Legal Document-Based, Domain-Driven Q&A System: LLMs in Perspective

Felipe Oliveira do Espírito Santo*, Sarajane Marques Peres*, Givanildo de Sousa Gramacho†, Anarosa Alves Franco Brandão‡ and Fabio Gagliardi Cozman‡
*Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brasil
Email: {felipeoes, sarajane}@usp.br
† Centro de Inteligência Artificial USP/IBM/FAPESP, Universidade de São Paulo, São Paulo, Brasil
Email: ggramacho1976@gmail.com
‡Escola Politécnica, Universidade de São Paulo, São Paulo, Brasil
Email: {anarosa.brandao, fgcozman}@usp.br

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650895

***Abstract*—Question Answering systems based on large language models are widely employed today, benefiting from continuous enhancements and improved performance. The legal domain has become a particularly active focus for Question Answering systems, given its complexity and social importance. This paper offers a discussion on how larger and smaller language models can be used to build a legal document-based Question Answering system. We present a novel model, named Cocoruta, generated by fine-tuning with a corpus of legal documents. In addition, we examine five LLMs as they answer questions related to the legal aspects of a specific domain – the Blue Amazon, a region of particular interest involving environmental issues. The results suggest that while LLMs are not yet of sufficient quality for use as core in legal context Question Answering systems, fine-tuning on specialized corpora imparts a beneficial bias to their legal discourse. Despite having fewer parameters, the Cocoruta model competes well with larger LLMs in this aspect.**

***Index Terms*—Large language models, LLM evaluation, legal Q&A systems, legal-document corpus**

I. Introduction

Question Answering (Q&A) systems have gained prominence among organizations and individuals, driven by the advancement of large language models (LLMs). Q&A systems integrate effectively with dialogue systems [6, 12], thus enabling one to build specialized and useful chatbots. The construction of such systems has been facilitated by open-source frameworks, popularizing their development [15, 14, 10, 12]. However, it is crucial to note that this ease is restricted by technical implementation issues. For example, a Q&A system based on LLMs and aimed at a specific domain or a certain textual style requires a training corpus and conducting specialized evaluation of the outputs generated by the system. Moreover, the inherent complexity of the inductive learning underlying LLMs imposes another barrier in this context. Often, the limitations of this type of computational learning [9] are unknown to users, preventing them from properly dealing with inaccurate or inadequate system responses. As a result, the apparent ease of implementing these systems can lead to frustrated expectations and expose the user to unwanted situations during or after interacting with the system.

This paper focuses on the study of LLMs as base for Q&A systems intended for a specific domain, with a particular emphasis on addressing legal queries. In particular, we present an LLM-based Q&A system named Cocoruta[1], which was tailored to respond to questions related to the legal aspects concerning the coast and the exclusive economic zone in Brazil, dubbed Blue Amazon[2].

The motivation behind such an study and the creation of Cocoruta stemmed from a sequence of unsatisfactory interactions in seeking information from the ChatGPT 3.5 system. To evaluate the feasibility of a chatbot system for disseminating useful information about the Blue Amazon to the general population, the ChatGPT 3.5 system was tested as a proof of concept. An example of a question posed to it led to a well-structured answer but with little useful information. One fo the questions was: "*What are the rules for building a resort on Ubatuba*[3] *beach?*" The response addressed the "impossibility of providing information about specific places", suggested seeking competent authorities and bodies, and warned about the need to follow local and environmental laws. Although correct, the answer is not helpful, given that it does not provide clear guidelines to direct the user. In this context, we hypothesize that a system with legal knowledge related to the Blue Amazon could provide more satisfactory answers.

Driven by the desire to develop a domain-specific system, a corpus of legal documents was created. A set of question-answer pairs derived from this corpus was then used to fine-tune the LLaMa 2-7B model, resulting in the Cocoruta 1.0.

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The fifth author was partially supported by CNPq grant 305753/2022-3.

[1]"Cocoruta" is the name given to a species of bird endemic to the Fernando de Noronha archipelago (Brazil), currently threatened with extinction. The name of the system was chosen as a way to honor biodiversity and support the defense of the conservation of the Blue Amazon.

[2]The term "Blue Amazon" (originally "Amazônia Azul" in Portuguese) was initially coined by the Brazilian Navy to draw a parallel between the region of Brazilian jurisdictional waters and the Amazon Rainforest. This comparison underscores the significance of both regions for Brazil.

[3]Ubatuba is a well-known coastal city in Brazil.

This novel system underwent both quantitative and qualitative evaluations. A portion of the tests conducted on Cocoruta 1.0 was also subjected to other LLMs (LLaMa 2-7B [15], GPT-3.5 Turbo-Instruct [1], the Maritalk 1.0 large[4], and Sabiá-7b [13]). This comparative analysis provided insights into the virtues and limitations of these LLMs in our scope of exploration. The LLMs tested were selected given the experimental nature of the research. Cocoruta 1.0 was fine-tuned from a 7 billion-parameter model; hence, we explored models of comparable size when available, to ensure fair comparisons. The primary contributions brought forth herein are:

- the introduction of Cocoruta 1.0, a legal document-based, domain-driven Q&A system, developed through the fine-tuning of LLM LLaMa 2-7B, using a corpus consisting of laws, provisional measures, decrees, ordinances, and other pertinent legal documents addressing national governance issues related to the Blue Amazon;
- a comprehensive and critical examination of the effectiveness of LLMs as the foundation for a legal document-based, domain-driven Q&A system. Such an examination focuses on a qualitative evaluation, covering the evaluation of the use of speech in legal language, the identification of obvious hallucinations, and the control of responses involving sensitive subjects such as racism, misogyny, national security, and related issues.

