Evaluating Large Language Models for Tax Law Reasoning

João Paulo Cavalcante Presa[0009−0004−4160−6495], Celso Gonçalves Camilo Junior[0000−0003−2553−8790], and Sávio Salvarino Teles de Oliveira[0009−0002−1203−5246]

Federal University of Goias (UFG)
joaopaulop@discente.ufg.br
celsocamilo@ufg.br
savioteles@ufg.br

Abstract. The ability to reason over laws is essential for legal professionals, facilitating interpreting and applying legal principles to complex real-world situations. Tax laws are crucial for funding government functions and shaping economic behavior, yet their interpretation poses challenges due to their complexity, constant evolution, and susceptibility to differing viewpoints. Large Language Models (LLMs) show considerable potential in supporting this reasoning process by processing extensive legal texts and generating relevant information. This study evaluates the performance of LLMs in legal reasoning within the domain of tax law for legal entities, utilizing a dataset of real-world questions and expert answers in Brazilian Portuguese. We employed quantitative metrics (BLEU, ROUGE) and qualitative assessment using a solid LLM to ensure factual accuracy and relevance. A novel dataset was curated, comprising genuine questions from legal entities in tax law, answered by legal experts with corresponding legal texts. The evaluation includes both open-source and proprietary LLMs, providing a assessment of their effectiveness in legal reasoning tasks. The strong correlation between robust LLM evaluator metric and Bert Score F1 suggests these metrics effectively capture semantic aspects pertinent to human-perceived quality.

Keywords: Legal Reasoning · Large Language Models (LLMs) · Legal Question Answering · Tax Law.

1 Introduction

The ability to reason over laws is essential for legal professionals, enabling them to interpret and apply legal principles to complex real-world situations. Legal questions often lack straightforward answers, requiring thorough analysis, comprehensive research, and synthesis of multiple sources to develop well-founded arguments or solutions. Tax law, in particular, is crucial because it influences how governments fund public services and impacts economic activity by shaping investment decisions and individual spending. However, interpreting tax law presents significant challenges for Natural Language Processing (NLP) due to the inherent complexity and ambiguity of legal language, constant updates, amendments, and the need to contextualize regulations within specific jurisdictions.

Large Language Models (LLMs) show significant potential in enhancing the legal reasoning process [23]. These models can process extensive legal texts, including statutes, case law, and legal opinions, to extract relevant information and address the complexities of tax law. By leveraging advanced generation techniques, LLMs can answer legal questions using specific legal datasets, such as court cases and legal precedents, thus providing comprehensive and relevant information to legal professionals [29]. However, there is a gap in understanding how LLMs reason over legal texts, as existing question-answering tasks typically contain answers directly extractable from the provided texts, whereas legal reasoning often requires deeper comprehension and application of legal principles to nuanced scenarios [2].

To address this gap, we developed a novel dataset comprising real questions posed by legal entities in the domain of tax law, answered by legal experts with supporting legal texts (gold passages). This dataset allows us to assess the legal reasoning abilities of LLMs, focusing on their capacity to understand complex legal questions, use relevant law articles, and generate accurate and coherent responses. The evaluation compares LLM-generated answers to expert responses using metrics such as ROUGE [19], BLEU [27], and semantic similarity [40], alongside assessments by a strong LLM [41], contributing to a understanding of LLMs' legal reasoning capabilities.

This research evaluates both open-source and proprietary LLMs in scenarios requiring comprehensive understanding and application of the law, distinct from the extractive approaches used in datasets like SQuAD [28] and TriviaQA [14]. Our dataset requires LLMs to comprehend and apply the law to generate appropriate answers, often involving complex vocabulary and contexts not directly mirrored in the texts [2].

This paper presents two significant contributions to the field of legal NLP, particularly within the challenging domain of tax law. First, it introduces a novel dataset consisting of real-world tax law questions, expert-crafted answers, and supporting legal texts, moving beyond extractive question-answering tasks and requiring models to demonstrate legal reasoning abilities. Using this dataset, we evaluate how well LLMs understand complex tax law questions and generate accurate, well-supported answers, providing a better understanding of current LLM capabilities and limitations in handling legal reasoning tasks.

2 Related works

While there are numerous works utilizing Large Language Models (LLMs) in the legal domain [5,22,23,25,39], our interest lies in those that apply LLMs for question and answer (Q&A) tasks. These works can be classified into three main categories: those using retrieval-augmented generation, those evaluating LLMs based on prior knowledge, and those performing fine-tuning and testing the models on Q&A tasks in the legal domain. Below, we discuss the key works found in each of these categories.

