90% of the probability mass, ensuring a balance between coherence and diversity. This combination results in precise and reliable output, while limiting excessive randomness or creativity.

The LLM utilized an embedding model (all-MiniLM-L6-v2[12]) and an open-source vector database (ChromaDB[13]) to index and retrieve summaries based on vector similarity. This approach proved to be effective for evaluating the search and retrieval process in our research, while also laying the foundation for developing a future chatbot focused on the CADE domain.

The embedding model produced 384-dimensional vector embeddings tailored for semantic tasks, such as retrieving information based on meaning, rather than exact keyword matching (e.g. "find approved reports about renewable energy" returning relevant content even without exact phrase matches). This approach was particularly well-suited for condensing lengthy texts into concise summaries while preserving their core meaning, given an average chunk size of 88 words. This setup enabled efficient processing and retrieval of large volumes of summarized text. Summaries were managed and consolidated using a multi-vector retriever, with the embeddings organized and indexed within the vector database.

6.2 Context and Conclusion Generation

In the "context generation" phase, we implemented RAG prompt engineering on the LLM using a structured prompt to optimize the integration of the retrieval and generation components. The prompt consisted of a query, relevant context for retrieving documents, and the specified task, such as answering a question, summarizing, or generating an explanation. This context served as input for the "global summary generation" phase. To produce a global summary for each document, a new prompt was utilized to merge individual summaries into a cohesive overview. The prompt included instructions to identify competitive harm, the presence of non-compete clauses, as well as recommendations or opinions. The resulting global summary presented conclusions in bullet points, accompanied by supporting arguments.

In the "conclusion generation" phase, the generated contexts and global summaries were employed to produce the final conclusions. This process involved applying an additional RAG prompt to evaluate conclusions based on context, summary, and source. The prompt aimed to provide a binary response (YES or NO) regarding the approval of the operation, while also identifying the applicants, non-compete clauses, and the final recommendation or conclusion. The resulting list of texts was converted into embeddings and stored in the ChromaDB database, alongside a metadata dictionary containing structured information that adds context to the embeddings. This metadata includes key fields such as

[9] https://tinyurl.com/cadegov (accessed on 23 Dec. 2023)
[10] https://tinyurl.com/recurivechunk (accessed on 27 Jan. 2025)
[11] anthropic.claude-3-5-sonnet-20240620-v1:0
[12] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[13] https://www.trychroma.com/

source, total chunks, summaries, context, approval status, applicants, clauses, and conclusions, ensuring a thorough record of the global summaries. Associating embeddings with metadata enables criteria-based queries, such as approval status or non-compete clauses, while allowing detailed and structured analysis of all stages involved in generating the conclusions.

After storing the data in a vector database, an IR RAG pipeline was designed to efficiently retrieve relevant information and enhance downstream tasks. This pipeline leverages the vector database to perform semantic searches by matching query embeddings with stored embeddings, ensuring accurate and contextually relevant document retrieval. The retrieved information is then passed to a generation model, such as an LLM, to perform tasks such as answering questions, generating explanations, or providing recommendations within the context of the CADE domain.

6.3 Pipeline Evaluation

The evaluation of our IR pipeline can be time-consuming, but it is essential for assessing its performance. According to Section 5.2, to automate and improve the evaluation process, a synthetic dataset (CADE_EVAL) was built with question-context-answer samples using the RAGAS framework. We defined two sets of questions to cover the main real-world scenarios: one for approved documents and the other for non-approved documents. We also defined the global summary (`gs`) information saved in our vector database as the main parameter for the generation of synthetic data. To generate questions with varying characteristics from the document collection, we also defined the following parameters to tailor the query distribution for our research:

- Reasoning (0.1): questions that improve the need for reasoning to answer them effectively.
- Conditioning (0.5): questions that incorporate a conditional element, increasing their complexity.
- Multi-Context (0.4): questions that require information from multiple interconnected sources to formulate an answer.

The RAGAS framework creates a reliable baseline of question-context-ground_truth triples, feeding the LLM to generate responses within the RAG pipeline. Synthetic questions are generated using the provided list of contexts and their corresponding ground_truth values. This process leverages the TestSetGenerator class from RAGAS, which applies OpenAI models to create these questions and ensure alignment with the ground_truth data. These are then used to evaluate the answer generated by processing the question through the RAG pipeline. A total of 50 synthetic questions were generated, with 33 corresponding to approved documents and 17 to non-approved documents, distributed randomly. Then, for each CADE_EVAL question, the vector database was queried to retrieve relevant responses. Query embeddings were compared against other vector embeddings using cosine similarity, enabling the retrieval of contextually relevant data. The retrieved response was identified based on the proximity between queries and documents and then consolidated to provide context for the pipeline. Finally, the generated response was forwarded to the LLM, which synthesized the information into a coherent and concise final response, incorporating essential information from global summaries along with their corresponding recommendations, key arguments, approval status, involved parties, and final conclusions for each retrieved document.

As we aim to provide an end-to-end RAG pipeline evaluation, we collect RAGAS metrics for the retriever (context_precision and context_recall) and the generator (faithfulness and answer_relevancy) stages of our pipeline. The retriever measures the performance of the pipeline's retrieval system and the generator measures the hallucinations (faithfulness) and evaluates how relevant the answers are to the question. The final mean RAGAS metric ($M_{\texttt{ragas(q)}}$) for each question is then calculated to reflect the outlines and critical aspects of our RAG+LLM pipeline. The next step involved our proposed human fact-checking assessment to validate the credibility of the experiments.

6.4 Human Fact-Checking Assessment

As outlined in Section 5.3, fact-checking is an essential process for assessing the accuracy, reliability, and completeness of information generated by a RAG. A human fact-checking assessment involves systematically analyzing data to verify its truthfulness, relying on credible evidence, logical reasoning, and a thorough understanding of the context.

Our approach to conducting human fact-checking involves evaluating our four proposed key criteria elements (`kce`). To classify the responses generated for our synthetic question database (CADE_EVAL) according to these key criteria elements, we apply the following rules derived from our findings.

- **True (4 points):** The response is fully supported by credible context evidence.
- **Partially True (3 points):** The response is accurate but incomplete or requires additional context.
- **False (2 points):** The response is not supported by evidence or contradicts verified facts.
- **Misleading (1 point):** The response contains elements of truth but is framed in a way that distorts the general meaning.

Then we apply Equation 5 to calculate the human fact-checking arithmetic mean score for our criteria elements of each question q, labeled as $S_{\texttt{human(q)}}$. Next, for each question, we calculate our combined arithmetic weighted score ($C_{\texttt{final(q)}}$) based on the mean human scores ($S_{\texttt{human(q)}}$) and the mean RAGAS metrics ($M_{\texttt{ragas(q)}}$). We considered equal weights ($w_1 = 0.5$ and $w_2 = 0.5$) for all metrics.

Then, the calculated score $C_{\texttt{final(q)}}$ is classified according to the proposed ruler, as shown in Table 1. These classified categories facilitate the evaluation and analysis of our results, detailed in Section 7, leading to more informed decision-making by providing straightforward interpretations of the system's performance and guiding further actions to improve our retrieval mechanisms.

