Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation

Isabella V. de Aquino[1], Matheus M. dos Santos[1], Carina F. Dorneles[1], Jônata T. Carvalho[1]

[1]Department of Informatics and Statistics
Federal University of Santa Catarina (UFSC)
P.O. Box 5064 – 88.040-370 – Florianópolis – SC – Brazil
isabella.aquino@grad.ufsc.br, matheus.m.santos@posgrad.ufsc.br
{carina.dorneles,jonata.tyska}@ufsc.br

Abstract. *Extracting information from unstructured data is a challenge that has drawn increasing attention over time due to the exponential growth of stored digital data in modern society. Large Language Models (LLMs) have emerged as powerful tools that benefit from this abundance and have shown remarkable capabilities in Natural Language Processing tasks. Nonetheless, these models still encounter limitations on extraction tasks. Retrieval Augmented Generation (RAG) is a novel approach that combines classic retrieval techniques and LLMs to address some of these limitations. This paper proposes a workflow that allows the assessment of RAG experimental setups, including the multiple possibilities of parameters and LLMs, to extract structured data from Brazilian legal documents. We validated our proposal with experiments using forty legal documents and the extraction of two target variables. The best results obtained with our workflow showed an average extraction accuracy of 90%, significantly outperforming a regular expression strategy, with 58.75% average accuracy. Furthermore, our results show that each extracted variable potentially holds an optimal combination of parameters, highlighting the context-dependency of each extraction and, therefore, the proposed workflow's usefulness.*

1. Introduction

The increasing digitization of judicial and administrative processes worldwide has led to massive production and storage of legal documents. These documents are commonly unstructured, complex and contain crucial information for lawyers, judges, and prosecutors. Extracting this information typically requires extensive human annotation and management in external systems such as relational databases. In line with this scenario, several efforts have been made to handle and process legal documents in various countries, for instance, explored in [Bach and et al. 2019] for extracting references from Vietnamese legal documents, and in [Vianna and et al. 2022] for examining the processing and summarization of Portuguese legal documents.

In particular, the Brazilian public legal sector is an example of an organization dealing with great amounts of documents; almost 200,000[1] public procurement processes were successfully contracted from 2020 to 2023 by the Brazilian Federal Government, in which each one of them requires thorough documentation to formalize every step of the process. As a result, retrieving and extracting specific information, such as legal processes, contract identifiers, and involved municipalities from these numerous complex documents, poses a demanding task if done manually.

Moving forward, information extraction (IE) is an extensively researched subject in legal domains to overcome the presented challenges and has been applied and evaluated through multiple approaches, such as traditional pattern matching

[1]https://portaldatransparencia.gov.br/licitacoes

[Cheng and et al. 2009]. Likewise, the work presented in [Kowsrihawat and et al. 2015] achieves expressive results in extracting variables in legal documents through a proposed framework utilizing regular expressions. Overall, IE in legal domains is a rapidly evolving field with the potential to transform how legal professionals work and automating information extraction can provide valuable insights to legal entities and potentially aid broader analyses to detect and prevent fraud and corruption.

Then, Large Language Models have emerged offering the promise of understanding and generating human-like text at scale and in the legal domain [Katz and et al. 2023]. However, despite their impressive performance and variety of applications, LLMs still face inherent limitations when extracting structured information from unstructured data sources. LLMs knowingly struggle with domain-specific or knowledge-intensive tasks [Kandpal and et al. 2023], have their performance degraded when dealing with relevant information in the middle of long contexts [Liu and et al. 2023] and tend to produce "hallucinations" [Huang and et al. 2023] when searching for information beyond their training data.

In response to these challenges, Retrieval Augmented Generation (RAG) has emerged as a promising approach for enhancing the capabilities of LLMs in information extraction tasks [Gao and et al. 2024]. By combining classic retrieval techniques with LLMs, RAG systems enable the retrieval of relevant information from external sources during text generation, thereby mitigating domain-specific and context window limitations of LLMs and improving the accuracy and coherence of the generated text.

