Towards a Corpus Methodology for LLMs in the Legal Domain

Aline Athaydes, Lucas B. Bulcão Mota, Fernando Humberto de Almeida Moraes Neto, Samuel Rios da Silva, Babacar Mane, Daniela Barreiro Claro, Marlo Souza, Andressa Beatriz Cardoso Lisboa

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

{alineathaydes,lucasbulcao,fernando.humberto,babacarm,}@ufba.br

{samuelrs,dclaro,msouza1}@ufba.br,andressalisboa28@gmail.com

Abstract. *The creation of high-quality Question-Answer (QA) datasets is critical for developing reliable legal AI systems, yet a significant gap exists between intrinsic textual metrics and real-world model performance. This paper introduces an end-to-end framework to bridge this gap. We first present a refined methodology for generating a legal QA dataset (V2) based on the Brazilian Consumer Protection Code (Código de Defesa do Consumidor - CDC), demonstrating its superiority over a baseline corpus using metrics such as MTLD and Shannon Entropy. We then assess its practical impact by fine-tuning a Qwen3-8B model with LoRA. The model's performance is evaluated on a novel, expert validated 76 question multiple choice benchmark. Results show that the fine-tuned model achieves perfect accuracy on the benchmark and surpasses the base model across text generation metrics including BLEU, METEOR and BERTScore. Our work offers a reproducible methodology for legal dataset construction and validation, providing empirical evidence that improvements in data quality yield tangible gains in downstream legal reasoning tasks.*

1. Introduction

With the growing adoption of Artificial Intelligence (AI) in legal services, including automated chatbots and document analysis tools, ensuring reliability and legal precision has become a critical objective. In Brazil, the Consumer Protection Code (Código de Defesa do Consumidor – CDC) [Brasil 1990] is a central reference for consumer rights. However, its complex structure, interpretative difficulty, and legal specialized dialect create obstacles for its direct application in computational systems.

High-quality and domain specific datasets are essential for training Natural Language Processing (NLP) systems. However, in the legal domain, especially in Portuguese, such resources remain scarce. The generation of synthetic question-answer (QA) datasets for legal applications poses several challenges: simplistic prompting strategies often lead to corpora with low lexical diversity, frequent repetition, unclear or imprecise information regarding normative principles and jurisprudence, and fundamental linguistic flaws (e.g., missing diacritics), undermining both model training and legal validity.

To address these issues, recent work has proposed improvements for prompt engineering techniques that control parameters such as temperature, frequency penalty, and tone to produce more coherent and diverse data. However, most efforts remain centered on dataset creation and intrinsic textual evaluations, such as entropy, lexical diversity (e.g., MTLD and TTR), readability (e.g., Flesch Reading Ease for Portuguese), and keyword density [Lucena et al. 2025]. The

critical question remains underexplored: *Do high-quality QA datasets lead to better-performing Large Language Models (LLMs)in downstream legal tasks?*

Many datasets lack external validation or structured benchmarks. Few include manually reviewed gold standard sets or expert validated multiple choice evaluations. This gap limits the assessment of LLMs in real legal reasoning tasks and undermines model comparability under standardized conditions.

This study directly addresses these gaps by building upon the foundational dataset from our previous work (Athaydes et al., 2024) [Athaydes et al. 2024]. We introduce an advanced prompting methodology to generate a refined QA corpus (*V2*) with superior quality. To demonstrate the impact of this improved dataset, we fine-tune the base Qwen3-8B language model [Yang et al. 2025] using the LoRA technique [Hu et al. 2021] on this dataset and employ it on a new gold-standard benchmark of 76 multiple choice questions validated by a legal expert. Performance is measured not only by accuracy (achieving 100% versus 98.7% for the base model) but also through established generation metrics such as BLEU [Papineni et al. 2002], METEOR [Banerjee and Lavie 2005], and BERTScore [Zhang et al. 2020]. Our findings open a research opportunity to achieve high-quality datasets and models for the legal domain.

This paper is organized in sections as follows: Section 2 describes the related work, section 3 presents our approach. Section 4 describes our results and discusses them in Section 5, and finally Section 6 concludes our work. The acknowledgments are presented at the end of the paper.

2. Related Work

The advancement of NLP in specialized fields like law depends on high-quality datasets and robust evaluation benchmarks. This section reviews prior work in these two critical areas within the Brazilian legal context.

