Brazilian Consumer Protection Code: a methodology for a dataset to Question-Answer (QA) Models

Aline Athaydes , Lucas Bulcao , Caio Sacramento , Babacar Mane , Daniela Barreiro Claro , Marlo Souza , Robespierre Pita

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{alineathaydes,lucasbulcao,caio.sacramento,babacarm}@ufba.br

{dclaro,msouza1,robespierre.pita}@ufba.br

Abstract. *This work introduces the methodology for building a new dataset based on the Brazilian Consumer Protection Code (CDC), focusing on question-answer (QA) models. The dataset collected legal data, including CDC articles, legal summaries, and court rulings from the Superior Court of Justice (STJ). Automated data extraction techniques using Python were employed, and advanced language models such as Llama3-8b-8192, Gemma2-9b-it, and GPT-4o-mini were used to generate question-answer (QA) structures. This work presents our methodology for creating such a dataset to be used by language models for training in the legal domain, particularly in the CDC domain.*

1. Introduction

The demand for judicial services in Brazil has grown enormously, especially with the high consumption rates in society. The customer-consumer relationship has become increasingly problematic and generated many legal issues, increasingly demanding the attention of legal professionals. Consequently, their practices are more costly, time-consuming, and even error-prone. One judicial challenge is facilitating the general public's access to legal information in easily accessible language. This holds particularly true for Consumer Protection Law (CPL). The large number of cases makes it unfeasible for the legal structure to provide a timely decision in their judgments.

The evolution of technologies over time is remarkable, with humans constantly seeking tools to facilitate work and optimize time. In the legal field, this reality is no different. The AI era has empowered how legal issues are addressed, particularly with Large Language Models (LLMs). Thus, new possibilities arise to optimize access to and interpret legal information.

Question and Answer (QA) models have been widely employed in the legal domain to provide legal tasks, such as simplifying questions, summaries, image captioning, etc. The Brazilian Consumer Protection Code (CDC, which in Portuguese means "Código de Defesa do Consumidor") is the main code of Consumer Protection laws in Brazil; despite being widely widespread, for instance, each commercial establishment must provide access to a copy for its customers, the CDC is still relatively unknown by the large population, which frequently raises questions from citizens and legal professionals regarding the rights of consumers and providers of services and goods. In this regard, a QA model on

CDC might simplify access to information, popularizing accessibility to the legal domain and simplifying legal comprehension by citizens. As demonstrated by the development of [Silveira et al. 2023], a model specifically designed for the Brazilian legal context, adapting language models to legal needs significantly improves accessibility to complex legal information, making it easier for non-experts to navigate within the law - a recent concern of Brazil's Federal Supreme Tribunal.

In the era of LLMs, a QA model must have access to vast data to provide valuable answers. Thus, our first step in advancing the legal domain of CPL in Brazil is building a dataset to act as the source for these answers, as discussed in [Rajpurkar et al. 2016], highlighting the importance of having structured data and the challenges faced in training QA models. The methodology of creating such a dataset is introduced in this work, as far as the evaluation is concerned, to acquire the quality, accuracy, and contextualization of the generated answers. This process involves a pipeline from collecting and organizing a large amount of data related to the CPL domain, including CDC's 118 articles, 40 Supreme Court's summaries, and 10,504 rulings related to Consumer Protection Law, and evaluating the data to provide a dataset with questions and answers that will be used to train a QA model in the future.

Within this work, we tried to answer the following Research Question:

- *How does the lack of specialized datasets focus on the Consumer Protection Code (CDC) impact the development of efficient Question-Answer (QA) models in the Brazilian legal domain?*
- *How do different language models (Llama, Gemma, and GPT) perform in generating questions and answers about the CDC, and what are the key performance differences between them?*

Our first main is to explore the creation of a structured dataset for QA in the domain of Consumer Protection Law, employing LLMs to create triples of questions, answers, and contexts based on the articles of the CDC, discussing challenges, data organization techniques, and impacts in the legal domain.

This article is structured as follows: Section 2 describes the related works on QA models applied to the legal domain, emphasizing the lack of datasets focused on the Consumer Protection Law. Section 3 provides the details of how we built the dataset, including data normalization, organization, and the process of generating questions and answers using language models in the GPT, LLaMA, and Gemma families. Section 4 compares the generated answers from the models and discusses the challenges and findings. Finally, the last section outlines the future directions of using this dataset to train QA systems for the legal domain, particularly within CPL-related issues.

2. Related Works

Developing pre-trained language models for legal domains is essential for creating efficient question-and-answer systems. The work of [Silveira et al. 2023] demonstrated the importance of adapting language models to the Brazilian legal context, utilizing various legal texts, such as the Federal Constitution and the Civil Code but it did not include the Consumer Protection Code (CDC). The absence of a dataset for the CDC emphasized the need to focus on CDC articles, legal summaries, and related court rulings to train and validate language models for protecting consumer rights.