This paper proposes a workflow that leverages LLMs within RAG pipelines to extract structured information from Brazilian legal documents related to fraud in public procurement processes. However, RAG is a data-driven general framework, and its setup can be demanding once several different parameter types are required to be set beforehand. Our objective is to demonstrate the effectiveness of RAG in overcoming the challenges associated with information extraction from complex, domain-specific documents and propose a workflow that evaluates and indicates the best RAG parameter configurations for extracting a given variable. We extracted and evaluated two different variables of interest in forty different Brazilian legal documents utilizing our workflow. The results showed the proposed methodology's effectiveness, which achieved an average accuracy of 90%, outperforming a baseline strategy based on regular expressions, which achieved 58.5%.

2. Related Work

Information extraction (IE) has become a significant explored subject [Doan and et al. 2006], keeping pace with the rapid increase of unstructured data availability in today's data-driven world. IE tasks permeate various aspects of information and its forms of representation and structure, including visual aspects [Sarkhel and et al. 2021], and consider different languages [Zhu and et al. 2012]. It traditionally can be done by applying various approaches, such as the ones focused on annotating [Boisen and et al. 2000] or filtering [Wachsmuth and et al. 2013].

On that matter, Artificial Intelligence and NLP accompany IE advancements and research; [Han and et al. 2023] analyze and evaluate IE using ChatGPT, ranking LLMs encountered limitations, while [Wei and et al. 2024] explores IE systems chatting with ChatGPT in zero-shot settings.

Finally, IE is highly useful in legal applications, which commonly deal with expressive volumes of unstructured information. [Bhattacharya and et al. 2019] automatically identifies rhetorical roles in Indian legal cases. Then, [Pereira and et al. 2024] introduces basic information extraction from Brazilian audit court documents integrating

LLMs in a retrieval-augmented generation workflow.

Given the foregoing, extracting information from legal documents commonly encounters difficulties, such as formatting and structure variability, complicating pattern-matching strategies. As for NLP-driven strategies, sole LLMs are greatly affected by irrelevant and longer context, a big aspect of legal documents. Our work addresses these bottlenecks by enabling contextualization of variables and overcoming the need to feed entire documents into prompts with RAG, highlighting the usefulness of the paper.

3. Method

Our method is structured around a main RAG pipeline for the extraction, executing multiple times iteratively across multiple possible parameter configurations of RAG parameters. This approach facilitates comprehensive comparisons across various parameter sets, potentially identifying an optimal configuration for extracting a given variable. The list below outline these parameters, grouping them by types:

- **Generation:** Parameters related to the generation step of RAG. The following parameter can be tested: Large Language Models (LLMs);
- **Chunking:** Parameters related to the documents chunking strategy. The following parameters can be tested: chunk size, which is the size of each of the split chunks of text; chunk overlap, which is the size of text overlap between adjacent chunks; and splitting strategy, which is usually the text splitter used to execute the chunking;
- **Embeddings:** Parameters related to the embeddings to be generated. The following parameter can be tested: embedding model, used to generate the embeddings from the documents' chunks, e.g. BERT models;
- **Retrieval:** Parameters related to the retrieval step of RAG. The following parameters can be tested: vector database, responsible to store and retrieve the embeddings and Top K value, which is the K-amount of retrieved chunks to serve as context on the extraction.

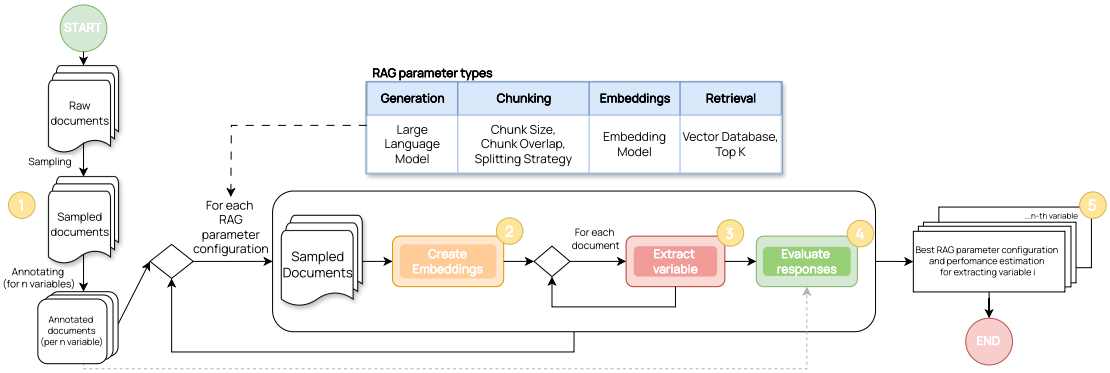


Figure 1. Main workflow overview.