2.1 Tax Law Applications of LLMs

Evaluating Q&A LLMs with Retrieval-Augmented Generation The LLeQA [21] dataset includes 1,868 legal questions annotated by experts, containing answers and legal references. This work applies the Retrieval-Augmented Generation (RAG) technique, retrieving statutory articles from an extensive corpus of Belgian legislation. The model's effectiveness is evaluated using the METEOR metric, demonstrating the feasibility of integrating information retrieval with LLMs to enhance the accuracy of legal responses. ChatLaw [5] addresses the creation of a large-scale language model for the legal domain, specifically in the Chinese context. This work combines vector database retrieval methods with keyword-based retrieval to increase the accuracy of responses. Integrating these techniques enables the model to provide more precise and contextually relevant answers.

Evaluating Q&A Legal Reasoning of LLMs LAiW [6] proposes a benchmark for evaluating the capabilities of LLMs in the Chinese legal context. The aim of this work is to test how well models can handle specific legal tasks. The results show that some legal-specific LLMs perform better than their general counterparts, although there remains a significant gap compared to GPT-4 [26]. LawBench [8] offers a comprehensive assessment of LLM capabilities in legal tasks, including Q&A. This work extensively tested 51 popular LLMs, including 20 multilingual, 22 focused on Chinese, and 9 specific to law. The conclusion is that while fine-tuning LLMs on specific legal texts brings some improvements, the models still need to be usable and reliable for complex legal tasks.

Fine-tuning and Evaluating Q&A Large Language Models FedJudge [37] uses Federated Learning (FL) to overcome data privacy challenges. This framework optimizes federated legal LLMs, allowing the models to be trained locally on clients, with their parameters aggregated and distributed on a central server. FedJudge is evaluated on Q&A tasks using metrics such as ROUGE, BLEU, and BertScore to compare the quality of generated answers. This work demonstrates that the model provides more precise and relevant answers in different legal contexts. DISC-LawLLM [38] employs large language models trained on supervised datasets in the legal domain and incorporates a retrieval module to access and utilize external legal knowledge. This system assesses objective and subjective perspectives using DISC-Law-Eval, a benchmark that includes legal question answering. Additionally, subjective evaluation is carried out using the GPT-3.5 model as a judge.

3 Metodology

This section outlines the methodology utilized in our study, with a particular emphasis on the model selection process. We also detail the data collection process, the creation of a relevant corpus, and the experimental setups of the selected models, including the specific prompts and parameters used. Furthermore, we detail the evaluation approach, discussing both the metrics employed and the strategy for subjective evaluation.

3.1 Dataset Collection

Our dataset consists of a series of tax law questions related to legal entities. The questions were selected from a collection that is annually updated by the General Coordination of Taxation (Cosit) [9] of the Brazilian Federal Revenue Service. The dataset includes over a thousand question-answer pairs, with most answers being supported by a relevant normative or legal basis. The granularity of the references in the answers is as detailed as possible, citing the specific articles of law or other regulations used to formulate the responses. The questions represent real taxpayer doubts, and experts in the Brazilian tax field craft the answers. Below, we will discuss how the dataset was created.

Selection of Questions We extracted a subsample from the comprehensive set of questions and answers provided by Cosit. In this selection process, we focused on questions that included responses with legal references rather than the entire regulation. Although the majority of responses included legal references, they were often elaborated by experts in a way that extended beyond the scope of the question or included excessive details such as tables and numerous examples. This complexity made them unsuitable for use in contexts like Retrieval-Augmented Generation (RAG). We excluded these overly detailed responses to ensure a fair evaluation with the LLMs. The initial outcomes of this selection process are depicted in the first three columns of Table 1.

Collection of Regulations (Gold Passages) After selecting the questions and their corresponding legal references, laws, and articles, we gathered each regulatory document referenced by the experts in their responses to the questions posed by legal entities. Although this task was time-intensive, it was essential for assessing the reasoning capabilities of LLMs in relation to legal texts. Upon completion of this stage, the dataset comprised the question, answer, reference to the regulation, and the regulation itself (gold passages). Table 1 presents the final dataset.