2.1. Datasets for Brazilian Legal NLP

The scarcity of structured, public datasets remains a major challenge in Brazilian legal NLP. Some initiatives have attempted to address this gap. The VICTOR dataset [Luz de Araujo et al. 2020] provides a large-scale corpus for classifying Supreme Court documents, while CDJUR-BR [Maurício et al. 2023] offers rich annotations for Named Entity Recognition (NER), similar to LeNER-BR[Luz de Araujo et al. 2018]. Despite their value, these resources were not designed for QA tasks, which require a direct mapping between questions and factual answers.

A recent attempt to address the QA gap was made by Athaydes et al. (2024) [Athaydes et al. 2024], who generated a synthetic dataset based on the Brazilian CDC. Although it represents an important step, the resulting dataset exhibits typical issues found in synthetic legal corpora, including low lexical diversity and notable linguistic flaws. These shortcomings emphasize the ongoing need for more refined data generation pipelines.

2.2. Evaluation of Large Language Models

Evaluating LLMs in the legal domain requires specialized benchmarks. While general benchmarks like MMLU [Hendrycks et al. 2021] assess broad knowledge, legal specific ones such as LegalBench [Guha et al. 2023] have been developed for English. For Brazilian Portuguese, however, there is a clear lack of such standardized benchmarks, making it difficult to compare models on local legal tasks.

To adapt models for specialized domains, parameter efficient fine-tuning (PEFT) methods are now standard. Techniques like Low-Rank Adaptation (LoRA) [Hu et al. 2021] are

commonly used to specialize models like the Qwen family [Yang et al. 2025] efficiently. For measuring the quality of generated text, metrics such as BLEU [Papineni et al. 2002], METEOR [Banerjee and Lavie 2005], and the semantically aware BERTScore [Zhang et al. 2020] are well-established in the field. Recent efforts have combined these methods to propose new resources tailored to the Brazilian legal domain, leveraging PEFT strategies and established metrics for evaluation.

3. Methodology

Our methodology is organized into two main phases. The first phase details the creation and intrinsic validation of a high-quality legal QA dataset. The second phase describes the experimental setup for evaluating the dataset's practical impact on a fine-tuned language model's performance.

3.1. Dataset Generation

Addressing the limitations identified in our prior work [Athaydes et al. 2024], our goal is to present a refined methodology for generating high-quality QA datasets in the legal domain. We take the corpus proposed in Athaydes et al. (2024) as our baseline, hereafter referred to as V1. That work provides a comprehensive description of the source documents used, which include the Brazilian CDC [Brasil 1990], legal summaries (*Súmulas*), and court decisions (*Acórdãos*). This paper introduces *V2*, a new version of the dataset generated from the same sources but employing an improved pipeline for prompt engineering and data preprocessing. The process for constructing this enhanced resource is described by the following structured pipeline.

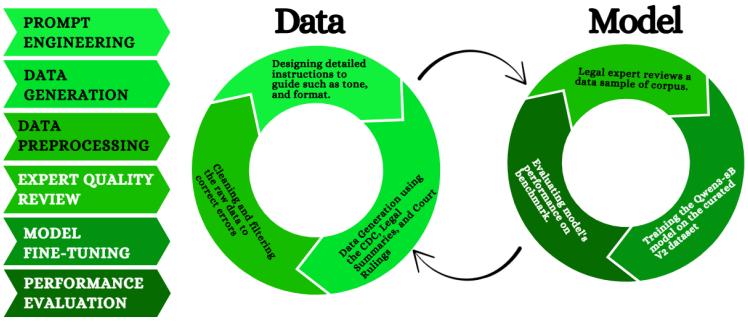

Figure 1. Overview of the dataset creation and evaluation pipeline.

3.1.1. Data Sourcing

We first developed a custom web crawler to gather a corpus of official documents to ground the dataset in a legal context. This corpus is composed of three primary sources from the Brazilian legal system:

- The full text of the CDC [Brasil 1990];
- Official legal summaries (súmulas) from the Superior Court of Justice (Superior Tribunal de Justiça - STJ);
- An extensive collection of full-text court decisions pertaining to consumer law.

These collected documents served as the foundational source material for the generation process.