This paper is structured as follows: Section II provides a brief background about LLMs and evaluation strategies; Section III introduces the Cocoruta 1.0 system; Section IV presents the criticism of LLMs; Section V discusses related work; and Section VI provides the final remarks.

II. Background

A. Large Language Models

Language models can be defined as either statistical or neural network-based machine learning models, possessing the capability to comprehend and generate human language. In technical terms, these models are specialized in the probabilistic modeling of sequences of words or tokens, facilitating the prediction of probabilities for future or missing words or tokens [2, 22].

Large Language Models (LLMs) represent the most recent evolution in language models. Their development is driven by the discovery that scaling pre-trained language models, whether by increasing model size or data size, often leads to improved model capabilities in subsequent tasks [22]. LLMs are trained with extensive volumes of text, sourced from diverse domains, and are constructed using architectures comprising hundreds of billions, or more, parameters. Technically, their development relies on self-attention modules organized within a Transformer architecture. LLMs are further characterized as foundational models, serving as a shared knowledge and skill base upon which engines can be built to implement diverse applications within the field of natural language processing. In this paper, our particular focus lies on the following families of LLMs:

- LLaMa 2: this family consists of LLMs ranging in size from 7 billion to 70 billion parameters and trained on an extensive multi-domain text corpus, comprising 1 trillion training tokens [15]. LLaMa 2-7B is the model used in this paper.
- GPT 3.5: a family of LLMs with models reaching up to 175 billion parameters and trained on a total of 300 billion tokens from several curated high-quality text datasets [1]. GPT-3.5 Turbo (Instruct) is the model used in this paper.
- Sabiá and Maritalk: family of models trained based on the LLaMa models using, in the case of Sabiá family[5], Portuguese-language datasets comprising 10.4 billion tokens. In this paper, we use the model Sabiá 7B [13] and Maritalk-large[6].

B. Evaluation techniques

In this research, two types of evaluation were employed: automatic quantitative evaluation and human qualitative evaluation. For the automatic evaluation, the resubstitution error strategy was utilized, and the indicators BLEU, ROUGE-N, BERTSCORE, and MOVERSCORE[7] were employed, based on the following definitions:

- resubstitution error [4]: it is applied if the training set is used to estimate an error rate, resulting in an optimistic estimate of the true error rate because the model is not tested on any samples that it has not already seen. The error value reflects the upper limit for the learning effectiveness achieved by a learning model.
- BLUE (Bilingual Evaluation Understudy) [11]: it was introduced for evaluating the quality of machine translations. This indicator evaluate the n-gram overlap between the evaluated text (hypothesis) and one or more reference texts, emphasizing a “precision” intuition. The outcome is on a scale from 0 to 1, where 0 signifies no n-gram overlaps, and 1 indicates identical texts in the comparison.
- ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation) [7]: it is part of the ROUGE set of metrics that was introduced to evaluate summarization and translation systems. Similar to BLEU, the ROUGE-N metric computes the proportion of common n-grams between the reference and the hypothesis, emphasizing a “recall” intuition. The ROUGE-1 was applied in this study.
- BERTSCORE [21]: it is calculated via a similarity score for each token in the hypothesis text with each token in the reference text, using BERT’s contextual representations. The multilingual BERT model was applied herein.
- MOVERSCORE [23]: it considers the displacement of specific entities within the generated text, attributing higher scores to texts that preserve the relative order and

[4]https://www.maritaca.ai/

[5]Details about Maritalk training are not available.

[6]https://www.maritaca.ai/

[7]There exist various criticisms of diverse quantitative indicators employed in evaluating responses to questions. The utilization of multiple indicators, coupled with human evaluation, seeks to mitigate evaluation biases.

cohesion of the entities found in the reference. The final score is a weighted measure that evaluates both, order of entities and their co-occurrence throughout the text. The MOVERSCORE was applied using the embeddings of the multilingual BERT model.

The human evaluation was conducted by analyzing responses to spontaneously generated questions and questions specifically designed to atacking the model (intentionally addressing sensitive topics such as racism, misogyny, etc.).

III. COCORUTA 1.0

The development of Cocoruta 1.0 required the compilation of a new corpus, formulation of Q&A datasets, fine-tuning of the LLaMa 2-7B model, and deploying a system. This section provides an overview of these components and procedures.

A. Corpus and Q&A dataset

In order to build the corpus, an ethically guided data scraping process[8] was executed on websites associated with entities within the Brazilian federal government[9]. This scraping operation involved the retrieval of laws, provisional measures, decrees, ordinances, and other pertinent legal documents addressing national governance issues. A total of 172,408 documents were successfully obtained, constituting a substantial corpus containing 67.2 million tokens. To tailor the corpus to focus on the Blue Amazon domain, documents were filtered based on words associated with the "ocean" domain (e.g. ocean, sea, navy, beach, island, estuary, archipelago, bay, continental margin, petroleum, pre-salt, blue amazon, exclusive economic zone continental shelf and similars words). A specialized regular expression was crafted to enact this filter. Consequently, the corpus has been refined to encompass 68,991 documents, totaling 28.4 million tokens.

In the current version of the corpus, aside from filtering, no curation work was undertaken. Consequently, the corpus encompasses archival documents, including those that have been annulled, as well as a compilation of legal documents from Imperial Brazil. Due to the presence of outdated information and content occasionally deemed unsuitable for contemporary social discourse, we exercise caution in refraining from making the corpus publicly accessible[10]. Harmful effects of using this type of information are discussed in Section IV.

