# Automatic Text Simplification for the Legal Domain in Brazilian Portuguese

Francielle Vasconcellos Pereira[1], Ana Frazão[2], and Viviane P. Moreira[1]

[1] Institute of Informatics, UFRGS Porto Alegre, Brazil
{fvpereira,viviane}@inf.ufrgs.br
[2] USP São Paulo, Brazil anarosapaiva@usp.br

**Abstract.** Legal and juridical documents such as rulings, laws, agreements, and contracts contain domain-specific terms and jargon, long and complex sentences that may be difficult to understand for laypeople without domain expertise, reading issues, or with a low education level. The simplification of these documents has been a concern for several years, aiming to democratize access to justice. Courts are already adopting simpler language, especially in documents aimed at laypeople, such as warrants and notifications, to enhance inclusion and clarity. Automatic textual simplification, a subfield of Natural Language Processing, seeks to make complex texts more accessible. This paper explores the task of automatic text simplification in Portuguese for the legal domain. The main challenge here is the lack of datasets containing complex sentences and their simplified versions. This work investigates how existing datasets, methods, and metrics used for text simplification perform applied to legal texts in Portuguese. We present qualitative and quantitative analyses using five models. The results show that GPT-based models have the best results, but fine-tuning with domain data is a viable open-source alternative.

**Keywords:** Automatic Text Simplification · Legal Texts · Natural Language Processing · Plain Language.

## 1 Introduction

Judicial and legislative texts have distinctive characteristics, such as the use of Latin and domain-specific language, that make them difficult to understand. The use of judicial language can be found in documents that describe laws, rulings, opinions, and other legal process documents. When it comes to forensic documents, specific terminology is necessary to understand the facts without leaving margin for different interpretations.

The concern about simplifying legal has been ongoing for some years. The need for the democratization of the Brazilian Judiciary, expanding access to Justice, has been raised by a few studies [1,4,18]. However, they focus on advocating that documents are written and published in a simplified manner. Writing simply without losing the entire essence of the text that the law requires is a practice that would need to be widely disseminated and encouraged in all courts in the country.

To encourage courts to use plain language, the CNJ (*Conselho Nacional de Justiça*) launched the National Pact of the Judiciary for Plain Language[3] and, through Ordinance No. 351/2023, the Plain Language Seal. Its purpose is to recognize and promote, across all segments of the judiciary and at all levels of jurisdiction, the use of clear and understandable language in the drafting of judicial decisions and in general communication with society.[4] This is a lengthy process that may take years to materialize, still, it can benefit from advancements resulting from the use of automatic tools for text simplification.

Over time, Brazilian Portuguese has evolved to exhibit distinct characteristics not typically found in European Portuguese. Brazil, with its vast territory and cultural diversity, encountered challenges in the development of its legal framework. The incorporation of influences from Roman-Germanic, French, and Portuguese legal traditions occurred within the context of adapting to the specificities of Brazilian society. Consequently, Brazil's legal system represents a fusion of these legal traditions and an adaptive approach to local issues. These characteristics undoubtedly influence legal writing practices, affecting the formats of documents, hermeneutics, and argumentation in various ways [10, Chapter 26].

The challenge addressed in this work is how to automatically simplify texts in the legal domain written in Brazilian Portuguese. Text simplification (TS) is a subfield of Natural Language Processing (NLP). It seeks to translate complex texts into a simpler language, aiming for accessibility. There are several reasons to simplify texts, such as covering the understanding of non-native speakers, people with a lower level of education, or cognitive disabilities.

We apply Large Language Models (LLM) to generate simplifications given a complex sentence from a legal text. We experimented with different approaches, including fine-tuning a T5 model and prompting strategies. In addition, we enhanced a fine-tuned model with reinforcement learning (RL) to assess whether this could improve simplification results.

The evaluation of TS is still an open problem. There is no single metric that can accurately gauge all nuances that are involved in simplifying a text. As a result, this work provides a quantitative analysis using four evaluation metrics. We also performed a qualitative analysis, which is crucial to allow insights into the generated simplifications.

The main contributions of this work can be summarized as follows:

1. an evaluation of five LLMs applied to text simplification in the Brazilian legal domain;
2. a quantitative assessment of the quality of the simplification using four evaluation metrics; and
3. a qualitative evaluation of the model's outputs by a human expert.

[3] https://www.cnj.jus.br/gestao-da-justica/acessibilidade-e-inclusao/pacto-nacional-do-judiciario-pela-linguagem-simples/

[4] https://www.cnj.jus.br/gestao-da-justica/acessibilidade-e-inclusao/pacto-nacional-do-judiciario-pela-linguagem-simples/selos/

Our findings showed that GPT models still have the best results for TS, even in this specific domain. The use of RL brought performance improvements to the fine-tuned. Apart from the GPT models, using only instructions and examples is not enough to yield good results.