Figure 1 depicts the overview of the proposed workflow, which will cover all combinations of parameter settings possible to extract the same chosen variable and compare the results among each other. A parameter configuration is a unique set of values for each of the RAG parameter types previously described due to iterating through all the available parameter options. Our proposal is based on the following steps:

- Step 1 establishes the beginning of our proposed workflow, initiated by sampling the documents available. We ensure all of them will have an expected value to be extracted for a given query and manually annotate every sampled document with its expected values. This step is the base of our further evaluation assessment step, represented by step number 4.

- Step 2 iterates all possible parameter configuration combinations within the selected options for each parameter type. Step 2 represents the embeddings creation for every sampled PDF document using the current configuration for chunking and embeddings. These embeddings will be used during the extraction step in the workflow.
- Step 3 constitutes the main RAG pipeline. It will retrieve the most relevant embeddings generated in the previous step related to a given query. It will insert them as the context in a prompt template and return direct responses containing or not the answer for the task.
- Step 4 evaluates all the extracted responses by comparing them directly to the foregoing annotated values, labeling as correct the responses that contain the exact expected information for a query.
- Step 5 outputs the best results parameter configurations and performance estimation for each extracted variable.

Lastly, a key aspect of this process is the iterative nature of the main pipeline execution. Various parameters are systematically altered based on a range of parameters values to be tested. This approach enables the evaluation across configurations and identification of the potential most suitable parameter setting for extracting a certain variable from the documents. It also aids decision-making in selecting from the potential options that can be applied to general RAG pipeline parameters by offering comparative results for each configuration and highlighting the best-obtained ones. In our proposed workflow, any of the previously stated parameters can be iteratively tested and analyzed by determining which options for each parameter type will be covered.

4. Experimental Evaluation

We analyzed and ran our experiments with forty selected Brazilian legal documents provided by Santa Catarina Government Agency for Law Enforcement and Prosecution of Crimes (MPSC), with an average of 26 pages and 60,000 characters each. Moreover, as mentioned in Section 3, while our proposed workflow allows the variation of any of the general RAG parameters, our experiments focused on alternating the parameters Large Language Models (Llama-2-7b, Llama-2-13b and Mistral-7B-v0.2), Chunk Size (128, 256, 512), Chunk Overlap Size (20, 50, 100, 200) and Top K (1, 3, 5, 8, 10, 12) and maintaining BERT Model, Splitting Strategy and Vector database as fixed parameters.

4.1. Data Preparation

The first step in the experiment, equivalent to step number 1 in Figure 1, is to prepare the available documents dataset to be used. This step is divided into two sub-steps: sampling and annotating. In the first sub-step, we filtered a smaller sample of the legal documents provided by MPSC, ensuring that all documents contained our study's analyzed variables. Then, to evaluate the accuracy of the experiments, the second sub-step was to manually annotated each selected document, mapping useful information including the variables to be evaluated. The annotations will be used to directly compare the model's responses to the constructed prompts, thereby assessing the accuracy of each extraction on every experiment.

4.2. Embeddings Creation

Embeddings were generated from the pieces of text parsed and chunked previously utilizing BERTimbau Large [Souza and et al. 2020], a BERT model pre-trained in Brazilian Portuguese, representing Step 2 in Figure 1. They were then stored in a vector database to manipulate and retrieve these embeddings. Chroma[2] was used as our option for vector database, a commonly chosen option for general RAG pipelines, highlighted for being open-source.

[2]https://github. com/chroma-core/chroma

4.3. Extracting variables

With the embeddings stored in the database, the next step was to embed the user's query, retrieve the most similar embeddings through a similar search, and finally use them as context on the prompt final form, represented by the template illustrated in Figure 2. These steps correspond to the main RAG pipeline, identified by step 3 in Figure 1.

```
## Instructions
You are a helpful AI assistant and provide the response in Portuguese to the question based on the provided context.
Use the following chunks of context to answer the question at the end. If it is not possible to answer the question from the
context, just answer that you didn't find the answer.
## Context built with Top K relevant retrieved chunks
CONTEXT: [RETRIEVED CHUNKS]
>>>QUESTION<<<: [USER QUERY]
>>>ANSWER<<<:
```

Figure 2. English translation of prompt template used in every experiment.