Our nearest approach is carried by [Jardim et al. 2023], who creates a question answering system centered on the CDC. Inspired by the dataset's methodology, we applied a similar strategy to generate questions and answers based on consumer protection law and jurisprudence. Their work showed that it is possible to structure a dataset from diversified textual sources, which was crucial in guiding the development of our dataset and focused on providing answers about consumer rights. Our approach is similar to [Jardim et al. 2023], where the authors created a question-answering dataset in the sports domain. The data extraction methodology and structuring it in a format suitable for question-answering systems served as a reference for our data organization.

Different from our approach, [Jardim et al. 2023] focuses on a sports-related domain rather than a legal one. While their methodology is valuable, they do not address the specific challenges and intricacies involved in legal contexts, particularly regarding the interpretation and application of legal documents like the CDC. Additionally, their work does not include court rulings or legal summaries, which are critical for providing a comprehensive understanding of consumer rights in Brazil.

From our knowledge, this is the initial dataset created for the CDC legal domain that has been included in a pipeline for Question-Answer (QA) tasks. Our work utilizes CDC articles, legal summaries, and court rulings to create a solid foundation for training language models that address consumer-related legal questions, filling a gap in the legal domain of consumer protection.

3. Method

Our data generation method is structured in a pipeline merging (1) consumer protection laws and jurisprudence, extracted from CDC, courts' rulings and summaries, (2) a data normalization process, (3) QA examples generation with LLMs, (4) Prompt Adjustments and Results.

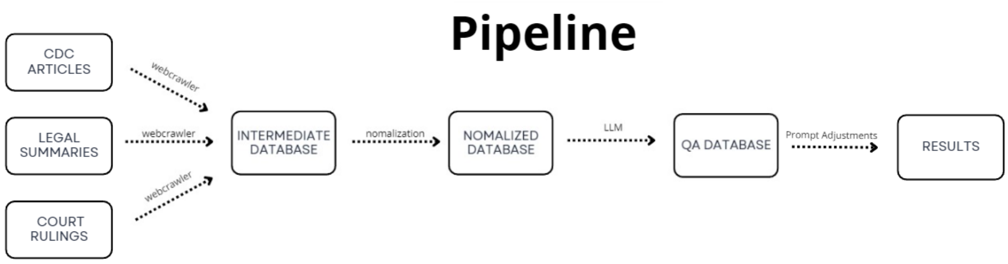

Figure 1. Methodology steps of our CDC dataset

3.1. Merging legal data of CDC

In the following, we discuss each step presented of the Pipeline depicted in Figure 1.

Firstly, we obtain the Brazilian Consumer Protection Code (CDC) directly from official websites through *web crawling* using *Selenium* with Python. The data is automatically structured on Titles, Chapters, Sections, and finally, Articles, based on the retrieved HTML structure. The complexity of the legal content was organized to make data manipulation easy for future steps.

Aiming to obtain a greater generality and applicability of our dataset, we expanded the initial set of laws extracted from the CDC with summaries and court rulings from the Superior Court of Justice (STJ). This expansion allows our dataset to better reflect the newest interpretation of consumer protection law and its applicability to different cases by incorporating relevant jurisprudence. We collected summaries and court rulings using the same web crawling approach, containing their respective numbers, descriptions, summaries, and detailed decisions. This process creates a more comprehensive and diversified dataset.

3.2. Data Normalization

After data collection, a textual normalization process was carried out to ensure consistency. This process is crucial to avoid issues during data manipulation, especially when performing automated analyses and feeding machine learning models. Managing accents and special characters in the Consumer Protection Code (CDC) legal texts, legal summaries, and rulings was one of the most challenging tasks we faced. These texts, often extracted from different sources, exhibited variations in encoding and formatting, which could compromise the quality of the final dataset.

We developed a Python function that utilizes the Unicode data library, known for efficiently handling strings. The function was designed to remove diacritical marks and special characters, transforming the text into a more standardized form and making it compatible with subsequent processing stages.

This normalization process is critical in legal contexts, where the accuracy of text representation can directly impact the interpretation of articles, summaries, and legal rulings. By standardizing the text and removing accents, we reduced inconsistencies.

The Superior Court of Justice (STJ) rulings represent a substantial part of the dataset and were particularly challenging due to their volume, with 10,504 rulings related to consumer protection. To facilitate processing and the generation of questions and answers, the rulings have been divided into 20 parts. We used the panda's library to handle these divisions and organize them into DataFrames.