Legislation Corpus In this stage, we collected over 30 documents, which included laws, instructions, decrees, and opinions. Each document contains up to thousands of articles comprising multiple provisions. These documents represent

Table 1: Questions and Answers with Legal References and Gold Passages

Question	**Answer**	**Reference**	**Gold Passages**
Which legal entities are exempt from presenting the ECF?	The following are exempt from presenting the ECF: I - those...	IN RFB No. 2004, 2021, art. 1, § 1.	Art. 1 The Fiscal Accounting Bookkeeping (ECF) shall...
What are the tax effects in case the ECF is corrected?	When the correction of the ECF shows a higher tax due...	IN RFB No. 2055, 2021, art. 148.	Art. 148. The credit related to tax administered...
Does the exemption from IRPJ depend on prior recognition?	No. The benefit of the IRPJ exemption does not depend on...	RIR/2018, art. 192.	Art. 192. The exemptions referred to in this Section...
Under what circumstances is an individual considered equivalent to a legal entity?	For income tax purposes, individuals are considered...	RIR/2018, art. 162, § 1, items I to III.	Art. 162. Individual enterprises are considered...
Are co-owners in property ownership subject to income tax?	Condominiums in property ownership are not subject to...	RIR/2018, art. 167.	Art. 167. Condominiums in property ownership shall...

a fraction of the Brazilian tax legislation and include the regulations that underpin the experts' responses in the dataset. It is important to note that these regulations are constantly being amended, and many provisions have been revoked. All revoked provisions were excluded up to the dataset creation date to ensure a high-quality corpus. Additionally, any questions that had their regulatory basis revoked were eliminated during the question selection phase. Figure 1 shows the corpus documents.

3.2 Experimental Setup

In this study, we conducted a comprehensive evaluation of large language models (LLMs) in terms of their ability to reason about laws, specifically focusing on corporate taxation for legal entities. We evaluated the LLMs using the datasets created in this paper, which were created from real-world questions and answers about the taxation of legal entities, which were provided by subject-matter experts.

We selected over 20 LLMs for evaluation, encompassing both proprietary and open-source models. The chosen models include notable examples such as Mistral AI, Llama, Gemma, Qwen, various community fine-tuned versions of these models, and a proprietary model. Each model possesses unique characteristics and capabilities, providing a diverse range of perspectives for our assessment.

ADI SRF nº 5, de 2001
ADN Cosit nº 4, de 1996
Decreto-Lei nº 1.381, de 1974
Decreto-Lei nº 1.510, de 1976
Decreto-Lei nº 1.598, de 1977
IN DPRF 21, de 1992
IN RFB nº 1.252, de 2012
IN RFB nº 1.520, de 2014
IN RFB nº 1.700, de 2017
IN RFB nº 2.004, de 2021
IN RFB nº 2.055, de 2021
IN SRF nº 213, de 2002
IN SRF nº 51, de 1978
IN nº 122, de 1989
Lei nº 6.404, de 1976
Lei nº 6.766, de 1979
Lei nº 9.249, de 1995
Lei nº 9.316, de 1996
Lei nº 9.430, de 1996
Lei nº 9.532, de 1997
Lei nº 9.718, de 1998
Lei nº 11.051, de 2004
PN CST nº 1, de 1983
PN CST nº 2, de 1983
PN CST nº 4, de 1981
PN CST nº 58, de 1977
PN CST nº 72, de 1975
PN CST nº 146, de 1975
Portaria MF nº 356, de 1988
RIR 2018

Fig. 1: Documents from the legislative corpus

In order to maintain consistency in our evaluations, we standardized the temperature parameter at 0.1 for all chosen models. This low-temperature setting was chosen to reduce randomness in the output, thereby encouraging more deterministic responses. Additionally, we did not impose a maximum token limit, allowing the models to generate responses without any constraints on their length.

A specific prompt (see Prompt Question Answer in Appendix A) was crafted to guide the models in reasoning about the law and generating appropriate responses. The prompt explicitly instructs the models to reason through the legal context provided and formulate an answer. If a model is unable to generate a satisfactory response, it is instructed to state that it does not know the answer.

The legal information needed to answer each question, like an article from a law or a legal document, is included in the prompt. This information is the same one used by the experts to create the reference answers, ensuring a fair basis for comparison. By using standardized prompts and incorporating relevant legal provisions, it ensures that the models have access to the same information as human experts. This enables a thorough evaluation of their reasoning capabilities.

It is important to note that the questions and the reference answers are presented in Brazilian Portuguese. This aspect of the study tests the models' reasoning abilities and evaluates their proficiency in generating accurate and contextually appropriate responses in the Portuguese language. Given that many LLMs are primarily trained on English-language datasets, assessing their performance on Brazilian Portuguese legal texts for understanding the applicability and limitations of these models in non-English-speaking jurisdictions.

Although the dataset used in this experiment contains a corpus suitable for Retrieval-Augmented Generation (RAG), our evaluation focused solely on the tasks of generation and reasoning. This decision was inspired by other prominent datasets, such as SQuAD 2.0 and HotpotQA, which also provide the expected passages alongside the ground truth answers, allowing for a direct assessment of the model's generation capabilities without the retrieval step. By concentrating on these aspects, we aimed to isolate and thoroughly evaluate the LLMs' ability to generate accurate and reasoned responses based solely on the provided legal context.