The LLMs options used in our experiments were the Llama-2 family chat models [Touvron and et al. 2023] and Mistral-7B-Instruct [Jiang and et al. 2023], all of them loaded locally with a RTX 3090 as the main GPU. This setup ensured privacy to handle the legal document's sensitive information, however, limited the involved models used in our study, making it impossible to handle bigger models on the current analysis.

4.4. Evaluation Metrics

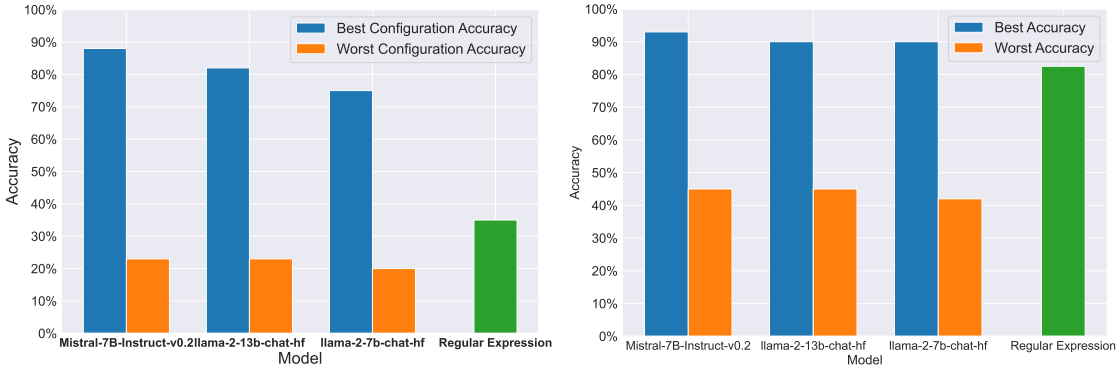
While several aspects of evaluation around RAG can be measured [Gao and et al. 2024], our work primarily concentrates on direct accuracy assessment. We specifically examine whether the generated response by the model precisely matches the annotated value associated with a particular document. This evaluation occurs in step 4 in Figure 1, which will divide the quantity of successfully matched extracted values by its annotation value by the total amount of documents.

4.5. Evaluated extracted variables

As previously stated, our work focuses on extracting two variables: public procurement process identifiers and municipalities of irregularity. The public procurement process identifier is a string that identifies a certain public procurement process for a municipality, and it is consistently presented in the format X/YYYY, where 'X' represents any numerical sequence and 'YYYY' denotes a four-digit year. The municipality of irregularity refers to the name of the municipality where fraud was committed through public procurement processes. Both variables have 145 associated experiments, one for every unique configuration possible interchanging the parameters detailed earlier in this Section.

5. Results and Discussion

As our baseline, we built a regular expression that looked for matches using the mentioned formats. When comparing the mode of the matches, extracting the variable with a regular expression reached a maximum of 35% accuracy against the best accuracy of 88% using our workflow when extracting public process identifiers and a maximum accuracy of 82.5% compared with our best accuracy of 93% when extracting municipalities. This comparison is visible in Figure 3, where the best obtained accuracies through our proposed method overcomes expressively the result obtained by the regular expression, when extracting public procurement process identifiers. For extracting municipalities, our experiments still bests the regular expression results by 10.5%. These results underscore the effectiveness of our method, overcoming regular expressions by contextualizing the