## 2 Related Work

Legal documents have complex and specific terms in any language. A compilation of existing datasets, methods, and metrics was investigated for TS in legal texts [11]. They reported the results in terms of readability, simplicity, similarity, presence of hallucinations, and fluency methods. Their analysis showed that most methods focus on split-and-rephrase, transforming larger and more complex sentences into shorter and simpler ones.

In general, research in languages other than Portuguese includes diverse methods, such as machine translation (both supervised and semi-supervised). The first relevant work in TS using a supervised method employed metrics as rewards in an RL algorithm called DRESS (**D**eep **RE**inforcement **S**entence **S**implification) [27]. Another work focuses on the replacement of complex words and on sentence splitting. They outperform other unsupervised methods for TS for different domains in English [6].

More recently, BLESS (Benchmarking Large Language Models on Sentence Simplification) evaluated 44 models for TS in three datasets (Wikipedia, news, and medical). They showed that some pre-trained models, even if not trained to TS, can perform well compared to the MUSS [15] state-of-the-art chosen as a baseline in the paper [13]. There is no consensus on what constitutes plain language, but there are principles that guide the practice of TS. The basic book of Plain Language theory [8], establishes 25 guidelines capable of guaranteeing the clarity of the text.

Brazilian Courts of Justice are working to implement plain language both in textual documents and in-person services provided to citizens. These initiatives aim to ensure that information is passed on clearly and objectively without the need to resort to third parties (such as lawyers). The main targets are documents that reach lay people, such as search warrants and subpoenas.[5]

There are few linguistic resources for TS in Portuguese [9,12,16], and some contributions to NLP and the legal domain made recently [19,23]. However, in Portuguese, most approaches are limited only to lexical simplification, not encompassing the entire *syntax* of the text. Lexical methods in TS are restricted to dealing with complex words in sentences. Just changing complex words is not enough to simplify legal sentences. Even in English, there is a lack of available data for TS in the domain, and the syntactic methods focus on splitting sentences into shorter ones.

An approach that considers the entire syntax of sentences in Portuguese was developed in PorSimples [14]. The project started in 2009, and in 2010, a

[5] https://comunicasimples.com.br/2023/01/03/juridiques/

rule-based system was developed and made available but is not currently being maintained. The authors also contributed with the only Portuguese parallel dataset containing original and simplified sentences.

A recent work presented Machine Learning methods for TS in judicial summaries of two Brazilian courts [3]. However, they reported results of readability metrics only, and those are not commonly used to evaluate TS systems.

This paper complements the previous studies by assessing TS in five LLMs applied to legal texts, comparing them with documents simplified by human specialists in Brazilian courts. It is unclear how existing models with different architectures can perform in Portuguese. The legal domain presents difficulties and a lack of data, even in English. We aim to answer whether a dataset for TS in Portuguese is enough to generalize and allow simplifying legal texts, and how some models can perform the task in this specific domain. To the best of our knowledge, this is the first work to evaluate the domain in Portuguese using metrics that take reference sentences into account.

## 3 Materials and Methods

This section describes the macro-steps for automatic TS we adopted in this work to answer the following research questions:
*RQ1* Are the main dataset for TS and main pre-trained models in Portuguese able to generalize and be applied to the legal domain?
*RQ2* How well do state-of-the-art and recently launched LLMs perform TS in the chosen domain and language?
*RQ3* Can RL improve TS results for fine-tuned models?

Figure 1 shows an overview of our pipeline. Our input comes from the three sources of data that were combined into a merged dataset. This dataset is then used in different ways: (*i*) for providing training data for fine-tuning a Transformer-based model and (*ii*) enriching prompts sent to generative models. The outputs produced by the models are evaluated against the reference simplified sentences.

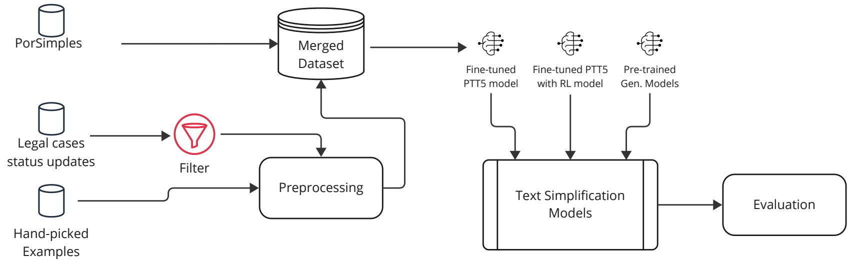

**Fig. 1.** Pipeline with the steps in our methodology.

### 3.1 Assembling a TS Dataset in Portuguese

The input instances for training and evaluating TS systems consist of a pair $\langle O, R \rangle$, where $O$ is the original (complex) sentence and $R$ is its simplified version. Unfortunately, there is no such dataset for the legal domain in Portuguese. As a result, we assembled a dataset for training and evaluating TS models, combining data from different sources (some in the legal domain and some generic). The following three sources of data were used.