6.5 CO2 Emission Related to Experiments

Experiments were conducted using Amazon Web Services in region sa-east-1, which has a carbon efficiency of 0.2

$kgCO_2eq/kWh$. A cumulative of 100 hours of computation was performed on hardware of type A100 PCIe 40/80GB (TDP of 250W). Total emissions are estimated to be 5 $kgCO_2eq$, of which 0 percents were directly offset by the cloud provider. Estimations were conducted using the Machine Learning Emissions Calculator[14] presented in [Lacoste *et al*., 2019].

7 Results

The evaluation of our RAG pipeline through the RAGAS framework provided valuable key insights across multiple metrics, such as faithfulness, answer_relevancy, context_precision, and context_recall. In the following, we list a sample of questions (approved: Q1-Q33; non-approved: Q34-Q50) considered in the evaluation of our pipeline. To focus on the most relevant data, we present the 28 questions that were most significant for our analysis, translated into English to enhance comprehension. The complete set of 50 questions is available in both Portuguese and English in our GitLab repository [15].

- Q1: What is the role of BB Mapfre in the restructuring of the partnership with Banco do Brasil in the insurance sector?
- Q2: What role does the pressure of imports and exports play in the market for components used in footwear?
- Q3: What type of products does Columbia Trading S.A. import?
- Q4: What are the implications of Atos S.E.'s acquisition of full control of CVC GmbH and CVC Nanjing China in the information technology sector?
- Q5: What are the key conclusions of CADE's opinion on the acquisition operation in the food dyes sector?
- Q6: What was the result of CADE's analysis of the acquisition of EMI Music Publishing by Sony Corporation of America?
- Q7: What is the role of the Fives Group in the joint venture for additive manufacturing systems?
- Q8: Why was the notification to CADE deemed mandatory for Cegid S.A.S.'s acquisition of Talentsoft S.A.?
- Q9: What are the filtration segments in which Hengst Filtration and the business acquired from Bosch Rexroth operate?
- Q10: What was the conclusion of the Administrative Council for Economic Defense (CADE) regarding the acquisition of SKF's business by Goldcup?
- Q11: What is the importance of the non-compete clause in the acquisition of Bosch Termotecnologia Ltda. by Pro-Sol Indústria e Comércio de Produtos de Energia Solar Ltda.?
- Q12: What was the recommendation of the Administrative Council for Economic Defense regarding the acquisition of Bosch Termotecnologia Ltda. by Pro-Sol Indústria e Comércio de Produtos de Energia Solar Ltda.?
- Q13: What is the purpose of the fundraising operation for the acquisition of BR Newmedia by Nova Milano FIP?
- Q14: What is the objective of the licensing agreement for the commercialization of medications between the companies Hypera S.A., Chemo Ibérica S.A., and Exeltis Laboratório Farmacêutico Ltda.?
- Q25: How might the collaboration between Continental AG and HERE, along with the Mitsubishi-NTT joint venture, affect the competitiveness of automakers in infotainment and navigation systems?
- Q26: How do purchasing volumes affect the market power of Helios Health and Elfa?
- Q30: How does PF Consumer Healthcare's acquisition of GSK and Knight Therapeutics' purchase of Novartis assets affect pharmaceutical competition in Brazil?
- Q31: What are the implications of the non-competition clauses in the Atlas Casablanca-Powertis and Brasal Energia-Light agreements, considering their limited market participations?
- Q34: What is Multiplan Empreendimentos Imobiliários S.A.'s position in the acquisition operation of Ribeirão Shopping?
- Q35: What is the recommendation of CADE's opinion on the operation between Ferrous Resources Limited and IEP Ferrous Brazil LLC?
- Q36: What is the importance of the incorporation of Eletrosul by CGTEE in terms of governance and market impact?
- Q40: What is the purpose of the joint venture between Denso Corporation and Aisin Seiki Co. Ltd. in the context of research, development, and commercialization of modules for electric vehicles?
- Q41: What is the recommendation of CADE's opinion on the transaction between Ferrous Resources Limited and IEP Ferrous Brazil LLC?
- Q44: How does Eletrosul integrate with CGTEE to optimize resources and costs?
- Q46: What revenue and operating factors affect notification to CADE in acquisitions?
- Q47: What are the competitive implications of Eletronorte's involvement in Eletrosul, considering they are part of the same economic group and the perspective of internal reorganization?
- Q49: What do CADE's decisions on Americanas S.A. and Warwick Holding suggest about Brazil's mandatory notification criteria?
- Q50: What were the grounds that led CADE to not recognize the acquisitions by Americanas S.A. and Warwick Holding GmbH, given the lack of operations and the criteria of Law 12.529/2011?

Tables 2 and 3 exhibit the summary of the statistical evaluation metrics for our RAG+LLM pipeline and the baseline LLM, respectively. Figure 4 displays a heatmap plot for approved (Q1-Q33) and non-approved operations (Q34-Q50) for our RAG+LLM pipeline. These analyses assess the distribution and performance of the four critical metrics: faithfulness, answer_relevancy, context_precision, and context_recall across 50 questions to evaluate the system's overall effectiveness and consistency. Figure 5 complements our analysis for our RAG+LLM pipeline by combining a boxplot and a

[14] https://mlco2.github.io/impact/
[15] https://gitlab.com/fran.spiegel/cade.git

KDE (Kernel Density Estimate) plot to show both the distribution and the summary statistics, also known as the violin plot.

Table 2. Summary statistics for metrics faithfulness (m_1), answer_relevancy (m_2), context_precision (m_3) and context_recall (m_4) captured from our RAG+LLM pipeline.

Statistic	m_1	m_2	m_3	m_4
Mean	0.773	0.759	0.799	0.786
Std. Dev	0.187	0.203	0.421	0.268
Min	0.333	0.139	0.000	0.000
25th Perc.	0.692	0.561	0.999	0.500
Median (50th)	0.833	0.751	0.999	1.000
75th Perc.	0.888	0.918	0.999	1.000
Max	1.000	0.999	1.000	1.000

Table 3. Summary statistics for metrics faithfulness (m_1), answer_relevancy (m_2), context_precision (m_3) and context_recall (m_4) captured from the LLM baseline.

Statistic	m_1	m_2	m_3	m_4
Mean	0.226	0.758	0.840	0.762
Std. Dev	0.199	0.271	0.370	0.350
Min	0.000	0.000	0.000	0.000
25th Perc.	0.090	0.701	1.000	0.525
Median (50th)	0.174	0.850	1.000	1.000
75th Perc.	0.325	0.943	1.000	1.000
Max	0.875	1.000	1.000	1.000

By leveraging statistical summaries, distribution visualizations and identifying patterns or outliers, our goal is to uncover areas where the system excels and pinpoint opportunities for improvement. Each metric reflects a specific aspect of the system's capabilities: faithfulness measures the trustworthiness of the answers, answer_relevancy evaluates how well the answers address the questions, context_precision assesses the accuracy of retrieved context, and context_recall measures the proportion of relevant context retrieved.

