LegalSim-PT: Building a Dataset for Legal Document Simplification in Portuguese Leveraging Linguistic Metrics

Arthur Scalercio[1], **Maria José Finatto**[2], **and Aline Paes**[1]
[1]Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil,
[2]Institute of Linguistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
arthurscalercio@id.uff.br, mariafinatto@gmail.com, alinepaes@ic.uff.br

Abstract

Document simplification has recently attracted increasing attention due to its broader practical applicability compared to sentence-level simplification. Beyond simplifying individual sentences, this task involves preserving fluency, conciseness, and coherence across the entire text, often incorporating summarization techniques. Despite its importance, research in this area remains largely concentrated on a few languages, particularly English. In this work, we introduce LegalSim-PT, the first large-scale Portuguese dataset for document simplification based on legal texts. To mitigate reliance on manual evaluation, we combined data augmentation strategies with readability, semantic similarity, and diversity metrics to select the most suitable document pairs. We conducted a comprehensive analysis of the resulting dataset, first characterizing its surface features and comparing them with those of existing simplification corpora. Next, we assessed its quality using automatic metrics, linguistic indicators, and human evaluations. Finally, we select representative models as baselines and fine-tune two models on LegalSim-PT, achieving improved performance in document-level simplification.

1 Introduction

Text simplification (TS) aims to make written information more accessible by reducing linguistic complexity while preserving meaning (Al-Thanyyan and Azmi, 2021). It aims to improve accessibility for a broad range of readers, including language learners, individuals with cognitive or literacy difficulties, and citizens who must interpret complex institutional documents (Martínez et al., 2024). Traditionally, TS research has focused on sentence-level simplification (Al-Thanyyan and Azmi, 2021), where the task is to rewrite each complex sentence in isolation. Although valuable, this approach does not capture broader discourse-level phenomena such as coherence, conciseness, and referential consistency, which are essential for understanding long documents (Vásquez-Rodríguez et al., 2024). Consequently, document-level simplification (DS) has emerged as a more practically relevant yet methodologically challenging task, requiring integration of summarization, compression, and reorganization techniques to maintain both global meaning and local fluency (Cripwell et al., 2023b).

The legal domain accurately illustrates these challenges (Garimella et al., 2022). Legal documents are characterized by complex syntactic structures, archaic or specialized vocabulary, and intricate discourse relations (Zhong et al., 2020). Such linguistic density poses barriers to transparency, citizen participation, and equitable access to justice (Vargas, 2023; Cabrera, 2024). Simplifying legal documents can enhance public understanding of laws, contracts, and institutional policies, thereby supporting efforts toward linguistic accessibility and open government. However, creating corpora that simplify legal language while preserving precision and normative validity requires careful control over both meaning and form, a challenge compounded by the limited availability of training data (Ariai et al., 2024).

Despite growing attention to document simplification, most existing research remains restricted to English and relies on a limited set of corpora that primarily support sentence-level simplification (Ryan et al., 2023a). In other languages, particularly Portuguese, the scarcity of simplification datasets represents a major obstacle to progress. Pioneering resources like PorSimples (Aluísio et al., 2008) have focused on sentence-level simplification in Brazilian Portuguese, but there are currently no *large-scale* datasets addressing document-level simplification in specialized domains, such as the legal domain.

In this work, we introduce LegalSim-PT, the first large-scale Portuguese dataset for document simplification based on legal texts. The dataset was con-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 392–405

structed through a data augmentation and filtering pipeline, designed to automatically generate and select simplification pairs that meet quality standards for readability, semantic preservation, and diversity (Cripwell et al., 2023a). The resulting corpus was then subjected to extensive analysis: (i) surface and linguistic features were compared with existing simplification datasets, (ii) dataset quality was assessed via automatic metrics, human evaluation, and qualitative assessment with LLMs, and (iii) the dataset's effectiveness was tested through model fine-tuning experiments.

We selected four representative models as baselines for the legal document simplification tasks, and we fine-tuned two of them on our dataset. Results demonstrate that models trained with LegalSim-PT achieve substantial improvements in document-level simplification performance, producing outputs that are simpler, more fluent and readable without compromising legal meaning.

This paper makes three main contributions:

1. We introduce LegalSim-PT, the first large-scale Portuguese dataset dedicated to document-level simplification in the legal domain. The resource is publicly available for research and educational purposes, adhering to the principles of openness and reproducibility [1].

2. We conduct a detailed quantitative and qualitative analysis of the dataset, comparing its characteristics with those of existing corpora and with texts generated by other LLMs.

3. We evaluated both representative and fine-tuned models on our dataset using automatic metrics. The results provide strong baselines, indicating that domain-specific fine-tuning on our dataset leads to consistent gains in simplicity, readability, and semantic preservation.

2 Related Work

In this section, we review works most closely related to our proposal, with particular emphasis on initiatives that expand simplification research to underrepresented languages, especially Portuguese.

2.1 Document Simplification and Evaluation

Recent work has highlighted the importance of document-level simplification, where models handle entire paragraphs or documents to maintain logical flow and referential clarity across sentences (Cripwell et al., 2023a). For example, some approaches incorporate explicit discourse planning or sentence splitting strategies to improve coherence (Cripwell et al., 2023b). Cripwell et al. (2023a) split sentences based on discourse structure to simplify text without losing connectivity, and Vásquez-Rodríguez et al. (2023) introduced a coherence-aware TS evaluation, showing that prior simplification systems rarely evaluated how well the output stays coherent across multiple sentences. Besides coherence, meaning preservation is also a key challenge in TS. Recent research has explored multi-objective training and constraints to ensure simplified outputs remain faithful, (Cripwell et al., 2024), for example, by jointly optimizing simplicity, readability, and coherence (Vásquez-Rodríguez et al., 2024).