***PorSimples*** [14] created four baselines referred to as PorSimplesSent. They contain a parallel corpus with two levels of simplification, *natural* and *strong*. The texts were extracted from Brazilian news articles. Each pair $\langle O, R \rangle$ is labeled according to the level of simplification, which can be $N$ (natural simplification) or $S$ (strong simplification). The difference between them is the degree of operations made into the original sentences. In $N$, splitting and inversion of clause ordering are applied discreetly, and in $S$, all mapped rules are applied in the simplification (rewriting and lexical substitution, change to canonical order, change to active voice, reordering, splitting, joining, and dropping). All combinations of the pairs in this dataset were used in the fine-tuning step, totaling 8,120 pairs. The combinations are O→N, O→S, and N→S, where the text on the left is equal to or more complex than the text on the right.

***Legal case status updates*** is a dataset assembled by JusBrasil,[6] a Brazilian company that provides access to information on legal cases. They used an OpenAI generative model to explain the meaning of status updates such as *"Remetidos os Autos (em diligência) para Central de perícia"*. There are 1,656 explanations. Each explanation is annotated by experts who rated their quality according to different aspects, namely: (*i*) whether the explanation is legally accurate and if it is somehow useful. The possible answers were 'yes', 'no', 'partially', or 'not applicable'; and (*ii*) the quality of the explanation. The possible answers were 'good', 'bad', 'partially', or 'not applicable'.

We discarded the 231 instances annotated with 'bad' quality or that were not considered legally accurate or useful. The remaining 1,424 instances were used in the fine-tuning step.

***Hand-picked examples.*** We manually selected 149 pairs of sentences $\langle O, R \rangle$ from materials published by justice courts that are involved in plain language projects promoted by the Brazilian government. The list of these sources can be found in.[7]

Instances from the three sources were merged and used as training data. The test data consists of 91 instances from the hand-picked dataset that are from the legal domain. The training, validation, and test splits were disjoint. Some statistics of the datasets are in Table 2. Some examples of the dataset structure are in Table 1, and statistics of the datasets are in Table 2.

[6] https://www.jusbrasil.com.br

[7] https://l1nk.dev/yghQP

**Table 1.** Example of the dataset rows used in the work

| **Level** | **Complex** ($O$) | **Simple** ($R$) |
|---|---|---|
| ORI→NAT | Autores de furto estariam migrando para o roubo. | Autores de furto estariam mudando para o roubo. |
| ORI→STR | O CNJ enviou a demanda de informações à SECTI. | O Conselho Nacional de Justiça pediu as informações à Secretaria de Tecnologia da Informação. |

**Table 2.** Statistics of the dataset used for fine-tuning

| **Number of instances** | **O→N** | **O→S** | **N→S** | **Total** |
|---|---|---|---|---|
| Total | 2,931 | 2,570 | 4,101 | 9,602 |
| Training | 2,631 | 2,318 | 3,692 | 8,641 |
| Validation | 300 | 252 | 409 | 961 |

### 3.2 Machine Learning Models

We employed four widely used Transformer-based models to perform TS. Both decoder-only and encoder-decoder models were used.

**PTT5**[5] is a T5 [20] model pre-trained on the BrWac corpus [24], a large collection of web pages in Portuguese. It improves the performance of T5 on Portuguese sentence similarity and inference tasks. It is available in three sizes (small, base, and large) and with two vocabularies. We used the base model as an initial baseline. PTT5 can also be used with instructions only (and no fine-tuning), but this approach did not work well on this model, as the final outputs were just a copy of the prompt. Thus, only the fine-tuning alternative of PTT5 is reported here and is labeled as **FT-PTT5**.

***Flan-T5-Large*** [7] is also a multilingual model from the T5 family. It has 700 million parameters and continues the training from an adapted T5-LM checkpoint. It uses a wide variety of labeled data to fine-tune with instructions, but the data is not specific to the TS task.

***GPT-3.5-Turbo*** and ***GPT4o*** The GPT models originate from OpenAI's research in NLP, with GPT-3.5-Turbo and GPT-4o featuring advancements such as increased parameters and context windows.

The GPT-3.5-Turbo and Flan models were chosen based on a benchmarking of generative models for TS [13]. To the best of our knowledge, GPT-4o was not tested for TS due to the recent launch. These three models were used based solely on prompting without any fine-tuning.