3.3 Evaluation Metrics

Our study evaluated large language models (LLMs) using a comprehensive approach integrating quantitative and qualitative methods. For the quantitative assessment, we employed the BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics [19,27]. In the field of natural language processing, these metrics are crucial for assessing the quality of text generation by comparing the models' responses to a predefined set of reference answers. Specifically, in the domain of questions and answers related to corporate taxation, these metrics provide a quantitative measure of how closely the generated responses align with the ideal answers.

Despite their widespread use, metrics such as BLEU, ROUGE [37], and METEOR [21], which are widely used for evaluating language models, they primarily provide a quantitative perspective and may not fully capture the accuracy of responses in question-answering scenarios [20]. This limitation arises because these metrics do not adequately assess the factual accuracy or the relevance of the generated responses, which are critical in determining whether the questions were answered correctly.

In order to address this gap, we adopted a more nuanced qualitative approach, utilizing the capabilities of a powerful language model as a surrogate for human judgment. Specifically, we employed GPT-4 to evaluate the performance of other models. This approach is premised on the notion that a robust LLM, such as GPT-4, can effectively emulate human judgment in evaluating responses [7,10,17,20,31,35,41] to open-ended questions, thereby providing a closer approximation to human evaluative criteria.

For the qualitative evaluation, we used a carefully designed prompt to assess the factual accuracy of the models' responses. The accuracy of each model was then calculated based on this assessment. The specific prompt used for this evaluation can be found in Prompt Evaluation in the Appendix A for further details.

4 Results

In this section, we will present the results of the experiments described in the Experimental Setup section. We used the metrics outlined in the Metrics section, both of which are situated in the Methodology (Section 3). We selected the main language models operating in Portuguese to evaluate how well they can reason and answer questions in the context of tax law, with the aim of identifying potential improvements.

Table 2: Model Performance Metrics

Model	ROUGE-L	BLEU	Bert Score F1	Acc. GPT-4
Mistral-7B-Instruct-v0.2	0.35	0.20	0.67	0.54
Mistral-7B-Instruct-v0.3	0.40	0.26	0.71	0.55
Mixtral-8x7B-Instruct-v0.1	0.38	0.24	0.70	0.53
Mixtral-8x22B-Instruct-v0.1	**0.44**	**0.30**	**0.73**	0.59
Llama-2-70b-chat-hf	0.38	0.19	0.69	0.49
Llama-2-13b-chat-hf	0.37	0.20	0.68	0.43
Llama-2-7b-chat-hf	0.32	0.14	0.65	0.34
Llama-3-70b-chat-hf	0.34	0.16	0.65	0.60
Llama-3-8b-chat-hf	0.35	0.15	0.65	0.54
Qwen1.5-110B-Chat	0.39	0.21	0.71	0.60
Qwen1.5-72B-Chat	0.41	0.24	0.71	0.62
Qwen1.5-14B-Chat	0.34	0.16	0.68	0.48
Qwen2-72B-Instruct	0.43	0.29	**0.73**	**0.64**
gemma-7b-it	0.40	0.22	0.70	0.45
Yi-34B-Chat	0.38	0.26	0.70	0.52
gpt-3.5-turbo	0.38	0.15	0.69	0.56
Platypus2-70B-instruct	0.41	0.29	0.70	0.57
vicuna-13b-v1.5	0.41	0.27	0.71	0.50
vicuna-7b-v1.5	0.37	0.23	0.69	0.39
openchat-3.5-1210	0.42	0.28	0.72	0.55
WizardLM-13B-V1.2	0.36	0.25	0.68	0.49
SOLAR-10.7B-Instruct-v1.0	0.36	0.23	0.70	0.51
OpenHermes-2p5-Mistral-7B	0.41	0.25	0.71	0.55

4.1 Model Performance Analysis

The latest versions of the Llama, Qwen, and Mistral families exhibit significant advancements compared to their predecessors. These models incorporate several architectural enhancements, including SwiGLU activation [30] and Grouped Query Attention (GQA) [4]. Both the Qwen2-72B-Instruct [1] and Llama-3-70b-chat-hf [3] models benefited from these improvements, particularly the modifications to the tokenizer and the inclusion of GQA, leading to notable performance gains. As a result, the Qwen2-72B-Instruct [1] model achieved the highest accuracy. Similar results have been observed in other LLM evaluation benchmarks [1], highlighting the superior performance of models incorporating these techniques.