3.3. Question Answer Generation with Language Models

To enrich our dataset, we used three language models for question-answer generation:

- **Gemma2-9b-it** [AI 2023]
- **Llama3-8b-8192**[Touvron et al. 2023]
- **GPT-4o-mini**[OpenAI 2023]

Adapted from [Jardim et al. 2023], we generated questions and answers based on the extracted CDC articles, summaries, and rulings. We focused on creating legally grounded questions and simulating scenarios. For this generation process, we connected our datasets to the models through their respective APIs and adjusted prompts, which coached the models to simulate common consumer doubts.

The development of [Silveira et al. 2023] directed our conceptual understanding to adapt pre-trained models to the Brazilian legal context. Unlike our approach, they do not include the CDC in their training, highlighting the importance of creating a dataset focused on the Consumer Protection Code, along with the STJ summaries and rulings.

3.4. Prompt Adjustments and Model Results

Adjusting the prompts was crucial in ensuring the quality of the generated questions and answers. To compare models, it was essential to create diverse questions and answers. For instance, there are three different text generation models: **Llama3-8b-8192**[Touvron et al. 2023], **GPT-4o-mini**[OpenAI 2023] and **Gemma2-9b-it** [AI 2023]. Every model received a carefully crafted prompt, with only minor alterations in keywords to reference CDC articles, legal summaries, or rulings.

The prompt adjustment was performed by requesting the generation of 10 to 30 questions for the Article of the CDC on the different models and evaluating the generated questions and answering, regarding informativeness and adequacy to the proposed use case, i.e. a user who is not a legal expert. Figure 2 shows the final prompt.

SYSTEM: Consider two distinct consumer groups: those who shop online and those who shop in physical stores. These consumers may have different levels of knowledge about the Consumer Protection Code (CDC), ranging from no knowledge to a basic or intermediate level, taking into account their shopping experiences and potential doubts. The questions should simulate real-life situations where they might need legal guidance or more details about how to ensure their rights. Focus on practical, everyday issues related to the CDC and situations these consumers might face. Never explicitly mention the SECTION in the questions or answers. The questions must be self-contained, meaning they should not require access to the section to be answered. Formulate questions technically, making them challenging and requiring a high level of understanding of the subject. Base the answers on LAWS, ARTICLES, RESOLUTIONS, and Rulings, explicitly integrating them into the ANSWER. For example: The Penal Code (art. X) stipulates the penalty...; Whenever there is mention of a law or article, integrate it into the ANSWER; Do not include personal opinions or speculations in the answers; Do NOT INCLUDE (Source:), (References:), (Based on section:), (Legal basis:) or similar. DO NOT CITE AUTHORS. DO NOT CITE THE SECTION.
USER: Assume the role of a Brazilian consumer with an issue involving online, and in-store purchases, or services provided by third-party companies, adopting the point of view of Brazilian consumers. Create a minimum of 10 and a maximum of 30 questions and answers based on " + {ementa} + ". Focus on real-life situations where they may need legal guidance, highlighting practical, everyday issues related to the Ruling. In addition to the questions and answers, bring the Ruling that supports the answer and its full description as context. Format the output exactly like this:
Question 1: [Text of question 1]
Answer 1: [Text of answer 1]
Context 1: [Description of Context´ 1]
...
Question {n}: [Text of question {n}]
Answer {n}: [Text of answer {n}]
Context {n}: [Description of Context {n}]

Figure 2. Prompt employed for generation fo the triples question-answer-context in the dataset

By using this procedure, questions and answers were generated consistently and then inserted into the corresponding data frame for each model.

4. Experiments and Results

In this section, we present our experiments and results with the three language models: **GPT-4o-mini**[OpenAI 2023], **Llama3-8b-8192**[Touvron et al. 2023] and **Gemma2-9b-it** [AI 2023]. We applied the same adjusted prompt to all models to ensure consistency in evaluating the results.

In this experiment, we provided CDC articles, legal summaries, and rulings as input to GPT-4o-mini, LLaMA3-8b-8192, and Gemma2-9b-it. Using an adjusted prompt, the models were fine-tuned to generate questions, answers, and corresponding contexts based on the legal data. We aimed to evaluate their ability to generate legally grounded responses in the consumer protection domain.

For each title, summary, and ruling, the models generated between 10 and 30 question-answer pairs. This allowed us to assess their performance in providing clear, accurate, and legally contextualized responses across various sections of the CDC.

4.1. GPT-4o-mini [OpenAI 2023]

We provide the following question to GPT-4o-mini:

Question : ”What should I do if the product I bought online arrived defective?”

Answer : ”You should contact the supplier immediately and request a replacement or repair of the product. If the defect is not resolved within 30 days, you can demand the substitution of the product, a refund of the amount paid, or a proportional reduction of the price.”

Context : Article 18 of the CDC: ”Suppliers of products and services are jointly liable for quality or quantity defects that render them unfit or inadequate for consumption...”

4.2. Llama3-8b-8192[Touvron et al. 2023]

We provide the following question to Llama3-8b-8192

Question : ”What is my right of recourse in case of defective products or services?”