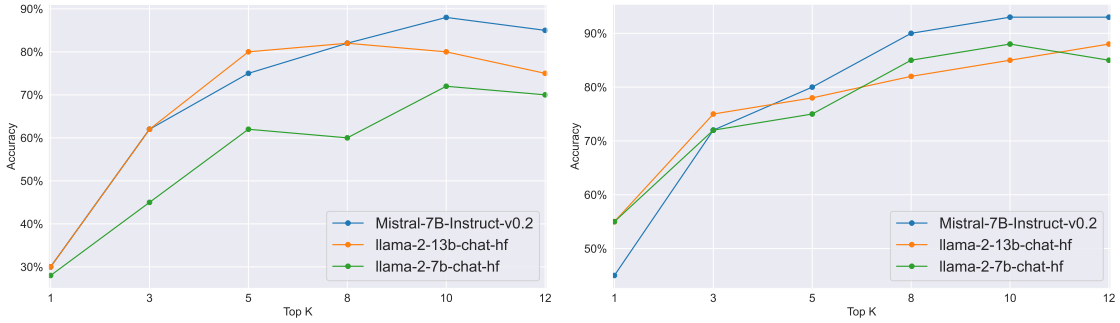


(a) Public Procurement Process Id. **(b) Municipality of Irregularity**

Figure 3. Best and Worst results per model vs Regular Expression

variables on the prompt fed to the LLMs. The built regular expressions for public procurement process identifier and municipality of irregularity are `\b\d+\/\d{4}\b` and `Município de ([A-Z][a-z]+(?:\s[A-Z][a-z]+)*)`, respectively.

Then, Figure 4 illustrates the Top K evolution and its impact on obtained accuracies on extracting both variables on fixed chunk sizes and chunk overlap. It suggests that the increase of Top k values directly impacts on the accuracy in pipelines with small chunks, increasing the probability of the retriever returning the accurate answer among the available embeddings.



(a) Public Procurement Process Id. **(b) Municipality of Irregularity**

Figure 4. Top K evolution with fixed chunk size as 128 and chunk overlap as 20.

6. Conclusion and Future Works

In conclusion, legal documents are often extensive and irregularly structured, and extracting relevant and structured data from these documents still poses a significant challenge. In this paper, we presented and evaluated a promising approach utilizing retrieval-augmented generation to extract two different variables of interest, obtaining an average accuracy of 90%, which overcame pattern matching measured accuracies in both scenarios. Our work addresses a common bottleneck for traditional extraction techniques — contextualization, and is part of a new paradigm of zero-shot IE, not requiring training or finetuning any models, representing a step forward on IE in legal domains.

Finally, our future works will focus on overcoming dataset and hardware limitations, in order to evaluate more expressive samples and include more robust Large Language Models. It will also be centralized in formalizing the proposed method as a RAG parameter evaluator framework for any type RAG pipeline for any system.

References

Bach and et al. (2019). Reference extraction from vietnamese legal documents. SoICT '19, page 486–493, New York, NY, USA. Association for Computing Machinery.

Bhattacharya, P. and et al. (2019). Identification of rhetorical roles of sentences in indian legal judgments.

Boisen, S. and et al. (2000). Annotating resources for information extraction. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Cheng and et al. (2009). Information extraction from legal documents. In *2009 Eighth International Symposium on Natural Language Processing*.

Doan, A. and et al. (2006). Managing information extraction: state of the art and research directions. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06', page 799–800, New York, NY, USA. Association for Computing Machinery.

Gao, Y. and et al. (2024). Retrieval-augmented generation for large language models: A survey.

Han, R. and et al. (2023). Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.

Huang, L. and et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Jiang, A. Q. and et al. (2023). Mistral 7b.

Kandpal, N. and et al. (2023). Large language models struggle to learn long-tail knowledge.

Katz, D. M. and et al. (2023). Natural language processing in the legal domain.

Kowsrihawat and et al. (2015). An information extraction framework for legal documents: A case study of thai supreme court verdicts. In *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 275–280. IEEE.

Liu, N. F. and et al. (2023). Lost in the middle: How language models use long contexts.

Pereira, J. and et al. (2024). Inacia: Integrating large language models in brazilian audit courts: Opportunities and challenges. *Digit. Gov.: Res. Pract.*

Sarkhel, R. and et al. (2021). Improving information extraction from visually rich documents using visual span representations. *Proc. VLDB Endow.*, 14(5):822–834.

Souza, F. and et al. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Touvron, H. and et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Vianna and et al. (2022). Organizing portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3388–3392, New York, NY, USA. Association for Computing Machinery.

Wachsmuth, H. and et al. (2013). Information extraction as a filtering task. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2049–2058, New York, NY, USA. Association for Computing Machinery.

Wei, X. and et al. (2024). Chatie: Zero-shot information extraction via chatting with chatgpt.

Zhu, W. and et al. (2012). Cross language information extraction for digitized textbooks of specific domains. In *2012 IEEE 12th International Conference on Computer and Information Technology*, pages 1114–1118.