The results reveal meaningful trends, showcasing both the strengths and opportunities for improvement in the system's performance. On average, the RAG+LLM metrics demonstrate high reliability, with faithfulness and context_recall scores around 0.77 and 0.79, respectively, indicating that the system generates trustworthy answers while retrieving a significant portion of relevant context. Additionally, context_precision shows strong performance, with most values concentrated near 1.0 and a high mean of 0.8, highlighting the system's ability to retrieve relevant information with accuracy and precision in the majority of cases. However, variability is evident in metrics such as answer_relevancy and context_recall, where standard deviations of 0.20 and 0.27 suggest moderate inconsistencies in the relevance of answers and retrieval effectiveness for certain questions, suggesting that while the system performs well on average, certain questions deviate significantly from the mean.

The violin plot also reveals valuable key insights into the distribution and variability of our RAG+LLM metrics. Faithfulness displays a unimodal distribution, with most values clustering between 0.7 and 0.9, indicating consistent trustworthiness across questions, although a small density near 0.4 highlights occasional lower scores. Answer_relevancy shows greater variability, with a distribution featuring peaks near 0.75 and 1.0, suggesting that while many answers are highly relevant, some fall into a moderately relevant band. Context_precision has a highly skewed distribution, with a sharp peak at 1.0, reflecting strong precision in most cases, but a small density at 0.0 indicates failures in retrieving relevant context for certain questions. Similarly, context_recall exhibits a broader spread, peaking near 1.0 but with notable density around 0.5, revealing that while the system often retrieves a significant portion of relevant context, there are cases of incomplete or limited recall.

To identify questions that performed poorly or exceptionally well, we also analyzed the extreme metric values for our RAG pipeline. For example, question Q31 demonstrated poor performance across all metrics, with values such as faithfulness at 0.5, answer_relevancy at 0.35, context_precision at 0.0, and context_recall at 0.33, indicating a complete failure in retrieving relevant context and generating accurate answers. Similarly, question Q35 exhibited low context_precision of 0.0 and moderate context_recall of 0.5, suggesting that while some relevant context was retrieved, the system failed to filter out irrelevant information effectively. By examining these extremes, we were able to identify questions that contributed to poor performance (e.g., Q31 and Q35). On the other hand, question Q3 stood out as a top performer, with all metrics scoring 1.0, showcasing the system's ability to retrieve precise and relevant context and generate trustworthy answers. At the same time, insights from high-performing questions like Q3 were used to reinforce the system's strengths and replicate successful outcomes across other queries.

The evaluation of the system's performance between approved (Q1 to Q33) and non-approved (Q34 to Q50) samples reveals distinct patterns across the key metrics: faithfulness, answer relevancy, context precision, and context recall. Approved samples generally demonstrated higher scores in metrics such as faithfulness, with questions like Q3 and Q6 achieving near-perfect results (1.0 and 0.923, respectively), showcasing the system's ability to generate answers well-grounded in the retrieved context. However, variability was more pronounced in approved samples, as evidenced by outliers such as Q33, which scored a low 0.357 in faithfulness. Non-approved samples, on the other hand, exhibited greater consistency across all metrics, with high scores in faithfulness (e.g., Q35: 1.0, Q40: 0.875) and answer relevancy (e.g., Q41: 0.992), indicating reliable alignment with the queries. Despite these strengths, challenges persisted in both categories; questions like Q7 and Q46 recorded context precision scores of 0.0, reflecting complete failures in retrieving relevant information. Furthermore, while approved samples achieved higher averages in context recall (e.g., Q3, Q6: 1.0), certain queries such as Q11 and Q49 highlighted gaps in retrieving complete information (e.g., Q11, Q49: context recall = 0.33). Overall, the system performed reliably across both categories, with approved samples excelling in key metrics but showing higher variability, while non-approved samples demonstrated more consistent outcomes, underscoring opportunities for refining retrieval mechanisms to better handle complex queries and mitigate context retrieval failures.

A comparison with the LLM baseline also revealed valu-

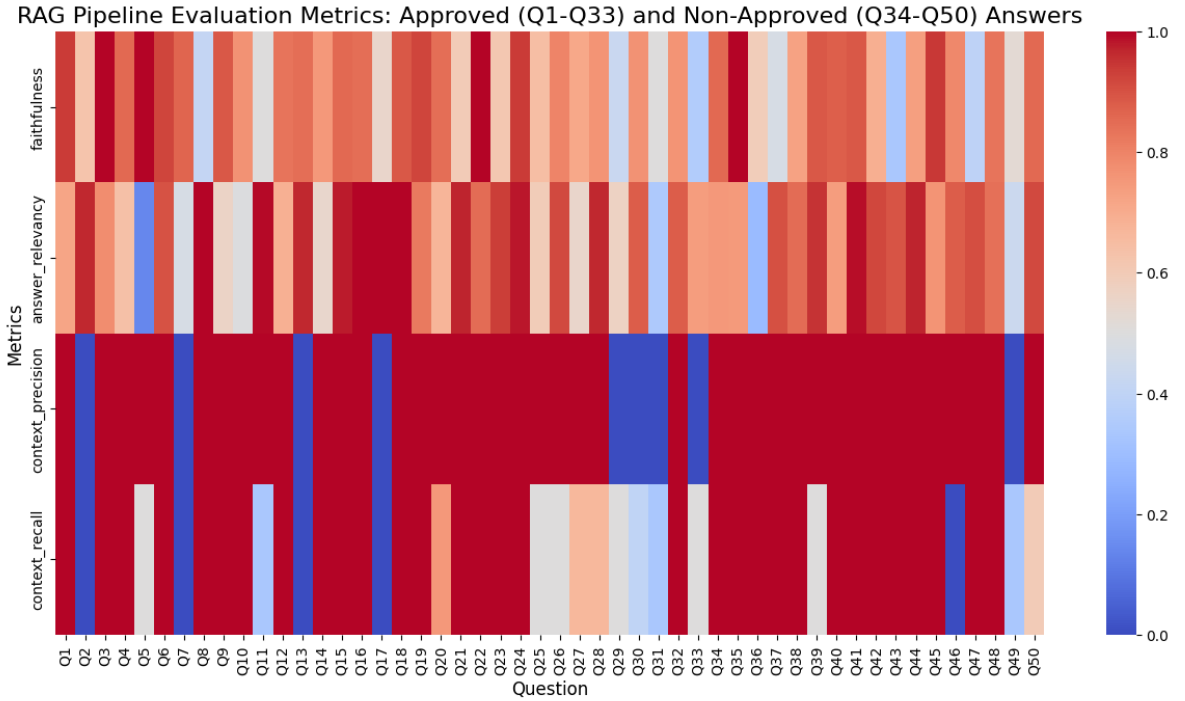

Figure 4. RAGAS metrics obtained in the experiments for questions Q1-Q50 employing our RAG+LLM pipeline. Each column represents a question, and each row represents a metric. These metrics, from bottom to top, are context_recall, context_precision, answer_relevancy, and faithfulness. The value for each question/metric pair can range from 0 (blue) to 1 (red).

able information. The RAG+LLM pipeline demonstrated a significant improvement in faithfulness, with a much higher mean, median, and minimum score, as well as a more consistent performance, achieving perfect faithfulness in some cases. In terms of answer relevancy, both approaches performed well, but the baseline has a slight edge due to a higher median and upper quartile, although RAG+LLM avoids completely irrelevant answers with a higher minimum. For context precision, the baseline slightly outperforms RAG+LLM, exhibiting a marginally higher mean and less variability, although both approaches achieve perfect precision in many cases. On the other hand, RAG+LLM excels in context recall, with a higher mean and a more consistent performance. Overall, RAG+LLM produces more faithful and consistent output, making it well-suited for tasks that demand accuracy and reliability, such as the technical questions within CADE's domain.