Recent studies have also evaluated automatic simplification in specialized domains such as healthcare (Wives and Finatto) and legal (Alves et al., 2023; Pereira et al., 2024; Garimella et al., 2022), examining how mainstream methods perform and exploring metric choices and limitations in those settings. Garimella et al. (2022) evaluates a range of unsupervised and supervised models on legal data, combining automatic metrics (for readability, meaning preservation, fluency, and hallucination) with expert human evaluation. Alves et al. (2023) compared a transformer and a neural machine translation method against MUSS baselines Martin et al. (2022) on a legal Brazilian dataset of about 200 documents. They found all approaches improved the Flesch Reading Ease (FRE) scores of complex legal documents, with the best models substantially increasing readability while keeping legal content intact. Pereira et al. (2024) examined how existing resources perform when applied to legal documents in Portuguese, given the lack of specific annotated datasets. The study combined qualitative and quantitative analyses using five different models. These demonstrate the applicability of neural DS in the legal domain, though the authors note that obtaining parallel legal simplification corpora remains challenging.

2.2 Document Simplification Datasets

Dataset creation is key to empirical TS research. Early efforts centered on English, e.g., Newsela (Xu et al., 2015) and WikiLarge (Zhang and Lapata, 2017), leaving other languages un-

[1] https://github.com/scalercio/legal-doc-simplification-data

derrepresented. Recent initiatives expanded and harmonized multilingual resources, like the Text Simplification Repository [2], which catalogs over 70 datasets across languages and domains, distinguishing parallel from comparable corpora.

A notable advance in harmonization resources is MultiSim (Ryan et al., 2023b), which integrates 27 simplification datasets in 12 languages (over 1.7 million pairs) into a unified benchmark, standardizing metadata and formats for multilingual modeling. Similarly, MultiCochrane (Joseph et al., 2023) and MultiMSD (Horiguchi et al., 2025) extend simplification into the medical domain, aligning manually simplified sentences in English, Spanish, and French, for example. The EASIER (Alarcón et al., 2021) and FEINA (Perez-Rojas et al., 2023) corpora contribute Spanish datasets focusing on accessibility and financial literacy, respectively. Additionally, the long-running Simplext program developed resources and methodologies based on Plain Reading principles to make information cognitively accessible, including alignment and generation components, and later system evaluation (Saggion et al., 2011). There are also initiatives for other languages like German and Japanese(Anschütz et al., 2023; Nagai et al., 2024).

Regarding Portuguese, early efforts at simplification, such as the PorSimples project (Aluísio et al., 2008; Aluísio and Gasperin, 2010), laid the groundwork by compiling simplification guidelines and corpora, targeting digital inclusion. Scalercio et al. (2025) gathered publicly available pairs of texts and their simplified versions from various Brazilian government agency websites. Although most texts consist of only a few sentences, given their domain similarity to legal language, this dataset was used for comparison.

2.3 Synthetic Dataset Development

LLMs are now widely used to produce complex–simple text pairs automatically (Yang, 2024; Vásquez-Rodríguez et al., 2024), enabling the creation of large-scale datasets at relatively low cost. However, ensuring semantic fidelity and factual correctness remains a major challenge, as synthetic simplifications may introduce omissions, distort legal or technical details, or generate overly generic rewrites (Devaraj et al., 2022).

To mitigate these issues, recent approaches incorporate fidelity-aware generation pipelines, which combine semantic similarity filtering (e.g., cosine or BERTScore thresholds), round-trip translation validation, and consistency-based filtering (Zhuo et al., 2023; Manakul et al., 2023). Other works employ human-in-the-loop protocols to refine and audit synthetic data (Kang et al., 2024), while readability- and perplexity-based reranking strategies are applied to balance linguistic diversity and semantic preservation (Chim et al., 2025). Together, these efforts highlight an emerging trend toward the controlled generation of synthetic data, where correctness and transparency are integral to dataset construction and evaluation.

3 LegalSim-PT: a Portuguese Legal Document Simplification Dataset

This section describes the creation process of LegalSim-PT and the statistics of the final dataset.

3.1 Dataset Construction

We followed standard document-level simplification procedures to construct LegalSim-PT, a new large-scale dataset for legal text simplification in Portuguese. Our dataset builds upon LegalPT (Garcia et al., 2024a), which comprises approximately twelve million Portuguese legal documents organized into eleven sub-datasets. From these, we selected the six sub-datasets pertaining exclusively to the Brazilian legal domain. Using this material, we generated simplified versions for a representative subset of texts with the Qwen3-4B model (Yang et al., 2025). This model was chosen for its compact size and strong performance, comparable to that of Qwen2.5-72B-Instruct across multiple benchmarks (Yang et al., 2025)[3].

To manage computational demands, we established a context window of 8,192 tokens and limited the maximum output at 4,096 tokens. In practice, this threshold proved highly effective: only a negligible fraction of documents (approximately 2%) exceeded it, ensuring that the vast majority of the corpus was processed without truncation.

From the original set of documents, we applied three filters to ensure the quality of the final dataset: (i) semantic consistency, (ii) readability, and (iii) diversity. To verify semantic consistency between each pair, we employed Sentence-BERT (Reimers and Gurevych, 2019) to compute sentence embeddings. Since the documents consist of multiple sentences, each one was split into smaller chunks,

[2] https://github.com/jantrienes/text-simplification-datasets

[3] https://qwen.ai/research

for which individual vector representations were generated. These vectors were then averaged to produce a single embedding per document. Based on empirical analysis, we retained only those document pairs with a cosine similarity greater than 0.8, ensuring that the simplified text remained semantically aligned with its source.

Since no annotated labels were available to evaluate the simplicity of the generated texts, and gathering them from manual human evaluation is costly and challenging, we adopted the Flesch-Kincaid Readability Index (Kincaid et al., 1975), which can be computed without reference data. This metric is grounded in the assumption that shorter words and sentences contribute to easier reading. We employed the Portuguese-adapted version of the formula (Leal et al., 2023). Although readability formulas such as Flesch and Flesch–Kincaid capture only surface-level features (e.g., sentence length and syllable count) and are not reliable proxies for overall simplification quality (Tanprasert and Kauchak, 2021), we report them as secondary descriptive metrics. Specifically, they serve as transparent, reproducible indicators of structural and lexical compression.

In addition, we applied a diversity filter to ensure that the number of unique tokens in each simplified document was lower than in its original counterpart. Table 1 presents the number of document pairs initially generated and those retained after applying the aforementioned filters.