From the filtered corpus, a collection of questions and answers (Legal Q&A Dataset) was created to enable the fine-tuning of the LLaMA model for the purpose of supporting a legal document-based, domain-driven Q&A system. The necessity to construct a comprehensive dataset required the use of an automated generation system. Some examples[11] of question-answer pairs are:

- *Question*: What is the Authorization for the Operation of Tourist Vessels in Conservation Units?
  Answer: The Authorization for the Operation of Tourist Vessels in Conservation Units is a document issued by ICMBio that allows the operation of tourist vessels within conservation units. To obtain the authorization, it is necessary to meet the following requirements: * Be a legal entity; * Possess a seaworthy vessel; * Have a qualified team; * Adhere to the rules and regulations of the conservation unit. The authorization is valid for a period of five years.
- 2: *Question*: What is the purpose of Decree No. 13,558, dated September 30, 1943?
  Answer: The purpose of the decree is to approve the project and budget for complementary works at the port of Itajaí.
- 3: *Question*: What types of fishing are prohibited in the Canavieiras Extractive Reserve[12]?
  Answer: According to Ordinance No. 1,124 of December 2018, the types of fishing prohibited in the Canavieiras Extractive Reserve are: purse seine fishing with motorized vessels, trawl fishing within a thousand meters from the coast, Red Lobster (Palinurus argus) fishing using trap nets, silk nets, and crab pots, fishing with nets on rocky areas within the Reserve, and trawl fishing with a 6-cylinder or larger engine.
- 4: *Question*: What are the beneficiary categories of the Pirajubaé Extractive Reserve?
  Answer: According to ICMBio Ordinance No. 532, dated August 15, 2017, the fishermen and fisherwomen beneficiaries of the Pirajubaé Extractive Reserve will be subdivided into three categories: * Category A – if the beneficiary primarily depends on the natural resources of the Extractive Reserve. * Category B – if the activity carried out by the beneficiary in the Extractive Reserve is a secondary source of income, and their other sources of income combined are up to 05 (five) minimum wages. * Category C – the beneficiary not covered by categories A and B, who uses the Extractive Reserve to maintain their tradition and cultural transmission to their descendants.
  The category recognition will be carried out by ICMBio, based on the criteria established in the Ordinance.

To generate question-answer pairs, the documents were segmented into excerpts referred to as "contexts", each comprising approximately 4,000 characters, with an overlap of 1,000 characters to maximize the overall coverage of the

[8]Ethical data scraping involves the automated extraction of data from web pages while adhering to the standards and guidelines set by the source sites, respecting, and avoiding server overload and copyright violations.

[9]CONAMA (National Environmental Council), ICMBio (Chico Mendes Institute for Biodiversity Conservation), Civil House, Official Gazette and Legislation Portal.

[10]The curation of this corpus is currently in progress, supported by the expertise of legal professionals and experts in Brazilian coastal matters.

[11]All questions and answers examples presented in this paper were originally formulated in Portuguese (either by models or human testers) and were translated into English by the authors to facilitate understanding of the discussion presented herein.

[12]The abbreviation "Resex" was used in the original question in Portuguese in place of "Extractive Reserve".

documents. Based on each context, three pairs of questions and answers were generated by generative models.

B. Fine-tuning

The LLM selected to constitute the first version of the Cocoruta system is the model LLaMa 2-7B. Opting for a model within the LLaMa family was mainly because it is open-source and ease of use. The decision to adopt the smallest model in the family was justified by the experimental nature of this research and considerations regarding computational and energy resources. The Cocoruta 1.0 system aims to address legal inquiries related to the Blue Amazon in Portuguese. Given the unique characteristics of the knowledge intended for the system, it was hypothesized to be useful to fine-tune the selected language model for the system's implementation.

Fine-tuning was performed using LoRA (Low-Rank Adaptation) as a way to make changes to a smaller number of parameters, reducing associated costs and preventing overfitting and catastrophic forgetting [8]. The fine-tuning process lasted 15 epochs[13] and was performed for 1.17% of the total trainable parameters, which is equivalent to approximately 119 million parameters. The following LoRA configuration was defined based on available resources and empirical testing:

- `r = 16`: dimension of the matrices to be updated.
- `lora_alpha = 8`: scaling factor applied to the matrices undergoing updates.
- `target_modules = all`: adjustment across all linear layers within the architecture.
- `bias = None`: treatment of the bias (or polarization) term is ignored.
- `task_type = CAUSAL_LM`: task of interest is "causal language modeling", involving the prediction of the next token based on a sequence of tokens.

C. System deployment

A interactive interface, similar to the traditional interface of a conversational agent, was built and made available as a service[14]. The necessity for creating the interface emerged during the qualitative testing phase of Cocoruta 1.0. In these tests, human users posed questions to the model, including attempts to attack it, as detailed in Section IV. Although Cocoruta 1.0 is not a dialogue system, the dialogue interface was designed to be engaging, enabling users to formulate new questions stimulated by answers to previous ones. The implementation of the interface involved adapting the OpenChat UI framework[15].

IV. Evaluation

This section outlines the outcomes of a quantitative evaluation for Cocoruta 1.0, accompanied by a qualitative evaluation of this model as well as other LLMs. This combined analysis facilitates an examination of how different LLMs respond to questions within the legal context of the Blue Amazon.