3.1.2. Prompt Engineering

Using the GPT-4o Mini model [OpenAI 2024], we generated QA pairs from legal documents. Prompts were carefully crafted to ensure thematic variety, self-containment, a technical and challenging tone, and inclusion of legal references.

To ensure high-quality output, the model's hyperparameters were empirically determined:

- `temperature = 0.8` (reduced creativity; default = 1.0)
- `frequency_penalty = 0.5` (discourages repetition; default = 0.0)
- `presence_penalty = 0.8` (encourages topic diversity; default = 0.0)

These specific values were chosen to find a balance between factual accuracy and linguistic fluency. A $temperature$ of 0.8 allowed for varied and natural-sounding responses without significantly increasing the risk of hallucination. The $frequency\_penalty$ and $presence\_penalty$ were calibrated to discourage repetitive phrasing and encourage the model to cover a wider range of legal topics, directly addressing a key weakness of the baseline dataset.

Recognizing that a legal dataset requires ongoing maintenance to remain accurate, we propose a concrete protocol to ensure its long-term reliability:

- **Expert Audits:** Periodic sampling of the dataset would be reviewed by legal experts to capture changes in legislation or interpretation.
- **User Feedback:** A basic reporting mechanism (e.g., "report this answer") would allow users to flag issues and support ongoing refinement.

The application of these prompt engineering techniques, combined with an emphasis on continuous maintenance, led to the construction of a more coherent and diverse dataset, with fewer outliers and more refined question-answer pairs.

3.1.3. Generated Dataset and Pre-processing

The prompt engineering process yielded an initial raw dataset generated by the GPT-4o Mini model [OpenAI 2024], structured as a collection of JSON objects. Each object included three keys: question, answer, and context, with the context field containing the specific legal excerpt used during generation.

This raw output, however, exhibited predictable artifacts and required a structured pre-processing pipeline to ensure usability. We combined manual review with automated scripts to address three primary categories of issues:

- **Formatting Issues:** Fixed merged fields, blank entries, and artifacts (e.g., colons, enumeration remnants).
- **Content Flaws:** Removed non-Portuguese text, corrected grammar/tokenization, and ensured correctness of legal source references.
- **Filtering and Duplicates:** Excluded entries with less 10 tokens and removed 3,100+ duplicate questions.

This process resulted in the final validated dataset *V2*, composed of 61,870 QA pairs. This version served as the basis for all subsequent evaluations and experiments.

3.2. Evaluation by Specialist

To establish a quality baseline for the *V2* dataset, we developed a structured evaluation protocol executed by a legal expert specializing in Brazilian consumer law. This qualitative assessment was designed to measure the dataset's quality across multiple dimensions of linguistic correctness and juridical soundness.

A random sample of 100 QA pairs was extracted from the *V2* corpus for manual analysis. For each pair, the expert rated it against a custom developed rubric composed of eight distinct dimensions. To ensure evaluative rigor and reduce subjectivity, each dimension was framed as a direct question for the expert:

1. **Technical Clarity:** Is the answer technically accurate and well formulated?
2. **Clarity for Non-Experts:** Is the language accessible to someone without legal training?
3. **Juridical Accuracy:** Is the legal reasoning correct and aligned with the CDC?
4. **QA Consistency:** Does the answer respond logically and directly to the question?
5. **Plausibility:** Is the question a realistic example of a consumer concern?
6. **Contextual Alignment:** Is the cited legal article appropriate for the QA pair?
7. **Expert Observations:** Notes on potential improvements or ambiguities.
8. **Need for Correction:** Does the pair require revision to meet quality standards?

For the first six dimensions, the expert assigned a score using a 5 point Likert scale (where 1 = Very Weak and 5 = Excellent). This process provided valuable, structured insights into the dataset's strengths and weaknesses. We acknowledge that this analysis was conducted by a single domain expert. However, the primary goal was to validate the overall quality of the corpus. The results of this validation were also used to select a high-confidence subset of 76 QA pairs, which formed the foundation for the gold-standard benchmark used in our downstream experiments.

3.3. Experimental Setup and Fine-Tuning

We fine-tuned the Qwen3-8B model [Yang et al. 2025], selected for its strong performance in multilingual and long context tasks, using the validated *V2* dataset. From 58,112 QA pairs, a portion was held out for validation. Fine-tuning was performed with LoRA [Hu et al. 2021] (rank=16, `lora_alpha`=32, learning rate=5e-5), allowing efficient adaptation to the legal domain with reduced computational cost.