Source data	#Created	#Filtered	%
acordaos_tcu	424.863	150.733	35,5
datastf	308.333	99.244	32,2
iudicium_textum	149.747	31.718	21,2
BRCAD-5	296.191	200.257	67,6
CJPG	298.393	167.921	56,3
tesemo_v2	735.239	316.471	43,0
Total	**2.212.766**	**966.344**	**43,7**

Table 1: Pair counts before and after the filtering.

Gov-Lang-BR Test Set Given the absence of other document-level simplification datasets in Portuguese, we adopted an existing corpus for comparison and evaluation: Gov-Lang-BR (Scalercio et al., 2025). Although originally developed for sentence simplification, this dataset also includes multi-sentence examples and texts from the legal domain. It was compiled by collecting original and simplified versions of documents from various Brazilian government websites, making it a robust resource for our evaluation.

3.2 Dataset Analysis

This section describes the main characteristics of LegalSim-PT.

Surface Statistics The LegalSim-PT dataset was analyzed and compared with the Gov-Lang-BR corpus, and the results are shown in Table 2.

		LegalSimPT	GovLangBR
# Documents	Original	966,344	1,703
	Simplified	966,344	1,703
Ave. # of sentences	Original	30.24	1.22
	Simplified	15.32	1.10
Ave. # of words	Original	628.48	33.40
	Simplified	204.39	19.35
Ave. # of characters	Original	3286.16	181.42
	Simplified	1038.03	104.19

Table 2: Statistics of LegalSimPT and GovLangBR.

Compression Ratio We examined how simplification affects document length by calculating the compression ratio, which is the number of characters in the simplified text divided by the number in the original. Figure 1 shows how these compression ratios are distributed across LegalSim-PT, along with a kernel density estimation for smoothing the curve.

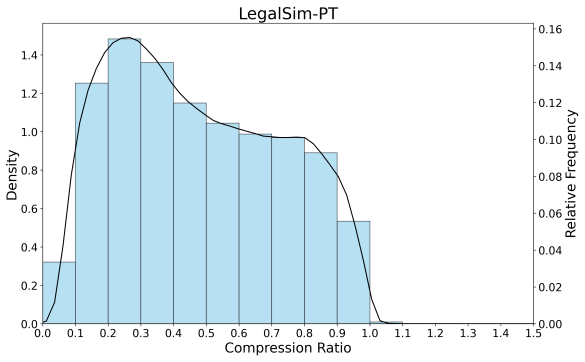


Figure 1: Compression rate distribution

As observed from the distribution, the compressions are relatively well distributed between 0.1 and 0.9, with a peak around 0.3. Although the simplified documents are, on average, considerably shorter than the original ones, all automatic and manual evaluations reported in Section 5 do not indicate any significant loss of semantic content. The compression rate distributions for each of the six subsets comprising LegalSim-PT are provided in the appendix A.

Document-Level Readability To evaluate the readability of LegalSim-PT, we computed the read-

ability scores using the Flesch readability index, the same metric used to filter out poor examples from our dataset. Table 3 lists the average readability scores of LegalSim-PT and GovLang-BR datasets.

	LegalSimPT	GovLangBR
Original (Source)	31.7	-12.1
Simplified (Target)	44.1	0.1

Table 3: Comparison of Flesch Readability Index

Although the documents from both datasets differ in terms of the average value of the metric, it can be observed that the gain in average readability for both datasets is quite similar. It is believed that the very low value of the readability metric for gov-lang-br is due to the complexity of this dataset, which consists of very long sentences and long and complex technical terms/jargon. Figure 2 illustrates the distribution of the difference in the readability metric between the simple and original documents.

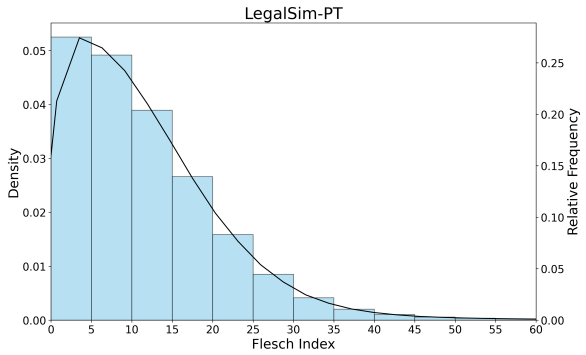


Figure 2: Flesch gain distribution

4 Experimental Setting

This section describes the experimental setting for evaluating the quality of our newly created dataset and establishing baseline results in the legal document simplification task.

4.1 Datasets

The **LegalSim-PT** dataset was randomly divided into training (946,350 document pairs), validation (9,997 document pairs), and test sets (9,997 document pairs). To further evaluate our construction strategy (Section 3.1), we also created two challenge sets, each comprising 200 example pairs.

The first, **challenge_good**, includes randomly selected examples from the last decile of both the semantic similarity and Flesch index increase distributions. Arguably, these examples demonstrate high semantic similarity between the original and simplified texts, accompanied by a substantial improvement in readability. Conversely, **challenge_hard** consists of examples from the first decile of the same distributions, representing pairs with low semantic similarity and minimal readability gains. The training and validation sets were used for training baseline models, whereas the test and challenge sets were reserved for quantitative and qualitative analyses.

4.2 Linguistic Metrics

To evaluate the quality of the simplifications in LegalSim-PT, we employed four reference-free linguistic metrics to assess examples from both the test and challenge sets. They are: (1) Average number of words to measure the size of the examples; (2) Ratio of passive to active voice verbs (P/A) to measure more direct constructions; (3) Proportion of adverbial clauses preceding the main clause (AdvL), capturing sentence structure tendencies; and (4) Ratio of fully developed to reduced relative clauses (D/R), reflecting syntactic simplifications. We consider the first to be fundamental due to the size of the documents in the dataset, while the other three were developed based on linguistic hypotheses about complexity that reflect structural complexity at the configurational level and are grounded in psycholinguistic research on language processing (Gibson, 1998; Charles, 2013; Corrêa et al., 2019). A higher P/A is associated with non-canonical word order and greater processing demands; simplification is therefore expected to reduce P/A by favoring active constructions. A higher proportion of adverbial clauses increases left-branching and delays the main clause, so simplification typically lowers AdvL. In contrast, an increase in the ratio of fully developed to reduced relative clauses (D/R) reflects greater explicitness and reduced syntactic compression, which is consistent with simplified, more transparent structures.