The performance analysis of the models revealed that model size significantly impacts the results, but this impact is not always straightforward. Larger models, such as Qwen2-72B-Instruct [1] and Mixtral-8x22B-Instruct-v0.1 [13], achieved superior performance, exhibiting the highest ROUGE-L, BLEU, Bert Score F1, and GPT-4 evaluated accuracy metrics. However, we observed that smaller models, such as Mistral-7B-Instruct-v0.3 [12] and OpenHermes-2p5-Mistral-7B [32], outperformed some larger models in specific metrics. For instance, Mistral-7B-Instruct-v0.3 attained a Bert Score F1 of 0.71, surpassing several larger models, and OpenHermes-2p5-Mistral-7B demonstrated remarkable performance with accuracy comparable to significantly larger models. These findings suggest that while larger models generally deliver better results due to their ability to capture more complex information, well-trained and fine-tuned smaller models can offer competitive performance in specific contexts. This trend indicates that the training quality and the model's suitability to the particular dataset are crucial factors that can mitigate the size disparity among models.

Although the volume of Portuguese data used in training these models has yet to be verified, the architectural and training improvements suggest the enhanced performance of the LLMs in Q&A tasks in corporate tax law. In the Mistral family, the Mixtral-8x22B-Instruct-v0.1 [24] model stood out with the highest scores in ROUGE-L, BLEU, and Bert Score F1, indicating the potential of the mixture of experts architecture [13] for legal texts in Portuguese.

The analysis of fine-tuned open-source models reveals significant improvements over the base models. The openchat-3.5-1210 [34] and OpenHermes-2p5-Mistral-7B [32], both derived from the Mistral-7B-v0.1 [12], showed notable increases in accuracy after fine-tuning. Similarly, the vicuna-13b-v1.5 and vicuna-7b-v1.5 models [41], fine-tuned from Llama 2 [33], also demonstrated advances in response accuracy. Furthermore, models such as WizardLM-13B-V1.2 [36], SOLAR-10.7B-Instruct-v1.0 [15,16], and Platypus2-70B-instruct [18], derived from Llama 2, improved the results of their base models. Notably, these fine-tuning processes were conducted on diverse datasets, not on the experimental dataset itself, yet still led to enhanced metrics within the experimental dataset. These improvements suggest that fine-tuning can effectively enhance the capabilities of Q&A and legal text generation tasks when applied to specific datasets.

4.2 Evaluation Metrics Analysis

The traditional metrics like BLEU and ROUGE may not fully capture the nuances needed for accurate question answering in Q&A tasks. The Bert Score F1 metric is extensively recognized for its alignment with human evaluation due to its capacity to capture deep semantic similarities between texts [40], surpassing the lexical matching capabilities of traditional metrics like ROUGE-L and BLEU [40].

While this study does not aim to prove that LLM as a judge for evaluation is aligned with human evaluation, recent studies have been exploring this alignment [20,31,35,41]. Our research evaluates the quality of responses generated by LLMs in legal domain Q&A tasks. The strong correlation between the LLM (GPT-4) Accuracy Evaluation and Bert Score F1, as evidenced by the Pearson (0.657) and Kendall (0.491) correlations (see Table 3), suggests that both metrics capture semantic aspects relevant to human-perceived quality. The results are in line with studies [40] recommending using Pearson and Kendall correlations to evaluate metric quality.

Furthermore, the Bland-Altman analysis, which is particularly suitable for comparing measurement methods [11], confirms that Bert Score F1 and LLM

Table 3: Correlation Matrices for Evaluation Metrics

Metrics	**Pearson**	**Kendall**
Accuracy (GPT4) ↔ Bert Score F1	0.658	0.491
Accuracy (GPT4) ↔ ROUGE-L	0.539	0.393
Accuracy (GPT4) ↔ BLEU	0.373	0.289
Bert Score F1 ↔ ROUGE-L	0.908	0.838
Bert Score F1 ↔ BLEU	0.781	0.666
ROUGE-L ↔ BLEU	0.821	0.670

(GPT-4) Accuracy Evaluation are more concordant, as shown by the lower variation in point dispersion and the narrower width of the limits of agreement (see Figure 2). In contrast, ROUGE-L and BLEU demonstrated higher mean differences and wider limits of agreement. Bert Score F1 exhibited a mean difference close to zero and narrower limits of agreement, indicating better concordance with LLM Accuracy Evaluation measurements.