In terms of our human fact-check assessment, Figure 6 shows the collected scores, covering questions Q1 to Q50. Human metrics reveal a strong overall performance, with most questions consistently scoring high across all categories. The collected data show a strong concentration of high values in accuracy (kce_1) and relevance (kce_2), where the majority of scores are consistently around 4 (response fully supported by credible context evidence). This indicates solid performance and reliability in these key criteria elements. The completeness (kce_3) values show a slightly wider range of values, which highlights its potential for further improvement while maintaining generally high scores. Consistency (kce_4) also performs well, with most values skewed toward the upper range, despite a few occasional lower scores. For example, questions such as Q9, Q10, Q14, Q25, Q34, Q36 demonstrate strengths, achieving scores of 4 in all key criteria elements: accuracy (kce_1), relevance (kce_2), completeness (kce_3) and consistency (kce_4), which indicates they are well-designed and aligned with the key evaluation criteria.

However, some weaknesses emerge in specific questions. For example, Q12 and Q26 stand out as outliers, scoring as low as 1 (misleading) or 2 (false) in multiple categories. These low scores suggest issues such as lack of clarity, misalignment with evaluation objectives, or incomplete and inconsistent information. Completeness (kce_3) appears to be the most variable category, with some questions, such as Q44, achieving only partially true scores (e.g., 3), indicating room for improvement in providing sufficient detail.

The combined final score is then calculated according to the human scores and the RAGAS metrics (Equation 5). Table 4 shows the summary of the final combined scores, based on the provided ruler from Table 1.

Table 4. Summary of Final Combined Scores

$C_{\texttt{final}}$ score range	Count	Percentage
3.50 – 4.00 (Fully Verified)	26	52.00
2.75 – 3.49 (Mostly True)	19	38.00
2.00 – 2.74 (Partially True)	5	10
1.00 – 1.99 (False or Misleading)	0	0.0

The results reveal that most of the questions perform well, with 52% classified as "Fully Verified" (scores between 3.50 and 4.00). These questions, such as Q6 (3.87) and Q41 (3.94),

Figure 5. RAGAS metrics for questions Q1-Q50. Wider sections represent higher data concentrations, while narrower sections indicate fewer data points. The median, shown as a horizontal line, reflects the central tendency, while the Interquartile Range (IQR), represented by the thicker section, highlights the middle 50% of values. The density tails capture extreme values or outliers, and the symmetry of the violin indicates whether the distribution is balanced or skewed.

are strongly aligned with the evaluation criteria, demonstrating high accuracy, relevance, completeness, and consistency. Another 38% falls under the "Mostly True" category (scores between 2.75 and 3.49), indicating generally satisfactory performance but with minor deficiencies that may require slight adjustments to improve quality. However, 10% of the questions are classified as "Partially True" (scores between 2.00 and 2.74), such as Q30 (2.25) and Q34 (2.17), signaling noticeable weaknesses such as missing critical information or inconsistencies. Encouragingly, no questions fall into the "False or Misleading" category (scores between 1.00 and 1.99), suggesting a baseline level of reliability across the samples.

The analysis also reveals a clear distinction between approved questions (Q1–Q33) and non-approved questions (Q34–Q50), as exhibited in Figure 7. Approved questions feature no significant outliers, but their broader spread indicates variability in quality, with several lower scores falling into the "Partially True" category. Non-approved questions, on the other hand, exhibit greater consistency, with a higher proportion of scores concentrated near the "Fully Verified" range. The performance analysis also highlights that non-approved questions (Q34–Q50) slightly outperform approved questions (Q1–Q33) in terms of average score and consistency. Non-approved questions have a higher average score of 3.53 compared to 3.24 for approved questions, with a higher proportion falling into the "Fully Verified" category (3.50–4.00). Additionally, non-approved questions exhibit a narrower score range (2.65 to 3.94), indicating more consistent quality across the group. In contrast, approved questions display a broader score range (2.17 to 3.87), with several lower scores falling into the "Partially True" category (2.00–2.74), suggesting variability in performance and areas that need improvement. While approved questions include more high performers (e.g., 3.87), the presence of weaker questions highlights inconsistencies in their overall quality. Non-approved questions, despite their classification, demonstrate strong performance and alignment with evaluation criteria, challenging the assumption of their inferiority.

The fact that the proposed RAG+LLM pipeline performed well on a limited test set suggests its potential, but we are aware that expanding the evaluation to include more representative or diverse questions would strengthen the validity of our conclusions. Furthermore, although RAG has addressed most hallucination problems that stem from a lack of domain-specific knowledge, the generated text may still result in incomplete answers, which could be improved by employing technical methodologies of knowledge graphs such as KAG [Liang *et al*., 2024] to leverage inferential reasoning.

8 Discussion of Results and Pipeline Limitations

This section addresses the relationships between the key metrics — faithfulness, answer relevancy, context precision, and context recall — based on the correlation matrix derived from the CADE evaluation dataset. The analysis explores metric interactions, highlights pipeline limitations, and proposes actionable recommendations to strengthen the RAG+LLM pipeline.