4.3 Human Evaluation

Furthermore, to thoroughly assess the quality of our dataset, we conducted a human evaluation focusing on three key aspects: adequacy, grammar, and simplicity. Simplifications are evaluated on a five-point Likert scale (1-5). Following Sun et al. (2021), in addition to lexical and structure simplification, we also adopted O-simplicity (Overall

simplicity with quality guarantee). O-simplicity indicates if the simplified document is simpler than the original, under the condition of quality guarantee, i.e., it also should read smoothly and retain the main meaning of the original document. The adequacy score measures whether the simplified version retains the main information, and the grammar score assesses whether the text is fluent and well-written. Three volunteer native Portuguese speakers, two with a background in linguistics and one law graduate, were asked to assess the document based on the above dimensions. We randomly selected 20 examples from each of our challenge sets and, for comparison purposes, our simple documents were compared with the simplifications generated by the LLMs Qwen2.5-7B-Instruct (Yang et al., 2024) and Bode (Garcia et al., 2024b). The first one is an instruction-tuned model that achieved the best results among 20 open-source LLMs in the Portuguese sentence simplification task (Scalercio et al., 2025), and Bode is a fine-tuned LLM for Portuguese prompt-based tasks, built for text generation tasks. Each simplification was rated once. In summary, the qualitative criteria are: (P1.) Simplicity with preservation of meaning and fluency; (P2.) Lexical simplification (replacement with simpler terms); (P3.) Structural simplification (reduction of syntactic complexity); (P4.) Meaning preservation (no essential omissions or irrelevant additions); (P5.) Grammatical correctness and fluency. More detailed instructions and questions can be found in Appendix B.

4.4 LLMs as evaluators

Human evaluation constitutes a robust and reliable strategy for assessing text quality; nevertheless, it is resource-intensive and requires specialized expertise in linguistics, law, or—ideally—both, a combination that is notably rare. To enable a qualitative assessment across a larger number of samples, we additionally employ LLMs as evaluators, following the LLMs-as-judges paradigm (Zheng et al., 2023). Specifically, we selected two LLMs that consistently appear among the top-ranked models in public leaderboards for this task [4]: GPT-OSS-120B and LLaMA3.3-70B, both accessed through the Groq API [5].

Both models were instructed to answer the same five evaluation questions designed for human assessment, with their definitions. The prompts included explicit instructions to return the output strictly in JSON format, facilitating automatic parsing of the responses. As in the human evaluation, each model received the three simplified versions of the original text for comparison. However, here the LLMs get all the 400 samples pertaining to the challenge sets described in Section 4.1. We run the models, changing only the temperature to be 0.0, to make the model more deterministic and less creative. We call them three times to account for other randomness in their behavior. Appendix C shows the system prompt used.

4.5 Baseline Models

We selected four representative models as the baselines for the legal document-level simplification task: (1) Qwen2.5-7B-Instruct (Yang et al., 2024), the same used during human evaluation; (2) Qwen3-1.7B (Yang et al., 2025): It is a reasoning model and the second smallest of the recent Qwen3 family models. (3) Bode (Garcia et al., 2024b), the same used during human evaluation; (4) Tucano-2b4-Instruct (Corrêa et al., 2025): It is a decoder-transformers natively pretrained in Portuguese, which achieved the best results in the sentence simplification among the portuguese LLMs (Assis et al., 2025). All the models were tested on our delineated test set. We also fine-tuned both the Qwen models for approximately two epochs, using Lora (Hu et al., 2021) and 4-bit quantization techniques. Experiments were conducted on an Ubuntu server equipped with two RTX 4090 GPUs (24 GB each). See Appendix D for inference and fine-tuning details.

5 Results

5.1 Linguistic Evaluation

We perform a large-scale linguistic analysis to try to understand what the LLM is doing or failing to do. We analyze the test set and both challenge sets. With this approach, we expect to measure the full spectrum of simplifications generated by the LLM. Initially, these sets were annotated morphosyntactically using the UDPipe model, which was trained on a scientific treebank (Straka et al., 2016). Then, we calculate the linguistic metrics delineated in Section 4.2.

When examining Table 4, we observed that the three sets follow the expected trend of developing

[4] https://www.prollm.ai/leaderboard/llm-as-judge, https://huggingface.co/spaces/AtlaAI/judge-arena

[5] https://groq.com/

Dataset	Linguistic Metrics			
	Avg Words	P/A	AdvL	D/R
Test Set				
Complex	620.0	.010	.30	.42
Simple	202.2	.017	.19	.53
Challenge Good				
Complex	576.7	.010	.29	.40
Simple	272.7	.013	.24	.50
Challenge Hard				
Complex	791.8	.009	.26	.37
Simple	150.7	.024	.14	.68

Table 4: Linguistic Metrics for three datasets. P/A is the ratio of passive to active voice; AdvL stands for the proportion of adverbial clauses preceding the main clause, and D/R is the Ratio of fully developed to reduced relative clauses.

reduced relative clauses (D/R) and moving adverbial clauses to after the main clause (AdvL). An unexpected trend observed across all three datasets, however, is the increase in the proportion of sentences in the passive voice. The test set metrics closely mirror those of the challenge-good set, whereas the challenge-hard set shows greater divergence, particularly in the simplified versions, largely due to stronger compression during simplification, which increases metric variability.

Thus, from these analyses, it can be inferred that the compression rate between original and simplified documents was a decisive factor in selecting examples for inclusion in the dataset. It is likely that examples with a compression rate higher than that of challenge-hard were discarded because they lacked the minimum semantic similarity required for inclusion, while the examples in challenge-hard lie precisely at the selection boundary, with semantic similarity values close to 0.8.