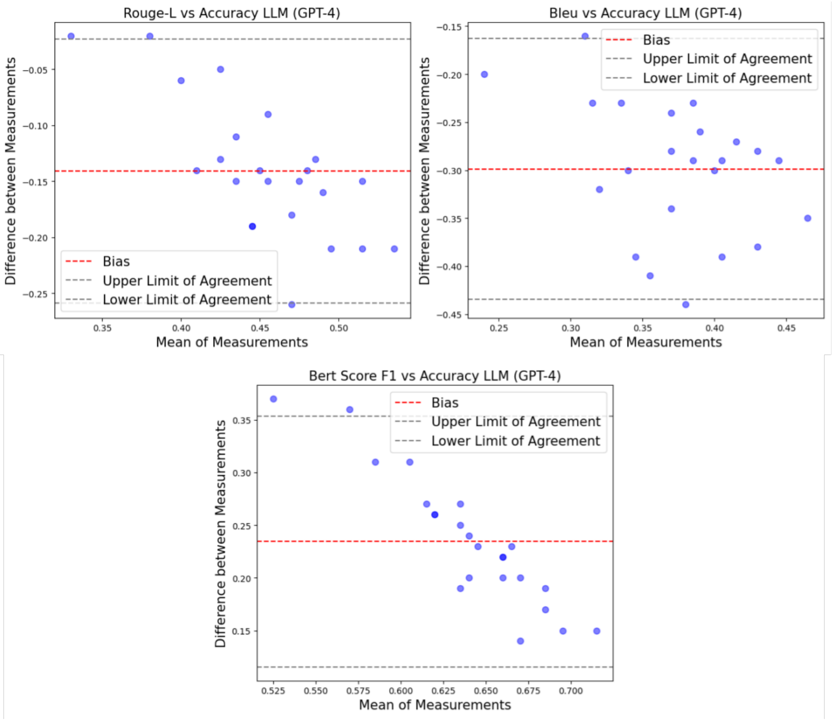

Fig. 2: Bland-Altman Plots for Metrics vs LLM (GPT-4) Accuracy Evaluation

These findings imply that LLM (GPT-4) Accuracy Evaluation, similar to Bert Score F1, could be a valuable and representative metric for assessing the actual performance of language models. While ROUGE-L and BLEU show higher correlations with Bert Score F1, the stronger correlation and concordance of LLM Accuracy Evaluation with Bert Score F1 indicate its potential alignment with human evaluation. This supports the development of evaluation metrics that more accurately reflect human-perceived quality, aligning with the direction of current research investigating the potential of LLMs used to align with human evaluation [20,31,35,41].

5 Conclusion

This study underscores the importance of tax law in society and the potential of language models to assist in its understanding and application. We developed a novel dataset of real-world tax law questions and expert answers in Brazilian Portuguese and conducted a rigorous evaluation of various language models. While our findings suggest that these models show promise in comprehending and reasoning about complex legal texts, further research is necessary to fully demonstrate their effectiveness in legal reasoning across a broader range of scenarios and tasks.

Our evaluation showed that advancements in model architecture have a noticeable impact on performance, and fine-tuning open-source models, even when done on diverse datasets rather than those specific to the legal domain, can still improve their ability to generate relevant and accurate responses. This suggests that continuous improvements and adaptations are valuable in enhancing the capabilities of language models in legal tasks.

For assessing model performance, we used Bert Score F1, known for its strong correlation with human evaluations in tasks involving descriptive and structural understanding, and a newer metric, LLM Accuracy Evaluation. While Bert Score F1 is already established as an effective measure aligned with human judgment, especially in descriptive tasks, our results showed that LLM Accuracy Evaluation also demonstrated strong correlation with Bert Score F1 through Pearson and Kendall correlations. The Bland-Altman analysis further confirmed that the LLM metric aligns closely with Bert Score, suggesting its potential as a reliable alternative in evaluations. However, it is important to note that while these findings are encouraging, the use of these metrics for reasoning-based tasks, such as those in this study, still requires further validation. The LLM metric is a promising tool, but more research is needed to establish its effectiveness fully, particularly in capturing the nuances of legal reasoning.

Limitations and Future Work A limitation of our study is that while it focused on evaluating the generation and reasoning capabilities of LLMs, it did not require the models to identify specific legal provisions as part of their responses. Our dataset includes a comprehensive corpus containing the necessary laws, enabling the application of Retrieval-Augmented Generation (RAG) techniques.

This allows models to retrieve relevant legal provisions and incorporate them into their answers, which is essential for a more complete statutory reasoning process. By providing the relevant legal articles, which do not directly mirror the answers to the questions, our study did assess a portion of the statutory reasoning by testing the models' ability to apply the law to generate accurate responses. Future work could leverage this corpus to explore the integration of RAG, aiming to enhance the models' ability to not only generate correct answers but also to identify and cite the appropriate legal provisions, thereby achieving a more robust and comprehensive statutory reasoning.