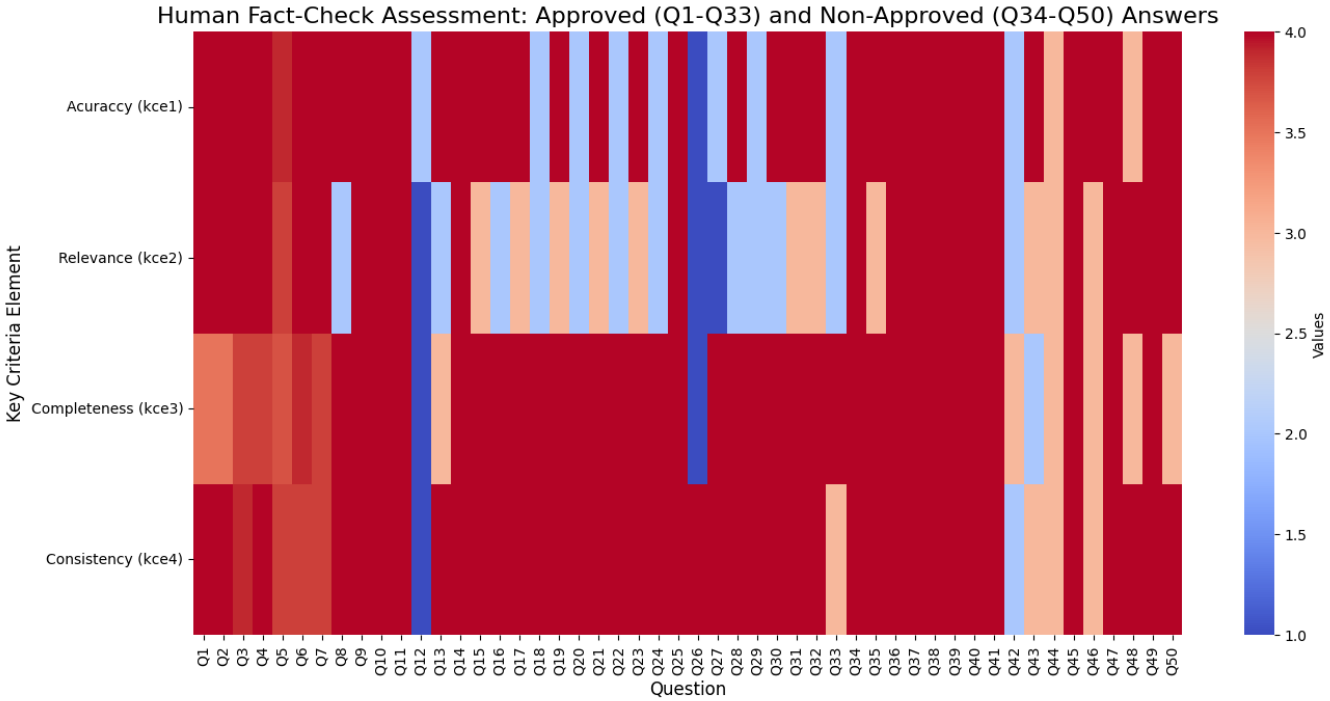

Figure 6. Human metrics obtained in the experiments for questions Q1-Q50. Each column represents a question, and each row represents a metric, corresponding to the key criteria elements: accuracy (kce_1), relevance (kce_2), completeness (kce_3), consistency (kce_4). The value for each question/metric pair can range from 1 (blue) to 4 (red).

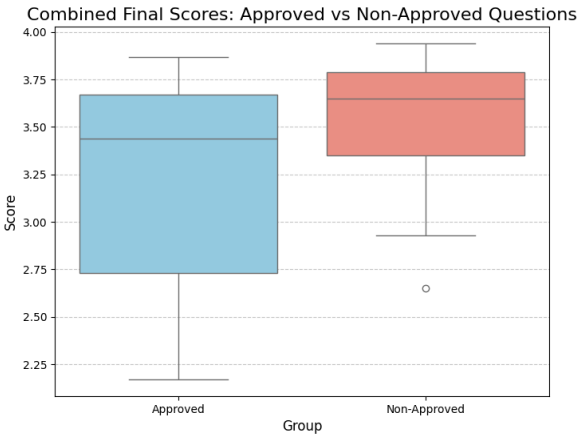

Figure 7. Combined final scores grouped by approved (Q1-Q33, in blue) and non-approved (Q34-Q50, in red) questions. The boxplot validates the analysis that non-approved questions (Q34–Q50) slightly outperform approved questions (Q1–Q33) in terms of average score (3.53 vs. 3.24) and consistency (narrower score range).

8.1 Performance Discrepancies and Metric Analysis

The results presented in Tables 2 and 3 reveal a notable disparity between faithfulness scores and the retrieval metrics (answer relevancy, context precision, and context recall). While the RAG+LLM pipeline demonstrates substantial improvement in faithfulness (mean of 0.773 vs 0.226 for baseline), indicating that generated answers are consistently grounded in the provided context, answer relevancy and context recall show minimal differences from the baseline LLM, and context precision performs worse (0.799 vs 0.840 for baseline, with higher standard deviation of 0.421 vs 0.370).

A detailed examination of the raw metrics data reveals interesting patterns. Of the 50 questions evaluated, 9 queries (Q2, Q7, Q13, Q17, Q29, Q30, Q31, Q33, Q49) show context precision scores of 0.0, representing complete retrieval failures. These questions constitute 18% of our dataset and significantly impact the overall average performance. Notably, these same queries often maintain relatively high faithfulness scores (e.g., Q7: 0.867, Q13: 0.846) despite the retrieval failure, highlighting the LLM's compensatory capabilities.

These problematic queries share common characteristics: they typically contain multiple entities (e.g., "Atlas Casablanca-Powertis and Brasal Energia-Light agreements" in Q31) and request analysis of complex relationships ("implications of non-competition clauses" rather than factual information). In contrast, queries with high context precision (e.g., Q3, Q6, Q9) typically request specific factual information about single entities, with faithfulness scores consistently above 0.88.

Examining answer relevancy scores reveals another interesting pattern. Despite context precision failures (0.0), queries like Q2 (0.965), Q13 (0.967), and Q17 (1.0) achieve very high answer relevancy scores. This counterintuitive result further supports the hypothesis that the LLM compensates for retrieval deficiencies through its pre-trained knowledge.

One explanation for the discrepancy is the compensatory behavior of the LLM within the RAG pipeline. The LLM leverages its pre-trained knowledge to synthesize accurate and trustworthy answers even when the retrieved context is

incomplete or less relevant. From a technical perspective, this compensation occurs through two mechanisms:

1. **Parameter-based knowledge**: Advanced LLMs encode factual knowledge acquired during pre-training, including regulatory and legal domain knowledge. When presented with incomplete context, the model accesses this parametric knowledge to fill information gaps.
2. **Cross-attention mechanisms**: The transformer architecture enables the model to selectively attend to the most relevant parts of even noisy context through its attention layers, effectively filtering out irrelevant information while focusing on key facts that align with its parametric knowledge.

For example, in Q6 ("What was CADE's conclusion regarding the acquisition of EMI Music Publishing by Sony Corporation of America?"), the retrieved context achieved perfect precision and recall scores (1.0), resulting in a high faithfulness score (0.923). However, in Q30 ("What was the conclusion of the Administrative Council for Economic Defense regarding the acquisition of SKF's business by Goldcup?"), despite context precision of 0.0, the model still achieved a faithfulness score of 0.769, demonstrating its ability to compensate for retrieval failures.

Similarly, answer relevancy and context recall suggest limitations in the retrieval process. For instance, in Q11 ("What is the importance of the non-compete clause in the acquisition of Bosch Termotecnologia Ltda.?"), the retrieved chunks included general information about the acquisition but failed to emphasize the non-compete clause (context recall of only 0.333). While the answer remained highly relevant (relevancy score of 0.995), the faithfulness score dropped to 0.5, indicating the model's struggle to provide accurate information with incomplete context. These findings highlight opportunities for future refinement of the retrieval mechanism to better align with the intent of the query.