5.2 Human Evaluation

The human evaluation followed the procedure described in Section 4.3. To enrich our analysis and cover as much as possible of the range of examples that make up LegalSim-PT, we randomly selected samples from our challenge sets.

The results of the human evaluation are presented in Table 5. The evaluations confirmed that the hard set has lower semantic similarity compared to the good set; however, the achieved similarity was still quite high in both cases and significantly superior to that of the other models, which also struggled more to preserve semantics in the hard set. Regarding simplicity, there was a clear tendency among evaluators to favor simplifications that involved higher compression of the original document. Overall, the Qwen2.5 simplifications consistently maintained a high compression rate, with relatively stable scores. In contrast, our simplifications from the good set had lower compression, which seems to have affected primarily the syntactic and lexical simplicity metrics. For the hard set, however, all our simplicity indicators were much higher, reinforcing confidence in the quality of the dataset. The text generated by the Bode model, on the other hand, struggled with the length of legal documents and lost a considerable amount of content. Finally, all models produced highly fluent texts with very few grammatical errors.

To demonstrate that human judges assign higher simplicity scores to more compressed simplifications, we computed Spearman's rank correlation coefficient (Zwillinger and Kokoska, 1999) between compression rate and human ratings. For structural simplicity, the correlation was -0.17. For lexical simplicity, the value was slightly negative -0.01, while for overall simplicity it was slightly positive 0.05. Content preservation showed a strong correlation with the compression rate (0.43), and grammar had a moderately positive correlation (0.15).

5.3 LLMs as Evaluators

Table 6 and Table 7 present the results of each evaluator per question and simplified version, considering the hard and good sets, respectively.

Across both sets, the results show a consistent pattern among the three simplification models and the two LLM evaluators. In both sets, considering the average of the answers to the five questions, the same ranking emerged: Ours $>$ Qwen2.5 $>$ Bode, indicating agreement between the evaluators. For the hard samples, our model maintained high performance, particularly in terms of meaning preservation (P4) and fluency (P5), with average scores ranging from 4.6 to 4.9. Moderate but still strong scores for simplicity (P1), lexical (P2), and structural (P3) simplification indicate balanced performance. Qwen 2.5 achieved moderate results, doing better on lexical simplification (P2) but losing coherence in simplicity (P1) and preservation (P4). Bode consistently underperformed, with low scores across most dimensions.

The same trends appeared in the good samples: our model nearly maximized meaning preservation and fluency, Qwen2.5 showed moderate gains, and Bode remained weak. Overall, our model performed better on hard samples, widening the gap with Bode, while Qwen's results were mixed:

Set		O-simplicity	Lexical simplicity	Structure simplicity	Meaning	Grammar	Avg. char-ratio	Avg. words
Good	Ours	3.9	3.6	3.7	**4.7**	4.8	0.60	322.8
	Qwen2.5	**4.1**	**4.0**	**4.4**	4.2	**4.9**	0.22	63.6
	Bode	2.8	3.2	3.6	2.8	4.4	0.29	61.8
Hard	Ours	**4.6**	**4.2**	**4.7**	**4.3**	**4.8**	0.28	152.4
	Qwen2.5	4.1	4.0	4.4	3.9	**4.8**	0.15	79.2
	Bode	2.4	2.9	2.9	2.1	3.4	0.15	108.8

Table 5: Human evaluation results on the 40 selected document pairs. Bold indicates the best result.

Models	Crit.	GPT-OSS-120B	LLaMA-70B
Ours	P1	4.123 ± 0.055	4.333 ± 0.043
	P2	4.083 ± 0.105	4.873 ± 0.105
	P3	3.927 ± 0.128	4.333 ± 0.043
	P4	4.637 ± 0.113	4.812 ± 0.031
	P5	4.767 ± 0.154	4.900 ± 0.048
Qwen2.5	P1	3.453 ± 0.098	3.333 ± 0.043
	P2	3.988 ± 0.149	4.273 ± 0.245
	P3	3.955 ± 0.280	3.470 ± 0.175
	P4	3.575 ± 0.180	3.810 ± 0.033
	P5	4.735 ± 0.126	4.268 ± 0.141
Bode	P1	2.027 ± 0.206	1.677 ± 0.141
	P2	2.125 ± 0.217	2.723 ± 0.193
	P3	2.220 ± 0.315	1.808 ± 0.110
	P4	2.285 ± 0.052	2.288 ± 0.221
	P5	4.157 ± 0.186	3.185 ± 0.544

Table 6: Mean $\pm$ std in the hard set across runs per criterion and simplification of LLMs-as-judges.

Models	Crit.	GPT-OSS-120B	LLaMA-70B
Ours	P1	4.097 ± 0.195	4.192 ± 0.093
	P2	3.772 ± 0.158	4.602 ± 0.098
	P3	3.848 ± 0.264	4.192 ± 0.093
	P4	4.780 ± 0.092	4.800 ± 0.198
	P5	4.797 ± 0.144	4.933 ± 0.115
Qwen2.5	P1	3.735 ± 0.100	3.222 ± 0.179
	P2	4.240 ± 0.121	3.928 ± 0.208
	P3	4.348 ± 0.189	3.403 ± 0.192
	P4	3.573 ± 0.089	3.677 ± 0.220
	P5	4.785 ± 0.159	4.360 ± 0.056
Bode	P1	2.435 ± 0.380	2.580 ± 0.381
	P2	3.365 ± 0.110	2.923 ± 0.325
	P3	3.562 ± 0.234	2.758 ± 0.430
	P4	2.355 ± 0.053	2.883 ± 0.211
	P5	4.527 ± 0.193	3.762 ± 0.453

Table 7: Mean $\pm$ std in the good set across runs per criterion and simplification for LLMs-as-judges.

stronger with LLaMA and weaker with GPT.

We computed agreement scores across runs and models. Weighted Cohen's κ showed strong intra- and inter-model consistency: overall κ = 0.74 (95% CI 0.43–0.88) between GPT-OSS-120B and LLaMA-70B, and 0.85 (95% CI 0.64–0.96) on hard samples—indicating substantial to near-perfect agreement.