A. Cocoruta 1.0 quantitative evaluation

The quantitative evaluation applied to Cocoruta 1.0 aimed to verify whether the LLaMa model, adjusted to the Legal Q&A Dataset, showed evidence of learning. The evaluation results stem from the application of quantitative indicators using the resubstitution error calculation strategy. Therefore, the evaluation is conducted on the dataset used in the fine-tuning procedure. Table I presents the the indicators values, for various limits on the number of tokens allowed in responses. The comparison is made between the original version of the LLaMa 7-2B model and the adjusted version Cocoruta 1.0.

TABLE I
Indicator values for model responses, considering different limits for token generation. BERT and MOVER refer to the BERTSCORE and MOVERSCORE indicators respectively

Model	ROUGE	BLEU	BERT	MOVER
	128 tokens			
LLaMa 2	0.24	0.02	0.67	0.48
Cocoruta 1.0	0.75	0.44	0.90	0.72
	256 tokens			
LLaMa 2	0.23	0.02	0.67	0.47
Cocoruta 1.0	0.80	0.62	0.91	0.77
	512 tokens			
LLaMa 2	0.22	0.01	0.66	0.45
Cocoruta 1.0	0.81	0.69	0.92	0.78
	1024 tokens			
LLaMa 2	0.20	0.01	0.65	0.44
Cocoruta 1.0	0.81	0.70	0.92	0.79

The values obtained for the indicators show a learning gain from fine-tuning[16], as evidenced by the higher quality of responses presented by Cocoruta 1.0 facing all indicators. It is notable that an increase in the number of tokens beyond 256 did not yield significant benefits for Cocoruta 1.0 and, in fact, showed a slight decline in the performance of the original LLaMa 2-7B[17]. This outcome is expected, considering that the average number of tokens in the Legal Q&A Dataset is 200 tokens. The lower performance of the original LLaMa 2-7B compared to Cocoruta 1.0 suggests a challenge for LLM in its original version when dealing with the specific characteristics of the problem represented in the Legal Q&A Dataset.

Quantitative indicators pose challenges in their interpretation within tasks associated with natural language generation. To provide an understanding about the behavior of these indicators in the context of this quantitative evaluation, consider the examples in Table II. This table illustrates two scenarios in which the ROUGE and BERT indicators indicated higher and lower answer quality. In the first example, the Cocoruta model mentions a series of fishing strategies that are not present in the ground truth, while the LLaMa model provides a response mentioning a smaller variety of prohibited fishing types, another proposal for a reserve and a specific river. In

[13]This training takes approximately 44 hours on an A100 GPU with 80GB of GPU RAM.

[14]For the reasons outlined in Section III-A, the Cocoruta interface has not been made publicly accessible. Access to the system is attainable through direct communication with the authors, only for scientific exploration purposes.

[15]https://github.com/imoneoi/openchat-ui

[16]According to the Wilcoxon test, there are significant differences between the evaluation obtained for LLaMa 2 and Cocoruta 1.0.

[17]The Wilcoxon test does not indicate significance in this variation.

the second example, the ROUGE indicator for the Cocoruta model achieves a higher value than in the first example. The LLaMa model does not present an improved evaluation for its answer, including a decline in performance according to the BERT indicator. The presence of a type of noise in the LLaMa model's answer may justify such a decline.

As illustrated in Table II, the examples highlight the limited interpretative and judgment capabilities of quantitative indicators in natural language discourse. The suitability of models for practical use necessitates a qualitative evaluation, as discussed in the subsequent section.

B. Putting LLMs in perspective

This section provides a critical analysis of the efficacy of LLMs as a foundation for constructing legal document based, Blue Amazon domain-driven Q&A systems. Two scales of LLMs are taken into account: (a) smaller models, with 7 billion parameters (LLaMa 2-7B, Sabiá 7B, and Cocoruta 1.0); (b) larger models with 120 and 175 billion parameters (Maritalk-large and GPT-3.5 Turbo Instruct, respectively). In the set of evaluated models, only Cocoruta underwent the fine-tuning process using the Legal Q&A Dataset.

The critical analysis was outlined based on a qualitative evaluation of answers provided by LLMs to a new set of questions (with no intersection with the questions on the Legal Q&A Dataset used in the fine-tuning procedure). Members of the research group, who were not directly involved in Cocoruta development and training tasks, were invited to use the system's chat interface to ask questions. The answers provided by the models were evaluate by one of the members. The evaluation of the answers considered three aspects:

1) If the answers adhered to the language used in legislation and normative contexts. Testers[18] were instructed to pose questions such as: what is prohibited, what are the rules, who is responsible for, since when is it permitted, what is the responsibility of the navy, the citizen, or the govern.
2) If the answers were evidently incorrect, with some degree of hallucination, or fall outside the Blue Amazon domain, including references to other geographic regions or knowledge domains, even when the question was appropriately directed to the correct domain.
3) If the answers contained inappropriate speech, such as inappropriate references to the era of slavery in Imperial Brazil, misogynistic or *LGBT+*phobic speeches, or discussing illicit actions. Testers followed the guideline for this evaluation item by posing questions that directly mentioned inappropriate speech or indirectly created likely situations for the use of inappropriate speech.

The qualitative evaluation comprised 147 questions. The set of questions covered topics directly or indirectly related to legal aspects (50 questions), as well as generalities within the Blue Amazon domain (39 questions). A portion of the questions aimed at attacking the model (49 questions). Some queries were submitted incompletely (2 questions), while others addressed issues beyond the scope of the Blue Amazon (7 questions). The remainder of this section is devoted to exploring the performance of the models in light of the three qualitative analysis aspects established in this research.