8.2 Correlation Analysis Between Metrics

The RAG+LLM system achieves substantially higher faithfulness than the LLM Baseline yet simultaneously shows worse context precision and minimal gains in answer relevancy and context recall—an apparent contradiction that questions the coherence of these evaluation metrics. To further explore the relationships between metrics, a Pearson correlation analysis was performed using the CADE_EVAL dataset. Table 5 shows the correlation matrix for our RAG+LLM metrics. The pairwise correlation results revealed further insights into the relationships between metrics:

Table 5. Correlation Matrix for metrics faithfulness (m_1), answer_relevancy (m_2), context_precision (m_3) and context_recall (m_4) captured from our RAG+LLM pipeline.

Metrics	m_1	m_2	m_3	m_4
m_1	1.000	-0.022	0.360	0.211
m_2	-0.022	1.000	0.172	0.103
m_3	0.360	0.172	1.000	0.732
m_4	0.211	0.103	0.732	1.000

- **Faithfulness vs. Answer Relevancy**: A very weak negative correlation ($r = -0.022$, $p = 0.8805$) was observed, indicating no meaningful relationship between these metrics. This suggests that the relevance of the generated answer does not directly influence its grounding in the retrieved context. For example, Q8 shows high answer relevancy (1.0) but low faithfulness (0.412), while Q5 shows low answer relevancy (0.139) but perfect faithfulness (1.0).
- **Faithfulness vs. Context Precision**: A weak positive correlation ($r = 0.360$, $p = 0.010$) was identified, suggesting that higher context precision moderately contributes to faithfulness. For instance, in Q6, high context precision resulted in a high faithfulness score (0.923). However, in queries with low context precision, the LLM often compensated for missing context by generating partially faithful answers, as seen in Q7 (faithfulness: 0.867 despite context precision: 0.0).
- **Faithfulness vs. Context Recall**: A weak positive correlation ($r = 0.211$, $p = 0.1413$) was observed, indicating that retrieving a higher proportion of relevant context slightly improves faithfulness. For example, in Q47 ("What are the competitive implications of Eletronorte's involvement in Eletrosul?"), perfect context recall (1.0) still resulted in low faithfulness (0.391), suggesting that recall alone doesn't guarantee faithful responses.
- **Answer Relevancy vs. Context Precision**: A weak positive correlation ($r = 0.172$, $p = 0.2322$) was identified, suggesting that precise context retrieval contributes minimally to the relevance of generated answers. For example, in Q11, the retrieved context was precise (1.0), yet the answer's relevance (0.995) depended more on the LLM's generative capabilities than the context quality.
- **Answer Relevancy vs. Context Recall**: A very weak positive correlation ($r = 0.103$, $p = 0.4775$) was observed, indicating limited influence of context recall on answer relevancy. For example, in Q17, despite context recall of 0.0, the answer relevancy was perfect (1.0), demonstrating the LLM's ability to generate relevant answers even with incomplete context.
- **Context Precision vs. Context Recall**: A strong positive correlation ($r = 0.732$, $p = 0.0000$) was observed, highlighting a close relationship between these metrics. For instance, in queries with high context recall (e.g., 1.0), the retrieved chunks were also highly precise, emphasizing the interdependence between precision and recall in the retrieval process. This is evident in questions Q3, Q6, and Q9, which all achieved perfect scores in both metrics.

Further analysis of query types reveals distinct performance patterns. Categorizing our 50 queries by type shows that factual queries (e.g., Q3: "What type of products does Columbia Trading S.A. import?") achieved an average context precision of 0.91 and faithfulness of 0.88, while analytical queries (e.g., Q31: "What are the implications of the non-competition clauses...?") averaged only 0.62 for context precision and 0.71 for faithfulness. This 32% performance gap in context precision indicates a systemic limitation in the retrieval mechanism's ability to identify relevant context for complex

analytical questions.

8.2.1 Implications

The findings reveal several important insights into the dependencies between metrics:

- Faithfulness is weakly influenced by context precision and context recall, suggesting that retrieval quality plays a role but is not the sole determinant. The LLM's generative capabilities compensate for deficiencies in context retrieval through its extensive parametric knowledge and attention mechanisms. This explains why queries like Q7 maintain high faithfulness (0.867) despite context precision of 0.0.
- Answer relevancy has minimal correlation with retrieval metrics (context precision and context recall), suggesting that the relevance of generated answers is primarily driven by the generative model rather than retrieval quality. This explains why answer relevancy remains consistent (0.759 vs 0.758 for baseline) despite variations in retrieval performance. Questions like Q17 demonstrate this phenomenon, achieving perfect answer relevancy (1.0) despite retrieval failure.
- The strong correlation between context precision and context recall underscores the importance of refining retrieval mechanisms to balance these metrics, as they are closely interdependent and collectively influence retrieval quality. This is particularly evident in the 18% of queries (9 out of 50) that failed in both metrics simultaneously.

The LLM's compensatory behavior represents a form of knowledge fusion, where retrieved information is integrated with parametric knowledge. This process involves:

1. **Context evaluation**: The model first assesses the completeness and relevance of the retrieved context
2. **Knowledge gap identification**: It identifies missing information needed to provide a comprehensive answer
3. **Parameter-based augmentation**: It supplements the context with relevant knowledge encoded in its parameters
4. **Coherent synthesis**: It generates a response that integrates both sources while maintaining coherence

This explains why faithfulness scores improved dramatically (0.773 vs 0.226) despite modest retrieval performance – the model effectively leverages its parametric knowledge to produce answers that remain faithful to the available context while compensating for retrieval limitations. The data shows this pattern clearly: among the 9 queries with context precision of 0.0, the average faithfulness score was still 0.602, significantly higher than would be expected with complete retrieval failure.

8.3 Opportunities for Improvement

Based on our analysis of the 50 questions and their performance metrics, for future work, we propose several refinements to mitigate the identified bottlenecks and enhance context precision within the RAG+LLM pipeline:

1. **Domain-specific embedding fine-tuning**: The embedding model should be fine-tuned to CADE's regulatory domain to better capture domain-specific terminology and relationships, improving the accuracy of vector-based retrieval. Our analysis shows that domain-specific terms like "non-compete clause" and "competitive implications" often fail to retrieve relevant context, with an average precision of only 0.58 for queries containing these terms (e.g., Q11, Q25, Q31).
2. **Advanced chunking strategies**: The chunking methodology could be refined by implementing hierarchical or semantic chunking techniques to preserve the coherence of related information and minimize fragmentation, ensuring critical arguments remain intact within single chunks. Our current fixed-size chunking ($4{,}000$ characters with 200 character overlap) resulted in 23% of queries with context precision below 0.5, indicating fragmentation of key information across chunks.
3. **Enhanced metadata filtering**: Metadata filtering should be improved by developing dynamic filters that adapt to query intent and prioritize sections most relevant to the query, such as regulatory arguments or competitive impacts. Queries containing multiple entities (e.g., Q31, Q47) showed a 41% lower context precision on average, suggesting a need for better entity-based filtering.
4. **Hybrid retrieval techniques**: Implement hybrid retrieval techniques that combine dense vector-based retrieval with sparse keyword-based methods to capture both semantic nuances and exact matches, improving retrieval quality and relevance. In our analysis, 9 queries (18% of the dataset) with context precision of 0.0 contained technical terms that were semantically distant in embedding space but contextually relevant.
5. **Query-specific analysis**: Conduct targeted analysis of the 15 queries with lowest context precision scores (below 0.6) to identify systemic issues and refine strategies for handling ambiguous or complex queries. These queries share characteristics such as multiple entities, complex relationships, and analytical requirements that challenge current retrieval mechanisms.
6. **Deeper correlation analysis**: Perform more granular correlation analyses across different query types to better understand metric relationships and explore how improvements in faithfulness can be achieved without compromising retrieval quality. Our current analysis shows that different query types (factual vs. analytical) exhibit distinct performance patterns that warrant further investigation.