By criterion, looking at all the samples, the highest concordance appears for meaning preservation (P4, κ = 0.86) and simplicity (P1, κ = 0.80), while lexical, syntactic, and fluency criteria (P2, P3, P5) show moderate alignment (0.5–0.67). Agreement varies by text: highest for Ours (κ = 0.62), lower for Qwen 2.5 (0.21) and Bode (0.29), reflecting greater divergence on weaker simplifications. Intra-model reliability is excellent, with mean pairwise above 0.9 for both models, confirming highly stable judgments across runs.

5.4 Automatic Evaluation

For automatic evaluation, we used D-SARI, Semantic Similarity (SIM), and Flesch metrics. D-SARI (Sun et al., 2021), derived from SARI (Xu et al., 2016), assesses additions, deletions, and retentions with length penalties. Flesch and SIM measure readability and semantic similarity, respectively, as also applied in dataset filtering (Section 3.1). Results for LegalSim-PT and GovLang-BR appear in Table 8.

For the sentence-level dataset, all models achieve high semantic similarity; however, Portuguese LLMs underperformed in simplicity metrics compared to the Qwen family. The fine-tuned models increased the readability metric but showed a decrease in the D-SARI metric, mainly due to a reduction in the deletion component (D_{del}). This suggests that fine-tuning tends to make models less likely to remove complex or unnecessary words, as measured against the reference text.

Finally, results on LegalSim-PT show the challenge of preserving document-level semantics. The 4,096-token context limit may explain why Portuguese LLMs struggle to preserve content in long documents. Qwen models achieved notably higher simplicity scores, and their fine-tuned versions showed clear gains in readability and D-SARI, highlighting LegalSim-PT's strong potential.

	LegalSim-PT						Gov-Lang-BR					
	SIM	Fles.	D-SARI	D_{add}	D_{del}	D_{keep}	SIM	Fles.	D-SARI	D_{add}	D_{del}	D_{keep}
Bode	.464	15.9	28.98	0.48	79.21	7.26	**.903**	-8.17	21.60	2.16	37.46	25.21
Tucano	.724	27.3	18.42	1.42	43.00	10.83	.872	-1.23	13.09	2.80	26.35	10.13
Qw2.5	.653	31.8	35.74	3.11	**83.33**	20.78	.873	0.71	**35.58**	6.22	**66.89**	**33.63**
Qw2.5FT	**.810**	**42.7**	**50.78**	**23.4**	81.18	**47.72**	.893	**6.46**	30.11	**7.05**	54.27	29.01
Qw3	.772	3.73	32.16	2.62	76.53	17.33	.864	4.14	30.44	4.43	63.12	23.77
Qw3FT	.780	30.6	37.22	10.7	74.63	26.30	.857	5.69	25.46	4.96	52.93	18.48

Table 8: Results on the LegalSim-PT and GovLang-BR test sets. Bold indicates the best result.

6 Conclusions

This study advances legal document simplification by introducing LegalSim-PT, a large-scale, high-quality dataset, along with a comprehensive evaluation framework combining quantitative and qualitative analyses. We benchmark several representative models, and the results confirm both the dataset's quality and the reliability of the metrics. Future work includes extending the framework to other domains, such as healthcare, incorporating discourse-level metrics, and investigating how compression rate affects the meaning preservation.

Limitations

A key limitation of this study is that the simplified dataset is entirely machine-generated. Although three native Portuguese speakers (two linguists and one law graduate) evaluated the outputs, they assessed only a representative sample and did not annotate or refine the full corpus. As a result, structural biases inherent to LLMs may remain undetected. Future work should incorporate professional legal editors in a human-in-the-loop framework to produce a fully human-refined gold standard. Nonetheless, adopting a synthetic pipeline is a pragmatic choice given the low-resource status of Portuguese in the legal domain, where large-scale expert annotation is costly and difficult to obtain.

Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and CAPES. We also thank the support of the CNPq National Institutes of Science and Technology, IAIA (grant 406417/2022-9), TILD-IAR (grant 408490/2024-1) and IAPROBEM (grant 408589/2024-8).

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Rodrigo Alarcón, Lourdes Moreno, and Paloma Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proc. of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proc. of the 8th ACM symposium on Document engineering*, pages 240–248.

Alexandre Alves, Péricles B. C. Miranda, Rafael Fe Mello, and André C. A. Nascimento. 2023. Automatic simplification of legal texts in portuguese using machine learning. In *Legal Knowledge and Information Systems - JURIX 2023: The 36th Annual Conference*, volume 379 of *Frontiers in Artificial Intelligence and Applications*, pages 281–286. IOS Press.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for german text simplification: Overcoming parallel data scarcity through style-specific pre-training. In *Findings of the ACL: ACL 2023*, pages 1147–1158.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *arXiv preprint arXiv:2410.21306*.

Gabriel Assis, Cláudia Freitas, and Aline Paes. 2025. Exploring brazil's llm fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):940–972.

Luis Cabrera. 2024. Babel fish democracy? prospects for addressing democratic language barriers through machine translation and interpretation. *American Journal of Political Science*, 68(2):767–782.

M Charles. 2013. Active and passive voice in research articles: An interdisciplinary study. *International Journal of Corpus Linguistics*, 18(3):279–318.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Letícia MS Corrêa, Erica dos S Rodrigues, and Renê Forster. 2019. On the processing of object relative clauses. *ExLing 2019*, 25:57.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the ACL: ACL 2023*, pages 13190–13206. ACL.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006. ACL.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. *CoRR*, abs/2404.03278.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345.

Eduardo A. S. Garcia, Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza, and Eliomar A. Lima. 2024a. RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In *Proc. of the 16th Int. Conf. on Computational Processing of Portuguese*, pages 374–383. ACL.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis C. S. Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024b. Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *Preprint*, arXiv:2401.02909.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proc. of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. Multimsd: A corpus for multilingual medical text simplification from online medical references. In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 9248–9258. ACL.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J. Ramanathan, Wei Xu, Byron C. Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 16662–16692. ACL.