3.3.1. Benchmark Construction

To evaluate the model's legal reasoning capabilities, we constructed a multiple choice benchmark grounded in expert validation. From the previously reviewed QA pairs, 76 were approved for use. For each, we used GPT-4 to generate plausible distractors (incorrect alternatives), which were then manually reviewed and refined by the legal expert. This resulted in a curated benchmark of 76 high-quality multiple choice questions serving as the evaluation standard.

3.3.2. Evaluation Metrics

We employed two complementary evaluation strategies:

- **Accuracy:** On the multiple choice benchmark, accuracy was calculated as the proportion of correct responses selected by the model against the expert defined answers.
- **Text Generation Quality:** On a held-out subset of *V2*, we evaluated the semantic and linguistic alignment between generated and reference answers using BLEU, METEOR, and BERTScore.

These metrics are defined as:

- **BLEU**: measures n-gram overlap between the generated output and the reference, favoring lexical and sequential matches.
- **METEOR**: combines exact, stemmed, and synonymous matches, weighted by precision, recall, and fragmentation, benefiting from a vocabulary consistent with the reference.
- **BERTScore**: computes semantic similarity using contextual embeddings, rewarding models that reproduce semantic and lexical constructions closely aligned with the reference set.

4. Results

Our experimental validation was conducted in two sequential phases. First, we performed a detailed intrinsic analysis to rigorously compare the linguistic and structural qualities of our dataset *V2* against the baseline corpus. Second, we performed a downstream task evaluation to measure the real world impact of our *V2* dataset on a fine-tuned language model.

4.1. Intrinsic Quality Analysis: V1 vs. *V2*(our proposal)

To demonstrate that our data generation methodology produced a superior corpus, we compared V1 Dataset and *V2* across a comprehensive set of quantitative and qualitative metrics.

4.1.1. Token Statistics

An analysis of structural properties (Table 1) shows that the *V2* entries are longer (125.67 vs. 109.29 avg. tokens) and more consistent, with a higher minimum token count (61 vs. 14). This reflects a deliberate choice to prioritize response completeness over V1 shorter but more numerous entries. The increased standard deviation in *V2* (23.72 vs. 20.35) also suggests broader thematic coverage.

Table 1. Token statistics for Datasets V1 and *V2*.

Token Statistic	**V1**	***V2***
Average Number of Tokens	109.29	125.67
Maximum Tokens	246	346
Minimum Tokens	14	61
Standard Deviation	20.35	23.72
Average Tokens per Question	19.25	22.62
Average Tokens per Answer	50.69	59.67
Average Tokens per Context	39.35	43.46

4.1.2. Entropy per Word

To measure information unpredictability, we used Shannon Entropy (H). The results in Table 2 shows our *V2* dataset has a higher average entropy (5.899 vs. 5.389). This indicates higher lexical diversity and less predictable content, aligning with the use of penalties in our prompt engineering to reduce repetition.

$$H = -\sum_{i=1}^{n} p(w_i) \log_2 p(w_i)$$

$$H = -\sum_{i=1}^{n} p(w_i) \log_2 p(w_i)$$

where H is the entropy, the sum is over the n distinct words in the text, and $p(w_i)$ is the probability of occurrence of the i-th word, w_i.

Table 2. Word-level entropy statistics for V1 and *V2*.

Metric	V1	*V2*
Average Entropy	5.389	5.899
Maximum Entropy	6.344	7.522
Minimum Entropy	1.500	4.140
Standard Deviation	0.236	0.272

4.1.3. Measure of Textual Lexical Diversity (MTLD)

As a robust measure of vocabulary variety, we calculated the MTLD score. As detailed in Table 3, the score for *V2* (140.32) is more than double that of V1 (64.26). This confirms a clear increase in vocabulary richness, demonstrating the effectiveness of the refined prompting.

$$\text{TTR} = \frac{V}{N}$$

where V is the Number of unique words (types) and N: Total number of words (tokens)

Table 3. MTLD statistics for datasets V1 and *V2*.