Dataset and Code Availability The dataset used in this study, as well as the code for reproducing the experiments and analyses, are publicly available. The dataset can be accessed at the following link: `https://github.com/joaopaulopresa/dataset`. The code can be found here: `https://github.com/joaopaulopresa/code`.

References

1. Qwen2 blog. `https://qwenlm.github.io/blog/qwen2/` (2024), accessed: 08/06/2024
2. Abdallah, A., Piryani, B., Jatowt, A.: Exploring the state of the art in legal qa systems. Journal of Big Data **10**(1), 127 (2023)
3. AI@Meta: Llama 3 model card (2024), `https://github.com/meta-llama/llama3/tree/main`
4. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebron, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4895–4901 (2023)
5. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)
6. Dai, Y., Feng, D., Huang, J., Jia, H., Xie, Q., Zhang, Y., Han, W., Tian, W., Wang, H.: Laiw: A chinese legal large language models benchmark (a technical report). arXiv preprint arXiv:2310.05620 (2023)
7. Du, Y., Wei, F., Zhang, H.: Anytool: Self-reflective, hierarchical agents for large-scale api calls. arXiv preprint arXiv:2402.04253 (2024)
8. Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., Ge, J.: Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289 (2023)
9. General Coordination of Taxation (Cosit): Questions and answers for legal entities 2023. `https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/declaracoes-e-demonstrativos/ecf/perguntas-e-respostas-pj-2023.pdf` (2023), accessed: 11/11/2023
10. Hackl, V., Müller, A.E., Granitzer, M., Sailer, M.: Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In: Frontiers in Education. vol. 8, p. 1272229. Frontiers Media SA (2023)
11. Haghayegh, S., Kang, H.A., Khoshnevis, S., Smolensky, M.H., Diller, K.R.: A comprehensive guideline for bland–altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. Physiological measurement **41**(5), 055012 (2020)

12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
13. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
14. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1601–1611 (2017)
15. Kim, D., Kim, Y., Song, W., Kim, H., Kim, Y., Kim, S., Park, C.: sdpo: Don't use your data all at once (2024)
16. Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., Kim, S.: Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling (2023)
17. Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J., Denny, P.: Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. arXiv preprint arXiv:2405.05253 (2024)
18. Lee, A.N., Hunter, C.J., Ruiz, N.: Platypus: Quick, cheap, and powerful refinement of llms (2023)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
20. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
21. Louis, A., van Dijck, G., Spanakis, G.: Interpretable long-form legal question answering with retrieval-augmented large language models. arXiv preprint arXiv:2309.17050 (2023)
22. Ma, S., Chen, C., Chu, Q., Mao, J.: Leveraging large language models for relevance judgments in legal case retrieval. arXiv preprint arXiv:2403.18405 (2024)
23. Martin, L., Whitehouse, N., Yiu, S., Catterson, L., Perera, R.: Better call gpt, comparing large language models against lawyers. arXiv preprint arXiv:2401.16212 (2024)
24. Mistral.ai: Introducing the mixtral-8x22b-instruct-v0.1 model (2024), `https://mistral.ai/news/mixtral-8x22b/`, accessed on: 15 May 2024
25. Niklaus, J., Zheng, L., McCarthy, A.D., Hahn, C., Rosen, B.M., Henderson, P., Ho, D.E., Honke, G., Liang, P., Manning, C.: Flawn-t5: An empirical examination of effective instruction-tuning data mixtures for legal reasoning. arXiv preprint arXiv:2404.02127 (2024)
26. OpenAI: Gpt-4 technical report. `https://cdn.openai.com/papers/gpt-4.pdf` (2023), accessed: 10/01/2024
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
28. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392 (2016)
29. S, L.J.G.W.W.J.Q.Z.Y.P.: Large language models in law: A survey (11 2023), `https://arxiv.org/abs/2312.03718`

30. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
31. Sottana, A., Liang, B., Zou, K., Yuan, Z.: Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
32. Team, I.: Openhermes-2-5-mistral-7b (2024), `https://github.com/inferless/OpenHermes-2-5-Mistral-7B`, accessed on: 15 May 2024
33. Touvron, H., A.: Llama 2: Open foundation and fine-tuned chat models (2023)
34. Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., Liu, Y.: Openchat: Advancing open-source language models with mixed-quality data. In: The Twelfth International Conference on Learning Representations (2023)
35. Wei, F., Chen, X., Luo, L.: Rethinking generative large language model evaluation for semantic comprehension. arXiv e-prints pp. arXiv–2403 (2024)
36. Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Jiang, D.: Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244 (2023)
37. Yue, L., Liu, Q., Du, Y., Gao, W., Liu, Y., Yao, F.: Fedjudge: Federated legal large language model. arXiv preprint arXiv:2309.08173 (2023)
38. Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Lin, W., et al.: Disc-lawllm: Fine-tuning large language models for intelligent legal services. arXiv preprint arXiv:2309.11325 (2023)
39. Zhang, R., Li, H., Wu, Y., Ai, Q., Liu, Y., Zhang, M., Ma, S.: Evaluation ethics of llms in legal domain. arXiv preprint arXiv:2403.11152 (2024)
40. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert (2020)
41. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)