Hong Jin Kang, Muhammad Ali Gulzar, Nanyun Peng, Miryung Kim, and 1 others. 2024. Human-in-the-loop synthetic text data inspection with provenance tracking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3118–3129.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Sidney Evaldo Leal, Magali Sanchez Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources & Evaluation*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: multilingual unsupervised sentence simplification by mining paraphrases. In *Proc. of the 13th Language Resources and Evaluation Conference, LREC 2022*, pages 1651–1664. ELRA.

Paloma Martínez, Lourdes Moreno, Hiram Ochoa, Alberto Ramos, and Mario Pérez-Enríquez. 2024. A tool suite for cognitive accessibility leveraging easy-to-read resources and simplification strategies. *CEUR-WS. org*.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for Japanese. In *Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476. ELRA and ICCL.

Francielle Vasconcellos Pereira, Ana Frazão, and Viviane P. Moreira. 2024. Automatic text simplification for the legal domain in brazilian portuguese. In *Intelligent Systems - 34th Brazilian Conference, BRACIS 2024, Proceedings, Part IV*, volume 15415 of *LNAI*, pages 31–45. Springer.

Nelson Perez-Rojas, Saúl Calderón Ramírez, Martin Solis-Salazar, Mario Romero-Sandoval, Monica Arias-Monge, and Horacio Saggion. 2023. A novel dataset for financial education text simplification in spanish. *CoRR*, abs/2312.09897.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing*. ACL.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023a. Revisiting non-english text simplification: A unified multilingual benchmark. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927.

Michael J. Ryan, Tarek Naous, and Wei Xu. 2023b. Revisiting non-english text simplification: A unified multilingual benchmark. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 4898–4927. ACL.

Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A. De Souza, Maria José Bocorny Finatto, and Aline Paes. 2025. Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477. ACL.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013. ACL.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14.

Verónica M Vargas. 2023. Analysis of barriers and proposals for inclusive access to justice for vulnerable groups. *Journal of Law and Epistemic Studies*, 1(2):20–24.

Laura Vásquez-Rodríguez, Nhung T. H. Nguyen, Piotr Przybyla, Matthew Shardlow, and Sophia Ananiadou. 2024. Simple is not enough: Document-level text simplification using readability and coherence. *CoRR*, abs/2412.18655.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proc. of the 2nd Workshop on Text Simplification, Accessibility and Readability*, pages 85–101. INCOMA Ltd., Shoumen.

Leandro Wives and Maria José Finatto. Usando llms para simplificar e representar documentos médicos. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados*, pages 415–425. SBC.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ziyu Yang. 2024. *Enhancing the Comprehension: Text Simplification Approaches and the Role of Large Language Models*. Temple University.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605. ACL.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proc. of the AAAI conference on artificial intelligence*, volume 34, pages 9709–9716.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Annual Meeting of the Association for Computational Linguistics (61st: 2023)*, pages 319–337. ACL.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

A Compression Ratio Distributions

Figure 3 shows the distributions for each of the six subsets that compose LegalSim-PT.

B Human Evaluation Details

During the human evaluation process of our dataset, 40 original documents and three simplified versions of each document were distributed to three evaluators. The origin of the simplified versions was not disclosed to them. The three simplifications of a given document were evaluated by only one evaluator. For each simplified version, five questions addressing simplicity, content preservation, and grammaticality were answered using a 1–5 Likert scale. Table 9 presents the five evaluation questions. An English translation of the questions was also provided.

Type	Question
General Simplicity	**PT:** O texto simplificado é mais simples que o texto original, sob a condição de garantia de qualidade? Ou seja, ele também deve ser fluido na leitura e preservar o significado principal do texto original. **EN:** Is the simplified text simpler than the original text while maintaining quality? That is, it should be fluent and preserve the main meaning of the original text.
Lexical Simplicity	**PT:** Qual é o grau de simplificação lexical, ou seja, o quanto as palavras originais foram substituídas por termos mais simples? **EN:** What is the degree of lexical simplification, that is, to what extent were original words replaced with simpler terms?
Syntactic Simplicity	**PT:** Qual é o grau de simplificação estrutural, ou seja, o quanto a organização e a complexidade sintática das sentenças foram reduzidas? **EN:** What is the degree of structural simplification, that is, to what extent were sentence organization and syntactic complexity reduced?
Content Preservation	**PT:** O documento simplificado preservou o significado principal do texto original? Fatores que podem impactar nessa métrica são a remoção de conteúdo indispensável ou a inserção de informações novas não contidas no documento original e que não são de senso comum. **EN:** Did the simplified document preserve the main meaning of the original text? Factors that may impact this metric include the removal of essential content or the insertion of new information not contained in the original document and not considered common knowledge.
Grammaticality	**PT:** O texto simplificado é gramaticalmente correto e fluente, garantindo que as sentenças sejam naturais e bem formadas? **EN:** Is the simplified text grammatically correct and fluent, ensuring that sentences are natural and well formed?

Table 9: Human evaluation questions.