Answer : ”According to Article 18 of the CDC, consumers have the right to receive products or services without defects. If the product or service is defective, the consumer can return to the company and request repair, replacement, or compensation, provided this is done reasonably and within the established deadline.”

Context : The Consumer Protection Code aims to protect consumers’ rights, ensuring they receive quality and safe products and services.

4.3. Gemma2-9b-it[AI 2023]

Question : ”I bought a phone online, and upon opening the box, I noticed the product was damaged. The store claims the damage occurred during shipping. Do I have the right to request an exchange or refund even if the damage was not caused by the store?”

Answer : ”Yes, It is within your rights to ask for an exchange or refund. Law No. 8.078/90, which establishes the Consumer Protection Code in Article 18, states that the product must be new, as advertised, and of quality. If the product arrived damaged, the store is responsible for providing it in perfect condition, regardless of the cause of the damage, as long as the consumer can prove the situation.”

Context : Article 18 of the CDC: ”The supplier is responsible for quality or quantity defects that render the products unfit or inadequate for consumption...”

4.4. Results and Discussions

In this section, we discuss the results of our experiments. To evaluate the models, we randomly selected 100 question-answer pairs from the dataset generated by each model and analyzed the proportion of incoherent, incomplete, or empty responses. These were classified into **adequate**, **regular**, and **irregular**. The dataset for this analysis is available in the **Data Analyze** repository, with code and dataset generation process documented [1].

4.4.1. Model Comparison

GPT-4o-mini[OpenAI 2023]: produced fewer QA pairs compared to the other models, but the quality was generally higher. The majority of the responses were classified as **adequate**, with a small portion considered **regular**. The answers were clear, well-founded in the CDC, and followed the prompt instructions accurately. **Llama3-8b-8192**[Touvron et al. 2023]: generated over 50,000 questions and answers, but many of these were regular. The responses were often generic and repetitive, although the model successfully followed the prompt instructions. **Gemma2-9b-it** [AI 2023]: faced significant challenges, producing the highest proportion of irregular responses, including empty and incoherent answers. It struggled with formatting, factual accuracy, and adherence to prompt instructions, highlighting its limitations in handling legal data. In a targeted analysis requesting 10 irregular examples, only Gemma provided results, while GPT-4o-mini and Llama3-8b-8192 returned none, further confirming Gemma's deficiencies.

4.4.2. Dataset Unification and Conversion

After generating questions and answers for the different models, we unified all the resulting tables into a single **DataFrame**. This data frame contained more than 5,000 questions and answers, which were then converted into **JSON** format, widely used in natural language processing (NLP) model training due to its simplicity and flexibility.

4.4.3. Architecture and Fine-Tuning Tool

We used the **Axolotl** tool to fine-tune the model and prepare the dataset, which proved efficient for adapting language models to specific legal domains. The tool facilitated fine-tuning, allowing us to reduce training costs and time. We used the **JSON** data structure to ensure the models received well-organized information, improving the precision and efficiency of the fine-tuning process.

Similar to [Silveira et al. 2023], we adapted our models to handle the vocabulary and complexities of Brazilian legislation. While the focus of this work is the **dataset construction**, it provides a solid foundation for future implementations of **Questions-Answer (QA)** systems in the domain of the **Brazilian Consumer Protection Code (CDC)**.

[1] https://github.com/FORMAS/DIGGER

5. Conclusions and Future Work

This work presented a methodology to create a dataset on the **Consumer Protection Code (CDC)**, incorporating articles, legal summaries, and court rulings. The complexity of the task required the use of various tools, including **Python**, **Selenium**, and advanced language models such as **Gemma2-9b-it** [AI 2023], **Llama3-8b-8192** [Touvron et al. 2023] and **GPT-4o-mini** [OpenAI 2023] to generate realistic and legally grounded questions and answers.

In future work, the goal is to use this dataset to train and fine-tune a model specifically designed for the **Question Answer (QA)** task related to the **CDC**. This model will have the ability to give precise responses to inquiries about consumer rights, facilitating both consumers and legal professionals with accessible legal information based on Brazilian law.

Acknowledgement

This work is partially financed by FAPESB CCE0022/2023, FAPESB TIC 002/2015, and CAPES Financial Code 001. We thank *Escavador.com* for their technical support.

References

AI, G. (2023). Gemma2-9b-it model documentation. `https://gemma.ai`. Accessed: 2024-09-12.

Jardim, P. C., Moraes, L. M. P., and Aguiar, C. D. d. A. (2023). Qasports: A question answering dataset about sports. *Repositório da Produção USP*.

OpenAI (2023). Gpt-4 technical report. `https://openai.com/research/gpt-4`.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 268–282. Sociedade Brasileira de Computação.

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.