Metric	V1	*V2*
Mean MTLD	64.26	140.32
Maximum MTLD	260.47	1372.00
Minimum MTLD	3.00	32.42
Standard Deviation	18.17	67.73

Dataset *V2* exhibits significantly higher MTLD scores than V1 (mean of 140.32 vs. 64.26), indicating a substantial improvement in lexical diversity. The minimum MTLD in *V2* (32.42) is notably greater than in V1 (3.00), showing that even its least diverse texts maintain acceptable complexity. The maximum MTLD of 1372.00 in *V2*, although likely an outlier, reflects the impact of improved prompting strategies. Furthermore, *V2*'s standard deviation is nearly four times larger, confirming a wider range of lexical variation that enriches the dataset.

4.1.4. Flesch Reading Ease (FRE-PT)

Readability analysis using the FRE-PT index shows *V2* is more linguistically complex, with an average score of 31.34 versus 43.67 for V1. This outcome is consistent with the generation of longer, more detailed, and technically precise legal text in the *V2* dataset.

$$\text{FRE-PT} = 248.835 - 1.015 \times \left(\frac{\text{words}}{\text{sentences}} \right) - 84.6 \times \left(\frac{\text{syllables}}{\text{words}} \right)$$

The Flesch Reading Ease formula was adapted for Brazilian Portuguese by adjusting its constant to 248.835 (from the English original of 206.835) to better reflect the language's structural features. The resulting scores classify text on a scale where higher values indicate easier readability (e.g., 90-100 for "very easy") and lower values denote more difficult text (e.g., 0-30 for "very difficult").

Table 4. FRE-PT statistics for datasets V1 and *V2*

statistics	V1	*V2*
Average FRE-PT	43.67	31.34
Standard Deviation	12.11	14.64
Maximum FRE-PT	100.38	88.38
Minimum FRE-PT	-19.82	-46.97

Dataset *V2* also shows greater variability in readability, with a higher standard deviation (14.64 vs. 12.11), likely due to the increased thematic diversity introduced through prompt engineering.

4.1.5. Density of Relevant Terms

An analysis of term density using TF-IDF showed a comparable usage of core legal terms in both datasets, indicating *V2* maintained its domain focus.

$$\text{Density} = \frac{\sum_{i=1}^{n} f(w_i)}{\text{Total number of words in the corpus}}$$

where the numerator, $\sum f(w_i)$, represents the total count of all predefined relevant words, and the denominator is the total word count of the corpus.

To identify the relevant words in both corpora, we applied TF-IDF (Term Frequency-Inverse Document Frequency) [Salton and Buckley 1988] finding the 100 words more relevants in each dataset. Following this, we calculated the relevant word density for each row in the dataframes.

Table 5. Average DPR for Datasets V1 and *V2*

Dataset	Average DPR
V1	0.2825
V2	0.2582

4.1.6. Vocabulary Divergence

To quantify vocabulary differences, we used the Jaccard similarity coefficient, which yielded a score of 0.5940. This low overlap confirms that *V2* features a distinct lexicon compared to the V1 baseline.

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$ are the lements shared by V1 and *V2* and $|A \cup B|$ are all the unique elements across both datasets.

By applying the formula to Datasets V1 and *V2*, we obtain:

Vocabulary Divergence(Jaccard) = 0.5940

This indicates a significant difference between the vocabularies of dataset V1 and dataset *V2*, most likely due to the modification of the penalty frequency hyperparameter, which increased the vocabulary variability in dataset *V2*.

4.1.7. Qualitative Analysis Summary

Finally, a qualitative analysis of 100 *V2* samples was conducted by a legal expert. The findings were positive, with high ratings for juridical accuracy (82%), question-answer coherence (85%), and clarity for laypeople (69%). The expert also noted that while 27% of items could be improved, the overall assessment confirmed *V2*'s quality and practical relevance. Based on this evaluation, 76 QA pairs were selected to compose the gold-standard benchmark used in our experiments.

4.2. Downstream Performance on Legal Reasoning Benchmark

Although intrinsic evaluations suggest that *V2* is of higher quality, we further assessed its practical impact by measuring how it enhances model performance. To this end, we fine-tuned a Qwen3-8B model on *V2* and compared its results with the base model using a legal reasoning benchmark. This downstream evaluation served as an external validation of the dataset's effectiveness in improving legal understanding and accuracy.

4.2.1. Benchmark Results

The primary evaluation focused on a multiple choice benchmark of 76 questions validated by a legal expert. As shown in Table 6, the Qwen3-8B model fine-tuned on the *V2* dataset reached 100% accuracy, correctly answering all items. This result indicates that fine-tuning improved the model's ability to select the correct legal interpretations, even when faced with distractors that mimic real world legal ambiguity.