1) Adherence to legal speech: The first evaluation conducted pertained to the models' ability to steer their responses towards the legal domain, employing discourse that explicitly references legal documents, authorities, institutions, or governmental bodies. Table III summarizes the frequency with which each model complied with the evaluation requirement for the 50 questions related to legal aspects.

In general terms, with the exception of the Sabiá model, the models demonstrated similar performance regarding the use of legal speech. The smaller models (LLaMA and Cocoruta) stand out for achieving comparable proficiency to the larger models. Cocoruta distinguishes itself from the other models by consistently employing discourse with explicit references to legal matters when the question pertains to permissions (e.g., whether visiting Fernando de Noronha is prohibited[19] or if constructing a resort on the beach in Ubatuba is allowed).

The discourse appropriate within the legal context, however, does not necessarily imply that the provided answer is correctly formulated and precisely references the correct regulatory documents or responsible institutions. By way of illustration, we refer to the responses listed in the Table IV. In their answers, the models Cocoruta, Maritalk, and GPT referenced the same law related to environmental crimes. These three models also appropriately contextualized their responses within the framework of criminal judgment. Cocoruta specifically noted the "Maritime Tribunal" in its answer, an autonomous entity with jurisdiction nationwide and auxiliary to the Judiciary. However, this entity is active and holds the power of judgment solely in matters of navigation. The models LLaMa and Sabiá cited other Brazilian institutions responsible for monitoring and investigation only. The articles and laws referenced by the LLaMa are incorrect; the articles address educational matters within public schools and universities, and the law pertains to a general budgetary provision.

Regarding the remaining 95 valid questions in the test set (encompassing generalities, attacks, and questions out of Blue Amazon domain), the Cocoruta model demonstrates the capability to employ legal speech in its responses in 31 instances (32%). This proficiency is followed by the Maritalk and LLaMa models, which maintain such a style of speech in 24 responses (25%) and 22 responses (23%), respectively. In contrast, the GPT models exhibit a lower inclination to utilize legal speech, recording rates of 15% and 4% in terms of responses adopting such a style, respectively. The use of legal speech is positive for a Q&A system grounded in legal documents. Therefore, the ability to accurately establish

[18] Testers, who are computer scientists and engineers, lacked legal knowledge or specific understanding of the Blue Amazon domain. Therefore, the evaluation did not analyze the correctness and precision of the answers.

[19] A famous Brazilian archipelago.

TABLE II

RELATIONSHIP BETWEEN THE MODEL'S ANSWERS, THE EXPECTED ANSWERS (GROUND TRUTH) AND QUANTITATIVE EVALUATION VALUES

Example 1	*Question*: What are the types of fishing prohibited in the Canavieiras Extractive Reserve? *Ground truth*: According to Ordinance No. 1124 of December 2018, the types of fishing prohibited in the Canavieiras Extractive Reserve are: purse seine fishing with motorized vessels, trawl fishing within a thousand meters of the coast, fishing for Red Lobster (Palinurus argus) using gillnets, silk nets, and traps or pots, fishing with nets over rocky areas within the Reserve, and trawl fishing with a 6-cylinder or larger engine.		
	Answers	ROUGE	BERT
Cocoruta	In the Canavieiras Extractive Reserve, it is prohibited to fish using doors trawl nets, pair trawls, midwater trawls, gillnets, explosives, and toxic substances. It is also forbidden to fish with nets longer than 400 meters or rectangular nets with sides longer than 100 meters.	0.385	0.723
LLaMa	Despite the Reserve being a public space open for recreational activities, the use of trawl nets is prohibited in the Canavieira River. Only capture with fishing rods or hooks is allowed.	0.222	0.698
Example 2	*Question*: How can a state obtain the delegation of authority to enforce federal laws on hunting and fishing? *Ground truth*: A state must request the delegation of authority from the Ministry of Agriculture, submitting the request along with a detailed and well-documented report on the organization of the hunting and fishing service existing in the state. The application will be subjected to analysis by the Hunting and Fishing Division of the National Department of Animal Production to provide an opinion on the appropriateness of granting the requested measure.		
	Answers	ROUGE	BERT
Cocoruta	A state can obtain the delegation of authority to enforce federal laws regarding hunting and fishing by submitting a request to the National Department of Animal Production of the Ministry of Agriculture, accompanied by a detailed report on the organization of the service relevant to this purpose. (*continues with a list of conditions*)	0.535	0.767
LLaMa	Article 102 of the Federal Constitution stipulates that states have the exclusive right to regulate hunting and fishing within their territory. In other words, they possess the authority to approve or reject national laws on the subject. Therefore, there is no requirement for the delegation of competence. However, it is still possible for a delegation of competence to occur between state and federal governments if there is mutual interest in the matter. (*continues by citing laws, and a list of "documents" and "external links" similar to Wikipedia verbetes*)	0.297	0.644

TABLE III

CAPACITY OF THE MODELS TO PROVIDE ANSWERS USING LEGAL SPEECH: CITING LEGAL DOCUMENTS (LAWS, DECREES ETC.), AUTHORITIES, INSTITUTIONS, OR GOVERNMENTAL BODIES. LEGEND: # – NUMBER OF ANSWERS

		Smaller models			Larger models	
	#	Cocoruta	LLaMa	Sabiá	Maritalk	GPT
Rules	10	7	10	4	8	8
Permissions	8	6	1	1	2	4
Responsabilities	10	9	10	9	9	10
Authorities and positions	11	10	10	9	8	9
Temporal issues	11	5	6	4	7	5
Total	50	37	37	27	34	36
	%	74%	74%	27%	68%	72%

associations with documents and institutions still needs enhancement in all models under analysis.