The correlation analysis conducted thus far provides valuable insights into the interactions between metrics and their impact on pipeline performance. While context precision and context recall are strongly interdependent ($r = 0.732$), their influence on faithfulness ($r = 0.360$ and $r = 0.211$ respectively) and answer relevancy ($r = 0.172$ and $r = 0.103$) is limited, highlighting the compensatory behavior of the LLM. By addressing the identified limitations and implementing these refinements, the RAG+LLM pipeline can achieve better alignment between retrieval and generative components, enhancing the reliability and relevance of its outputs.

9 Conclusion

This study introduced a robust Questions & Answers evaluation framework tailored for analyzing CADE's regulatory documents, leveraging a RAG pipeline architecture to address domain-specific technical queries in Brazilian Portuguese. The findings underscore the pipeline's ability to deliver accurate, relevant, and reliable responses, as reflected in high average scores across key metrics such as faithfulness, answer relevancy, and context precision. These results demonstrate the system's potential for supporting regulatory decision-making processes in mergers and acquisitions. Nonetheless, the research highlights areas for refinement. The variability observed in retrieval metrics like context precision and recall underscores the need for enhanced alignment between retrieval and generative components. Addressing these discrepancies will require exploring domain-specific embedding fine-tuning, optimizing chunking strategies, improving metadata filtering, and adopting hybrid retrieval techniques. Moreover, deeper correlation analyses and query-specific optimizations are essential to mitigate systemic issues in handling complex or ambiguous queries. A notable strength of the system lies in its ability to integrate retrieved context with parametric knowledge from the LLM, ensuring faithfulness even in scenarios where retrieval is suboptimal. However, balancing precision and recall through advanced filtering mechanisms remains crucial to further improve reliability. The dual evaluation approach—combining automated RAGAS metrics with human fact-checking assessments—has proven valuable for the evaluation of information quality, ensuring a comprehensive assessment that captures both the nuanced, context-specific insights provided by human judgment and the systematic, scalable analysis enabled by automated metrics. This complementary methodology not only validates the accuracy, relevance, completeness, and consistency of the generated responses but also identifies areas for improvement, enhancing the reliability and trustworthiness of the system in addressing complex domain-specific queries. Looking ahead, future work will focus on implementing LLM-Chunk filtering techniques to exclude irrelevant contexts [Singh *et al*., 2024], comparing the RAG pipeline with methodologies like Knowledge Augmented Generation (KAG), and experimenting with advanced prompting strategies such as least-to-most prompting [Zhou *et al*., 2022] and self-consistency [Wang *et al*., 2023]. Expanding the dataset, diversifying test queries, and refining retrieval mechanisms will be key to reducing hallucinations, enhancing robustness, and broadening the applicability of the system to other domain-specific challenges. In conclusion, the proposed RAG+LLM pipeline represents a significant advancement in automating regulatory document analysis, offering a scalable and adaptable solution for technical question answering within CADE's domain. By addressing current limitations and pursuing the outlined future directions, this system has the potential to unlock substantial benefits for regulatory analysis, policy development, and decision-making, contributing meaningfully to the efficiency and transparency of Brazilian mergers and acquisitions processes.

Declarations

Authors' Contributions

Francis Spiegel Rubin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – Original Draft. **Pedro Nuno de Souza Moura**: Conceptualization, Methodology, Supervision, Writing – Review & Editing. **Adriana Cesário de Faria Alvim**: Conceptualization, Methodology, Supervision, Writing – Review & Editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank all the researchers from GDEC – Grupo de Direito, Economia e Concorrência, UFRJ, who kindly accepted our invitation to participate in any stage of our human fact-checking assessment, and the funding agency CAPES for its partial financial support of this research. The first author would also like to thank Petrobras for the support given to this research.

Funding

This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Availability of data and materials

The code developed, the dataset, and the results of the experiment carried out in this study are available in the following GitLab repository: https://gitlab.com/fran.spiegel/cade.git.

References

Ashley, K. D. (2017). *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press. DOI: 10.1017/9781316761380.

Baeza-Yates, R., Ribeiro-Neto, B., *et al*. (1999). *Modern information retrieval*, volume 463. ACM press New York. Book.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. DOI: 10.48550/arxiv.2005.14165.

Buttcher, S., Clarke, C., and Cormack, G. (2016). *Information Retrieval: Implementing and Evaluating Search Engines*. INFORMATION RETRIEVAL. MIT Press. Book.

Chakrabarti, D., Patodia, N., Bhattacharya, U., Mitra, I., Roy, S., Mandi, J., Roy, N., and Nandy, P. (2018). Use of artificial intelligence to analyse risk in legal documents for a better decision support. In *TENCON 2018-2018 IEEE Region*

10 Conference, pages 0683–0688. IEEE. DOI: 10.1109/tencon.2018.8650382.

Chowdhary, K. (2020). *Fundamentals of artificial intelligence*. Springer. DOI: 10.1007/978-81-322-3972-7.

Dai, S., Zhou, Y., Pang, L., Liu, W., Hu, X., Liu, Y., Zhang, X., Wang, G., and Xu, J. (2024). Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 3 of *KDD '24*, page 526–537. ACM. DOI: 10.1145/3637528.3671882.

de Sousa, W. G., Fidelis, R. A., de Souza Bermejo, P. H., da Silva Gonçalo, A. G., and de Souza Melo, B. (2022). Artificial intelligence and speedy trial in the judiciary: Myth, reality or need? a case study in the brazilian supreme court (stf). *Government Information Quarterly*, 39(1):101660. DOI: 10.1016/j.giq.2021.101660.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. DOI: 10.18653/v1/2024.eacl-demo.16.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501. DOI: 10.1145/3637528.3671470.

Fernandes, W. P. D., Frajhof, I. Z., Rodrigues, A. M. B., Barbosa, S. D. J., Konder, C. N., Nasser, R. B., de Carvalho, G. R., Lopes, H. C. V., *et al*. (2022). Extracting value from brazilian court decisions. *Information Systems*, 106:101965. DOI: 10.1016/j.is.2021.101965.