A Appendix: List of Prompts Used

This appendix provides a list of the prompts utilized throughout the experiments detailed in this study. The inclusion of these prompts ensures the transparency and reproducibility of the experimental procedures, allowing other researchers to replicate the study and verify the findings.

Prompt Question Answer:

```
Use the following pieces of legal information from laws to answer the user’s question.
If the answer is not clear in context, try to figure out by interpreting the information.
If you don’t know the answer, just say that you don’t know, don’t try to make up an answer.
Context: {context}
Question: {question}
Do not quote the "contextual information" provided in the answer, do not say "according to the
       information" or anything like that, use the information only to answer the question.
Only return the helpful answer below and nothing else.
Answer the question in Portuguese.
Helpful answer:
```

Prompt Evaluation:

```
Evaluate the AI-generated response based on the following criteria:
1. Verify if the AI Response is contained within the Expert Response, meaning there are no
      contradictions. Ignore different terms or small additional or missing information.
2. The Expert Response may contain more information than requested in the question. If the information
       in the Expert Response is not necessary to answer the question, do not use it to evaluate the
      AI Response.
3. If the AI Response contains more information than the Expert Response, it should not be considered
      for evaluation as long as the information is correct.
4. Check if the response answers the question. Ensure the response provides the information requested
      in the question and is sufficient. For example, if the question can be answered with a simple "
      No," that is acceptable.
Include reasoning that justifies the Evaluation. If the criteria are met, return 'CORRECT.' If any of
      the criteria are not met, return 'WRONG.'
The Evaluation should be a JSON object with keys 'result' and 'reasoning.'
Examples:
1.
### Question:
Are the earnings from technical consulting services provided by a legal entity domiciled in Brazil to
      its parent company abroad subject to transfer pricing legislation?
### Expert Response:
Firstly, it is necessary to distinguish whether the provision of services in Brazil involved
      technology transfer. If technology transfer is proven with the consent of the National Institute
       of Industrial Property (INPI), the transaction will not be subject to transfer pricing rules as
       established by art. 55 of IN RFB No. 1,312, of 2012. In this case, the deduction of such
      expenses is subject to the limits established by arts. 362 to 365 of RIR/2018. If there is no
      technology transfer, these services become subject to transfer pricing rules.
### AI Response:
Yes, they are subject to transfer pricing unless there is technology transfer with INPI consent.
### Evaluation:
{{
  "reasoning": "The AI response aligns with the expert's response, correctly addressing the question
        without contradictions, although it is shorter.",
  "result": "CORRECT"
}}
2.
### Question:
What should be considered as 'accrued considerations'?
### Expert Response:
For the purposes of art. 175 of Normative Instruction RFB No. 1,700, of 2017, accrued considerations
      are considered due considerations.
### AI Response:
Due considerations.
### Evaluation:
{{
  "reasoning": "The AI response covers the main points mentioned by the expert without presenting
        contradictions, though it is less detailed.",
  "result": "CORRECT"
}}
3.
### Question:
Is there a deadline for offsetting rural activity tax losses?
### Expert Response:
There is no deadline for offsetting rural activity tax losses.
### AI Response:
The deadline is 7 days from the date of the loss, which can be extended up to 30 days for offsetting
      rural activity tax losses.
### Evaluation:
{{
  "reasoning": "The AI response is incorrect because it mentions '7 days' and 'up to 30 days,' which
        contradicts the expert's response.",
  "result": "WRONG"
}}
4.
### Question:
Can the negative CSLL calculation base be offset against results determined in subsequent periods?
### Expert Response:
Yes. The CSLL calculation base, when negative, can be offset up to 30% of the results determined in
      subsequent periods, adjusted by the additions and exclusions provided for by law.
### AI Response:
No. The CSLL calculation base can be offset against results determined in subsequent periods.
### Evaluation:
{{
  "reasoning": "The AI response contradicts the expert's response, providing an opposite answer
        regarding the possibility of CSLL offset.",
  "result": "WRONG"
}}
Now think step by step and make this Evaluation:
### Question:
{questao}
### Expert Response:
{resposta_especialista}
### AI Response:
{resposta_ia}
### Evaluation:
```