2) Incorret answers or hallucinations: In this analysis, the correctness of answers was primarily evaluated with respect to the purpose of the question. For instance, questions employing the structure "what are the rules" necessitated an answer explicitly mentioning at least one rule, guideline or requirement. Questions structured as "since when" demanded a temporal reference in the answer. Answers featuring references to international geographic locations or explicit mention of the region of the Amazon Rainforest were deemed incorrect. Answers displaying a context entirely disconnected from the question were categorized as hallucinations. Table V presents the performance of the models in this evaluation criterion. Only questions related to legal aspects, about generalities or beyond the scope of Blue Amazon were included in this evaluation. Questions aimed at attacking the model were included in the next aspect of evaluation.

The larger models are equally proficient and exhibit well-formulated and comprehensive answers. Among the smaller models, Cocoruta and Sabiá display more correct answers than the LLaMa model, indicating that fine-tuning on data related to the specific problem is beneficial. In the case of the Sabiá model, the fine-tuning data is in the Portuguese language, and not specifically derived from legal documents and those oriented towards the domain of the Blue Amazon. For this reason, the Cocoruta model outperforms it in the set of questions focused on the legal aspect. Furthermore, in favor of the fine-tuning performed on the Cocoruta model is the absence of noise in the provided answers, unlike the LLaMa model (7 answers with noise) and Sabiá model (13 answers with noise). The noise identified pertains to orthographic issues (multiple repetitions of a single letter in various words), references to documents claimed as "supposed" sources for the provided answers, inclusion of repeated sentences within responses, and the incorporation of random sentences or character sequences.

Note that the correctness of the answers was not evaluated in relation to the real legal and scientific context. The complexity and specificity of the subject require the involvement of a board of experts to conduct this evaluation[20]. However, the analysis conducted anticipates the challenge for the models in handling factual information, and even with the fine-tuning performed for Cocoruta, this difficulty persists. Table VI lists the responses provided by the analyzed models to a question with reference to a temporal landmark.

To understand the question in Table VI, we conducted

[20]An evaluation involving experts is part of the next steps in this research.

TABLE IV
ANSWERS PROVIDED BY THE MODELS EXHIBITING APPROPRIATENESS TO LEGAL SPEECH BUT WITH INACCURACIES REGARDING THE REFERENCED DOCUMENTS OR INSTITUTIONS.

	Question: Who is responsible for judging cases related to environmental crimes on the Brazilian coastline?
	Answers:
Cocoruta	The Maritime Tribunal is the entity responsible for judging cases related to environmental crimes on the Brazilian coastline, as stipulated by Article 12 of Law No. 9.605/1998.
LLaMa 2-7B	The Federal Public Ministry (MPF) is the entity vested with the authority to investigate environmental crimes outlined in Articles 207-A and 213 of the 1988 Federal Constitution. The MPF can be called upon by the State or civil society to conduct investigations into suspected environmental crimes. Furthermore, in accordance with Law No. 6.405/76, the MPF is empowered to initiate preventive inquiries when there is sufficient evidence of the commission of an environmental crime.
Sabiá	Ibama, which is the body responsible for environmental inspection in Brazil.
Maritalk	The entity responsible for judging cases related to environmental crimes on the Brazilian coastline is the Judiciary, specifically the Federal Judiciary. (...) In these instances, jurisdiction for the judgment lies with the Federal Judiciary, as outlined in the Federal Constitution and the Environmental Crimes Law (Law No. 9.605/1998). (*continues*...)
GPT	The entity responsible for judging cases related to environmental crimes on the Brazilian coastline is the Judiciary, more specifically, federal judges. In Brazil, environmental crimes are regulated by Law No. 9.605/1998, known as the Environmental Crimes Law. This law delineates the behaviors deemed as crimes against the environment, as well as the corresponding penalties for each offense. (*continues*...)

TABLE V
THE MODELS' CAPABILITY TO PROVIDE CORRECT ANSWERS, BY EVALUATING THE ADHERENCE OF THE ANSWERS TO THE CONTEXT OF THE QUESTION, THE PRESENCE OF CORRECT GEOGRAPHICAL INFORMATION, AND THE ABSENCE OF HALLUCINATIONS. LEGEND: #: NUMBER OF CORRECT ANSWERS

		Smaller models			Larger models	
	#	Cocoruta	LLaMa	Sabiá	Maritalk	GPT
Legal aspects	50	39	27	29	48	47
Generalities / out of domain	46	27	25	34	44	44
Total	96	66	52	63	92	91
	%	68%	54%	65%	95%	94%

research on the national day mentioned in the question. As far as we could ascertain, June 8th was approved as the National Day for Combating Ocean Pollution in Brazil by the Brazilian Federal Senate in 2019. The date was chosen because it also corresponds to World Oceans Day. As illustrated in Table VI, none of the models were capable of providing the correct answer. All answers allow for the inference of an incorrect date regarding the approval of the mentioned law or the commemorative day in question. The GPT model also provides an associated explanation, which is likewise incorrect.

3) Inappropriate speech: The third aspect addressed in the qualitative evaluation pertains to the ability of the models to handle sensitive topics appropriately (with appropriate speech). The 49 questions formulated for this evaluation, considered as attack questions directed at the models, were categorized into four groups: (a) questions designed to prompt the models to discuss race and racism; (b) questions intended to encourage the models to address gender, particularly examining potential misogynistic speeches; (c) questions touching on subjects related to permissions, crimes, and punishments; (d) questions concerning national security, drug trafficking, corruption, and topics of a sexual nature. Table VII provides information on the vulnerability of the models concerning these subjects.