We also evaluated the quality of generated answers on a held-out test set. The fine-tuned model surpassed the base version in all metric, BLEU, METEOR, and BERTScore, demonstrating more fluent, coherent, and semantically accurate outputs. These improvements are essential for legal applications that require clarity and precision in automated responses.

Table 6. Evaluation of Qwen3-8B on Multiple Choice and Text Generation Tasks.

Model	Accuracy	BLEU	METEOR	BERTScore (F1)
Qwen3-8B (Base)	98.7% (75/76)	0.0128	0.1968	0.6606
Qwen3-8B (Fine-Tuned)	**100% (76/76)**	**0.0956**	**0.2980**	**0.7534**

5. Discussion

Our results validate the core hypothesis: higher intrinsic quality in legal QA datasets improves downstream performance. *V2* surpassed the V1 baseline in structure, lexical diversity, and information richness, and these enhancements yielded gains in legal reasoning accuracy.

The modest increase in multiple choice accuracy (98.7% to 100%) contrasts with the more substantial improvement in generation metrics (BLEU, BERTScore). This indicates that while

strong base models like Qwen3-8B already perform well on classification tasks, fine-tuning on curated datasets sharpens their linguistic and semantic precision, critical in user facing legal applications.

Regarding the generation metrics, while the fine-tuned model clearly outperforms the baseline, the low absolute value of BLEU (0.0956) warrants consideration. This is characteristic of tasks where semantic adequacy is more important than exact lexical overlap. As BLEU is a precision oriented metric based on n-gram matching, it tends to undervalue correct answers that are paraphrased. In contrast, the much higher BERTScore (0.7534) provides stronger evidence that our model produces legally coherent and contextually aligned answers.

Finally, we acknowledge the limitations of this study. The evaluation setup, using a consistent synthetic source for training and testing, ensures a controlled comparison but may amplify the model's alignment to a specific linguistic style. Furthermore, the use of a single model architecture and a relatively small, domain specific benchmark (76 questions) means that further research is needed to assess the generalizability of these findings.

6. Conclusion

This work presented an end-to-end methodology for generating, validating, and testing legal QA datasets. Our refined dataset (*V2*), guided by both linguistic metrics and expert review, demonstrated superior intrinsic quality and produced tangible improvements in model performance.

Fine-tuning a Qwen3-8B model on *V2* led to perfect benchmark accuracy and improved text generation. These results highlight the importance of data-centric strategies in legal NLP and suggest that intrinsic improvements can lead to real-world gains.

As next steps, we plan to conduct a targeted ablation study to isolate the effects of different dataset characteristics and model configurations on performance. We also intend to expand our benchmark to cover more legal topics and to incorporate multiple legal experts in the validation process, addressing the single-evaluator limitation and ensuring greater robustness.

Acknowledgment

This work was partially supported by FAPESB through grants TIC 0002/2015, CCE 0022/2023, and INCITE PIE0002/2022. The authors also express their gratitude to Escavador for its support and to all members of FORMAS for their valuable contributions.

References

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (qa) models. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 493–500, Porto Alegre, RS, Brasil. SBC.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y., and Voss, C., editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Brasil (1990). *Código de Defesa do Consumidor*. Presidência da República - Casa Civil. Lei nº 8.078, de 11 de setembro de 1990.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Lucena, D., Souza, E. P., Albuquerque, H., Da Silva, N., Oliveira, A., and de Carvalho, A. (2025). Performance analysis of llms for abstractive summarization of brazilian legislative documents. *Conference on Digital Government Research*, 1.

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020). VICTOR: a dataset for Brazilian legal documents classification. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, page 313–323, Berlin, Heidelberg. Springer-Verlag.

Maurício, A., Pinheiro, V., Furtado, V., Neto, J. A. M., Bomfim, F. C. J., da Costa, A. C. F., Silveira, R., and Aragão, N. (2023). Cdjur-br: A golden collection of legal documents from brazilian justice with fine-grained named entities. *arXiv preprint arXiv:2305.18315*.

OpenAI (2024). Gpt-4o: Openai's multimodal model with improved efficiency and reasoning. `https://openai.com/research/gpt-4o`.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.