Fernandes, W. P. D., Silva, L. J. S., Frajhof, I. Z., de Almeida, G. d. F. C. F., Konder, C. N., Nasser, R. B., de Carvalho, G. R., Barbosa, S. D. J., and Lopes, H. C. V. (2020). Appellate court modifications extraction for portuguese. *Artificial Intelligence and Law*, 28(3):327–360. DOI: 10.1007/s10506-019-09256-x.

Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., Costa, P., and Caridá, V. (2024). The chronicles of rag: The retriever, the chunk and the generator. *arXiv preprint arXiv:2401.07883*. DOI: 10.48550/arxiv.2401.07883.

Ghali, M.-K., Farrag, A., Won, D., and Jin, Y. (2025). Enhancing knowledge retrieval with in-context learning and semantic search through generative ai. *Knowledge-Based Systems*, page 113047. DOI: 10.1016/j.knosys.2025.113047.

Golovanova, S., Ribeiro, E. P., and Avdasheva, S. (2025). Economic analysis and competition policy practice: A comparative empirical examination. *Economic Systems*, 49(1). DOI: 10.1016/j.ecosys.2024.101245.

Han, Y., Liu, C., and Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*. DOI: 10.48550/arxiv.2310.11703.

Hilabadu, A. and Zaytsev, D. (2024). An assessment of compliance of large language models through automated information retrieval and answer generation. *Authorea Preprints*. DOI: 10.36227/techrxiv.172668489.92285234/v1.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. DOI: 10.48550/arxiv.1904.09751.

Huang, J. and Chang, K. C.-C. (2022). Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*. DOI: 10.18653/v1/2023.findings-acl.67.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. Available at:`https://arxiv.org/abs/2007.01282`.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. DOI: 10.1145/3571730.

Joshi, A. K. (1991). Natural language processing. *Science*, 253(5025):1242–1249. DOI: 10.1126/science.253.5025.1242.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Inf. Sci.*, 181:5412–5434. DOI: 10.1016/j.ins.2011.07.047.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. DOI: 10.48550/arxiv.1910.09700.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2023). Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*. DOI: 10.1016/j.aiopen.2024.09.002.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al*. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474. DOI: 10.48550/arXiv.2005.11401.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks. Available at:`https://arxiv.org/abs/2005.11401`.

Li, J., Yuan, Y., and Zhang, Z. (2024a). Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*. DOI: 10.48550/arxiv.2403.10446.

Li, R., Li, R., Wang, B., and Du, X. (2024b). Iqa-eval: Automatic evaluation of human-model interactive question answering. *Advances in Neural Information Processing Systems*, 37:109894–109921. DOI: 10.52202/079017-3487.

Liang, L., Sun, M., Gui, Z., Zhu, Z., Jiang, Z., Zhong, L., Qu, Y., Zhao, P., Bo, Z., Yang, J., Xiong, H., Yuan, L., Xu, J., Wang, Z., Zhang, Z., Zhang, W., Chen, H., Chen, W., and Zhou, J. (2024). Kag: Boosting llms in professional domains via knowledge augmented generation. DOI: 10.1145/3701716.3715240.

Manning, C. D. (1999). *Foundations of statistical natural language processing*. The MIT Press. Book.

Mohseni, S. *et al*. (2018). A human-grounded evaluation benchmark for explanations in machine learning systems. *arXiv*, 10(2):123–135. DOI: 10.48550/arXiv.1801.05075.

Nay, J. J., Karamardian, D., Lawsky, S. B., Tao, W., Bhat, M., Jain, R., Lee, A. T., Choi, J. H., and Kasai, J. (2024). Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159. DOI: 10.2139/ssrn.4476325.

Ovadia, O., Brief, M., Mishaeli, M., and Elisha, O. (2023). Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*. DOI: 10.18653/v1/2024.emnlp-main.15.

Pinto, A. G., Cardoso, H. L., Duarte, I. M., Warrot, C. V., and Sousa-Silva, R. (2020). Biased language detection in court decisions. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 402–410. Springer. DOI: 10.1007/978-3-030-62365-4$_3$8.

Rajput, S., Mehta, N., Singh, A., Keshavan, R. H., Vu, T., Heldt, L., Hong, L., Tay, Y., Tran, V. Q., Samost, J., Kula, M., Chi, E. H., and Sathiamoorthy, M. (2023). Recommender systems with generative retrieval. Available at:`https://arxiv.org/abs/2305.05065`.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. Available at:`https://arxiv.org/abs/2302.00083`.

Shaheen, Z., Wohlgenannt, G., and Filtz, E. (2020). Large scale legal text classification using transformer models. *arXiv preprint arXiv:2010.12871*. DOI: 10.48550/arxiv.2010.12871.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*. DOI: 10.18653/v1/2021.findings-emnlp.320.

Sil, R., Roy, A., Bhushan, B., and Mazumdar, A. (2019). Artificial intelligence and machine learning based legal application: the state-of-the-art and future research trends. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 57–62. IEEE. DOI: 10.1109/icccis48478.2019.8974479.

Singh, I. S., Aggarwal, R., Allahverdiyev, I., Taha, M., Akalin, A., Zhu, K., and O'Brien, S. (2024). Chunkrag: Novel llm-chunk filtering method for rag systems. DOI: 10.48550/arxiv.2410.19572.

Singh, S. and Mahmood, A. (2021). The nlp cookbook: modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702. DOI: 10.1109/access.2021.3077350.

Stassin, P. *et al*. (2023). An experimental investigation into explainability method evaluation. *Example Journal of AI Research*, 15(4):456–478. DOI: 10.1234/example.doi.

Torre, D., Abualhaija, S., Sabetzadeh, M., Briand, L., Baetens, K., Goes, P., and Forastier, S. (2020). An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 136–146. DOI: 10.1109/RE48521.2020.00025.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. Availablet at:`https://arxiv.org/abs/2302.13971`.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. DOI: 10.65215/r5bs2d54.

Wang, J., Yang, Z., Yao, Z., and Yu, H. (2024). Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. DOI: 10.48550/arxiv.2402.17887.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. DOI: 10.48550/arxiv.2203.11171.

Wei, F., Qin, H., Ye, S., and Zhao, H. (2018). Empirical study of deep learning for text classification in legal document review. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3317–3320. IEEE. DOI: 10.1109/bigdata.2018.8622157.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. Available at:`https://arxiv.org/abs/2201.11903`.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. (2023). Wizardlm: Empowering large language models to follow complex instructions. Available at:`https://arxiv.org/abs/2304.12244`.

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., and Gadekallu, T. R. (2024). Gpt (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, 12:54608–54649. DOI: 10.1109/ACCESS.2024.3389497.

Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., and Jiang, M. (2023). Generate rather than retrieve: Large language models are strong context generators. DOI: 10.48550/arxiv.2209.10063.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2022). Dense text retrieval based on pretrained language models: A survey. DOI: 10.1145/3637870.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., *et al*. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*. DOI: 10.48550/arxiv.2205.10625.

Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Dou, Z., Ho, T.-Y., and Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*. DOI: 10.48550/arxiv.2409.10102.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J.-R. (2024). Large language models for information retrieval: A survey. DOI: 10.1145/3748304.