The evaluation of this aspect is the most complex to be carried out due to its high subjectivity. Two particular situations generated answers that were considered inappropriate to some extent in this research, but could be subject to questioning:

- the answers for the set of questions (c) in all models. In this case, answers deemed inappropriate explicitly mentioned the denial of any distinction between individuals based on race or gender for questions addressing people in general, without explicitly mentioning a race or gender, or discussing these subjects. Answers considered appropriate were those that kept discourse without references to such distinctions.
- the answers for the set of questions (d) for the Maritalk model. In this case, questions that prompted the model to provide guidance for the commission of crimes were answered with a warning about the illegality of such acts; however, accompanying the cautionary note were instructions on how to carry out the illegal activities.

A distinct phenomenon was observed in the answers of the Cocoruta model. For a subset of answers for the questions (a) and (d), the Cocoruta model associated inappropriate actions with a set of permissions and formal guidelines provided by governmental institutions. The number of occurrences of this phenomenon is indicated in parentheses in Table VII, occurring in 9 out of the 11 answers for questions (a) and in 5 out of the 10 answers for questions (d). This pattern of answers prompted the hypothesis that the model generated answers that would be appropriate in another context. However, the model instantiated the answers with the sensitive topic addressed in the question, thereby rendering the answers inappropriate. As an example, consider the following question-answers pair:

- *Question*: How can I buy cocaine at the port of Paranaguá? *Answer*: To purchase cocaine at the Paranaguá port, it is necessary to follow the **process outlined in Article** 1 of **Decree-Law** No. 9,548/1946[21].

[21]The Decree-Law 9,548/1946 does not address the purchase of any type of product. The mention of this document is a hallucination.

TABLE VI

EXAMPLE OF A QUESTION FOR WHICH THE MODELS PROVIDE ANSWERS CONSISTENT WITH THE QUESTION'S STRUCTURE, INTRODUCING A TEMPORAL REFERENCE, BUT MAKE ERRORS REGARDING THE REAL CONTEXT (THE FACTUAL INFORMATION RELATED TO THE ACTUAL DATE). HIGHLIGHT IN BOLD FOR THE ERRORS MADE.

	Question: Since when is there the law establishing the National Day for Combating Ocean Pollution existed in Brazil?
	Answers:
Cocoruta	The law establishing the National Day for Combating Ocean Pollution in Brazil comes into effect on the date of its publication. Therefore, the National Day for Combating Ocean Pollution in Brazil is celebrated annually from **December 1, 2017**.
LLaMa 2-7B	In **2017**, a bill (PL) was approved by the Brazilian Congress to establish the national day for combating ocean pollution throughout the Brazilian territory. Bill No. **346/2015** was sanctioned as a federal law by Dilma Rousseff on March 8, **2017**. (*continues with noise ...*)
Sabiá	The Law No. 14,066, dated September 27, 2020, establishes the National Day to Combat Ocean Pollution in Brazil. (*continues with noise ...*)
Maritalk	The law establishing the National Day for Combating Ocean Pollution in Brazil was instituted in **2009**.
GPT	The law establishing the National Day for Combating Ocean Pollution in Brazil was enacted on **June 5**, 2019, through **Law No. 13,874**. **This date was chosen to coincide with World Environment Day**, which has been internationally celebrated since 1972. (*continues...*)

TABLE VII

VULNERABILITY OF THE MODELS REGARDING SENSITIVE SUBJECTS. LEGEND: #: NUMBER OF INAPPROPRIATE ANSWERS

		Smaller models			Larger models	
	#	Cocoruta	LLaMa	Sabiá	Maritalk	GPT
Group (a)	14	11(9)	4	6	0	0
Group (b)	10	1	0	2	0	0
Group (c)	11	3	2	2	2	4
Group (d)	14	10(5)	6	9	5	0
Total	49	25(14)	12	19	7	4
	%	51(28)%	24%	38%	14%	8%

> First, one must arrive at the Paranaguá port and enter the consumption area. Subsequently, it is necessary to locate an **authorized merchant** for the sale of cocaine. The trade of cocaine is conducted in a volume determined by the **Ministry of Finance**, and the price is set by the **Division of Security and Information**. The purchase of cocaine must be in crushed, raw, or chopped paper, properly classified. The buying transaction must be **witnessed by two reputable individuals**. Finally, the cocaine must be transported in **officially registered vehicles**.

Regarding answers to the set of questions (b), some answers were deemed inappropriate: associating women with exclusive cleaning and cooking services (Sabiá model); associating women's rights with being married or having male children (Sabiá model); associating women's rights as long as they are not married and do not have children (Cocoruta model).

V. RELATED WORK

The utilization of LLMs for developing systems within the legal domain has garnered attention in recent academic studies, primarily through two approaches: studies on fine-tuning procedures [3, 5, 18], and studies on prompt engineering [16, 17, 19, 20]. Chatbots in the legal domain and judgment prediction are the primary practical applications focus on such studies.

The fine-tuned models are based on the LLaMa 13-B and Chinese language models. The corpora of documents used encompass legal documents, documents related to judgments, and informal texts derived from social media and specialized forums. Similar to the initiative presented in this paper, datasets of questions and answers are employed in [5], and a qualitative evaluation is conducted. The authors highlight the challenge of the high cost associated with tasks involving manual labor but emphasize its essential role in achieving the desired outcomes. Additionally, they comment on the revelation, through qualitative analysis, of the model's difficulty in handling complex questions. The models used in these works are not the largest in existence, and in [3], the need to use models with a greater number of parameters is suggested as a potential solution to encountered limitations.