C LLM as Evaluators Details

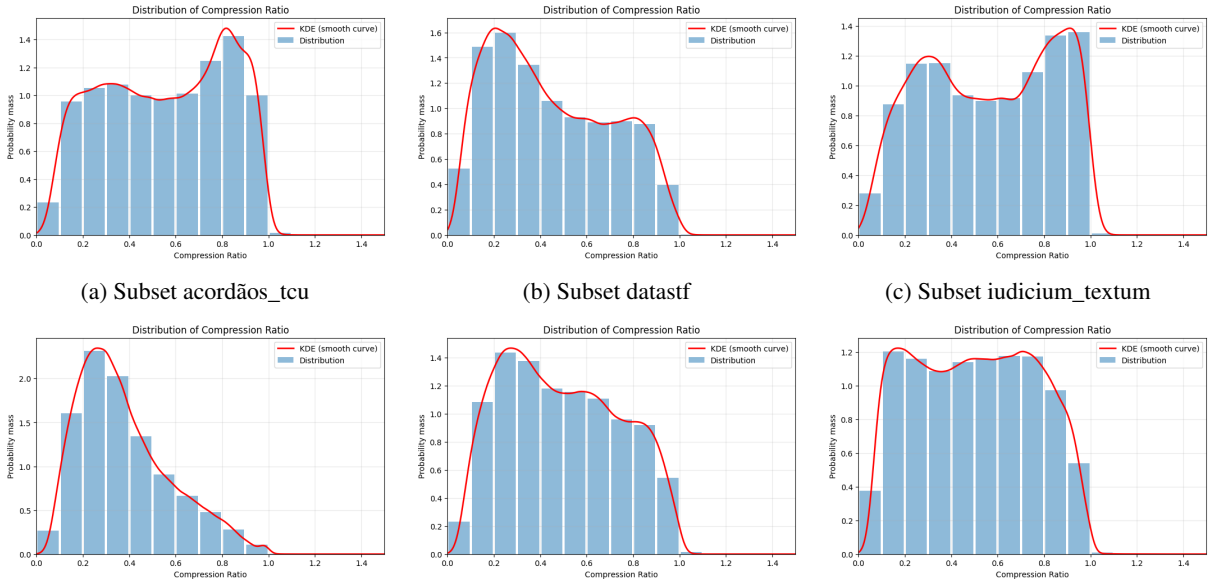
Figure 4 shows the system prompt used when employing LLMs as evaluators. The full prompt code is available in our GitHub repository.

D Inference and Fine-tuning Details

We provide technical details about the four reference models evaluated in the LegalSim-PT and Gov-Lang-BR benchmarks. Two of these models were pretrained primarily on Portuguese data (Bode and Tucano), while the other two belong to the Qwen family (Qwen3-1.7B and Qwen2.5-7B).

The Qwen3-1.7B model was executed using the Unsloth framework (Daniel Han and team, 2023) for optimized inference and fine-tuning efficiency. The remaining models were implemented using the Hugging Face Transformers library. Table 10 presents the model names along with some inference parameters.

Both Qwen models were also fine-tuned. For Qwen2.5-7B-Instruct, we used QLoRA-style (Dettmers et al., 2023) parameter-efficient training with 4-bit NF4 quantization (double quantization enabled) and bf16 computation. Gradient checkpointing was activated to reduce memory usage, and optimization was performed with Paged AdamW 8-bit. LoRA adapters were applied to the attention projection layers (q_proj, v_proj) with rank r=16, α =16, and dropout 0.05. Training was



(a) Subset acordãos_tcu (b) Subset datastf (c) Subset iudicium_textum

(d) Subset BRCAD-5 (e) Subset CJPG (f) Subset tesemo_v2

Figure 3: Probability distributions of compression ratios for the six LegalSim-PT subsets.

conducted for about 1.5 epochs with per-device batch size 1 and gradient accumulation of 24 steps (effective batch size = 24), learning rate 2×10^{-4}, cosine scheduler with 100 warmup steps. Qwen3-1.7B was fine-tuned using Unsloth with 4-bit quantization. Training used LoRA adapters with rank r=16, α =16, and dropout 0.05, targeting the attention and MLP projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj), with Unsloth-optimized gradient checkpointing enabled. We trained for about 2.5 epochs with per-device batch size 16 and gradient accumulation 16 (effective batch size = 256), learning rate 2×10^{-4}, linear scheduler with warmup ratio 0.03, weight decay 0.01. For both Qwen models, only LoRA parameters were updated during training, the maximum sequence length was capped at 4,096 tokens for training and 2,048 for evaluation, and the best model was selected based on validation loss. More details about inference and fine-tuning procedures can be found in our github repository.

Prompt

Você é um avaliador linguístico especializado em simplificação de textos e preservação semântica. Sua tarefa é avaliar o quão bem cada uma de três versões simplificadas atende a cinco critérios linguísticos. Use SEMPRE a escala 1–5 (1=ruim, 5=excelente) e gere SAÍDA ESTRITAMENTE EM JSON.
Critérios: P1. Simplicidade com preservação do significado e fluidez. P2. Simplificação lexical (troca por termos mais simples). P3. Simplificação estrutural (redução de complexidade sintática). P4. Preservação do significado (sem omissões essenciais ou adições irrelevantes). P5. Correção gramatical e fluência.
Para CADA item fornecido (com id), avalie as versões 1, 2 e 3 e produza o seguinte JSON por item: "id": "ID_DO_ITEM", "Versao_1": "P1": "nota": int, "justificativa": "string", "P2": "nota": int, "justificativa": "string", "P3": "nota": int, "justificativa": "string", "P4": "nota": int, "justificativa": "string", "P5": "nota": int, "justificativa": "string", "Comentário_geral": "string" , "Versao_2": ... mesmo formato ... , "Versao_3": ... mesmo formato ... , "Ranking_geral": "ordem_melhor_para_pior": ["Versao_X", "Versao_Y", "Versao_Z"], "justificativa": "string breve"
A saída FINAL deve ser um ARRAY JSON com um objeto por item, na MESMA ORDEM de entrada. Não inclua explicações fora do JSON. Não use markdown nem blocos de código. Se algum texto for muito curto para avaliar, ainda assim dê notas e explique a limitação na justificativa. Seja determinístico e consistente entre itens. Evite aleatoriedade.

Figure 4: System prompt provided to the LLMs. The original prompt is written in Portuguese.

Model	**Inference Configuration**
`recogna-nlp/bode-7b-alpaca-pt-br`	Quantization: 8-bit ; Context length: 4,096 tokens; Max new tokens: 1,024; Decoding strategy: Greedy.
`TucanoBR/Tucano-2b4-Instruct`	Precision: FP16 ; Context length: 4,096 tokens; Max new tokens: 1,024; Decoding strategy: Sampling .
`Qwen/Qwen2.5-7B-Instruct`	Quantization: 4-bit ; Context length: 32,768 tokens; Max new tokens: 2,048; Decoding strategy: Greedy .
`unsloth/Qwen3-1.7B-bnb-4bit`	Quantization: 4-bit ; Context length: 4,096 tokens; Max new tokens: 1,024; Decoding strategy: Greedy .

Table 10: Reference models and inference parameters used in the LegalSim-PT and Gov-Lang-BR experiments.