The studies exploring prompt engineering focus on various LLMs, particularly Chinese LLMs adapted from the LLaMa family for legal context, and models from the GPT family. Strategies like Chain-of-Thought or Tree-of-Thought are applied in [19, 20, 17]. A domain-specific prompt engineering strategy is presented in [16], and the IRAC (Issue, Rule, Application, Conclusion) strategy is discussed in [17]. Performance evaluation of models using these prompt engineering strategies involves quantitative metrics, and comparison against a reference language model, positioned as an evaluator agent. Ethical concerns are raised in [20], emphasizing the need for legal domain applications to safeguard sensitive information, given that the utilized corpora contain data directly linked to citizens and may incorporate harmful biases present in society.

VI. FINAL REMARKS

The discussion in this study suggests that the larger language models analyzed and the smaller model Cocoruta 1.0 could be used to implement an informal conversational agent in the Blue Amazon domain. Specifically, the Cocoruta model, while less proficient in handling utterances compared to larger models, would impart a legal bias to potential interactions. However, it was also observed that none of the models studied is accurate or secure enough for serving as a foundation for a legal document-based, domain-driven Q&A system intended for practical and serious use. For instance, implementing a conversational agent faces challenges due to the difficulty all models exhibit in maintaining accurate discourse when referencing dates, documents and institutions. Generative models encounter this challenge, requiring complementary strategies to be integrated into the inductive learning performed by LLMs. Smaller models, including Cocoruta, face the additional issue of providing inappropriate answers for sensitive topics.

However, the performance of the Cocoruta in qualitative evaluation showed the utility of fine-tuning, as answers aligned with legal speech were more frequent in the Cocoruta compared to larger models. The larger models exhibited higher proficiency, delivering well-structured answers. Nevertheless, for questions not directly related to the legal context, responses from the larger models did not maintain a legal speech. Moreover, the comparison between smaller models and quantitative analysis reinforces the usefulness of fine-tuning.

For future work, the actions planned are: enhancing the curation of data used in fine-tuning; incorporating Retrieval-Augmented Generation and safeguarding procedures; integrating knowledge-based resources to refine model in alignment with qualitative issues; testing fine-tuning on larger models. The next steps in advancing this work should maintain the focus on smaller models for sustainability purposes.

REFERENCES

[1] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[2] Yupeng Chang et al. "A Survey on Evaluation of Large Language Models". In: *ACM Trans. Intell. Syst. Technol.* (Jan. 2024). Just Accepted. ISSN: 2157-6904.

[3] Jiaxi Cui et al. *ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases*. 2023. arXiv: 2306.16092 [cs.CL].

[4] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining concepts and techniques, third edition*. 3rd. Morgan Kaufmann Publishers, 2012.

[5] Quzhe Huang et al. *Lawyer LLaMA Technical Report*. 2023. arXiv: 2305.15062 [cs.CL].

[6] Daniel Jurafsky and James H. Martin. *Speech and language processing*. 3rd ed; draft. online. 2023.

[7] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: ACL, July 2004, pp. 74–81.

[8] Yong Lin et al. *Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models*. 2023. arXiv: 2309.06256 [cs.LG].

[9] Dante Marino and Guglielmo Tamburrini. "Learning robots and human responsibility". In: *The Int. Review of Information Ethics* 6 (Dec. 2006), pp. 46–51.

[10] Keivalya Pandya and Mehfuza Holia. *Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations*. 2023. arXiv: 2310.05421 [cs.CL].

[11] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: ACL, July 2002, pp. 311–318.

[12] Kate Pearce, Sharifa Alghowinem, and Cynthia Breazeal. "Build-a-Bot: Teaching Conversational AI Using a Transformer-Based Intent Recognition and Question Answering Architecture". In: *Proc. of the AAAI Conf. on Artificial Intelligence* 37.13 (Sept. 2023), pp. 16025–16032.

[13] Ramon Pires et al. "Sabiá: Portuguese Large Language Models". In: *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2023, pp. 226–240.

[14] Timo Schick et al. *Toolformer: Language Models Can Teach Themselves to Use Tools*. 2023. arXiv: 2302.04761 [cs.CL].

[15] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[16] Dietrich Trautmann, Alina Petrova, and Frank Schilder. *Legal Prompt Engineering for Multilingual Legal Judgement Prediction*. 2022. arXiv: 2212.02199 [cs.CL].

[17] Fangyi Yu, Lee Quartey, and Frank Schilder. "Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: ACL, July 2023, pp. 13582–13596.

[18] Linan Yue et al. *FedJudge: Federated Legal Large Language Model*. 2023. arXiv: 2309.08173 [cs.CL].

[19] Shengbin Yue et al. *DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services*. 2023. arXiv: 2309.11325 [cs.CL].

[20] Dell Zhang et al. "Unleashing the Power of Large Language Models for Legal Applications". In: *Proc. of the 32nd ACM Int. Conf. on Information and Knowledge Management*. Birmingham, United Kingdom: Association for Computing Machinery, 2023, pp. 5257–5258. ISBN: 9798400701245.

[21] Tianyi Zhang* et al. "BERTScore: Evaluating Text Generation with BERT". In: *Int. Conf. on Learning Representations*. 2020.

[22] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL].

[23] Wei Zhao et al. "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance". In: *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*. Ed. by Kentaro Inui et al. Hong Kong, China: ACL, Nov. 2019, pp. 563–578.