### 3.3 Fine-tuning and Reinforcement Learning Approach

Fine-tuning is the process of adjusting a pre-trained language model on a specific dataset for a particular task or domain. In this work, the TS task is the target. The fine-tuning was done using the PTT5 model.

**Fine-Tuning Procedure** To fine-tune PTT5 to perform TS, we followed the splits described in Table 2. The batch size and batch per device in training were set to 4 sentences due to infrastructure limitations. We used 3 steps of gradient accumulation and a batch size of 64 per device for evaluation. The learning rate was 1e-4, weight decay was 0.01, and training was for 100 epochs with checkpoints at every 3000 steps. The loss function was used to choose the best model, and fp16 (half-precision floating point) was set to true. The SARI metric was calculated during validation.

**Reinforcement Learning** This fine-tuned model was used to apply Reinforcement Learning (RL) based on DRESS [27]. We refer to the resulting model as **FT-PTT5 + RL**. However, DRESS used an LSTM to generate the outputs and then applied RL. In the context of text tasks, RL is used as additional training for pre-trained models. A common technique is to linearly interpolate the RL reward with cross-entropy loss to avoid erroneous training due to the large action space.

During RL, we used metrics that do not rely on reference simplifications. These metrics are different from the ones used for evaluating the quality of the predictions in the test set (Section 3.5) since one cannot optimize and evaluate using the same metrics.

1. **FKGL** — Flesch-Kincaid Grade Level is used to evaluate the readability of a text, indicating the level of education necessary to understand it easily. It considers the average number of words per sentence and the average number of syllables per word, offering a score corresponding to years of schooling to understand a text. As SARI is the main metric to automatically evaluate outputs in TS, FKGL was calculated as a simplicity reward in RL.
2. **SAMSA** — Semantic Annotation for Machine Reading focuses on the ability of the system to understand and extract accurate semantic information from complex texts. Using semantic annotations, such as named entities and relationships between concepts, SAMSA seeks to quantify the quality of the system interpretation by identifying whether it correctly captures the essential meaning of the text [21]. SAMSA assigns high scores to split sentences, which is a simplification action cited as one of the most important in TS [14].
3. **Levenshtein Distance** — calculates the difference between two strings. This measure quantifies the minimum number of operations necessary to transform one sequence into another, where valid operations are insertion, deletion, or replacement of a single character. This metric is widely used in word processing applications such as spelling correction and plagiarism detection, providing an efficient way to compare and measure similarity between different strings. The calculation of this reward aims to preserve the meaning of the original sentence.

In the RL step, the model is trained to increase a reward. The PTT5 fine-tuned model is seen as an agent; it reads the source sentence and then takes an action according to a policy. The agent generates the output sequence as the simplified text; a reward is calculated, and the reinforcement updates the agent state. The reward $r(\hat{Y})$ used in this work followed the base reward of DRESS [27] and is calculated as in Equation 1 according to three perspectives: simplicity ($r_S$), relevance ($r_R$), and fluency ($r_F$):

$$r(\hat{Y}) = \lambda^S r^S + \lambda^R r^R + \lambda^F r^F \tag{1}$$

where $r^S$ is the FKGL score for simplicity, $r^R$ is the Levenshtein Distance for meaning preservation and $r^F$ is the SAMSA score for fluency. $\lambda^S$, $\lambda^R$ and $\lambda^F$ are constants $\in [0, 1]$ and were set as 1, 0.25 and 0.5, respectively.

### 3.4 Generative Approach

The lack of a parallel dataset to allow fine-tuning or transfer learning into the several versions of LLMs available makes the generative models a good alternative. The instructions do not need many examples to enable the model to understand the main goal of a task. This is exactly the scenario we have for the Portuguese legal domain. We adapted the prompt from BLESS [13], which uses the few-shot in-context learning and achieved good results in three domains in English. The prompt we used was written in Portuguese, but the translated instruction is presented in Figure 2.

Reescreva a frase complexa para facilitar o entendimento para pessoas que não são da área jurídica. Você pode fazer isso substituindo palavras complexas por sinônimos mais simples, excluindo informações sem importância, reordenando informações e/ou dividindo uma frase complexa longa em várias outras mais simples.

Complexa: {complex example of the domain}
Simples: {simplified version of the complex}

Complexa: {complex input of interest}

**Fig. 2.** Structure of the instruction submitted to the generative models.

### 3.5 Evaluation Metrics

The evaluation of TS is an open problem. Metrics designed for readability and levels of text understanding are mainly used in linguistics research about the task, but they are not adequate to assess simplification results.

FKGL is an example of such a metric. Its score is highly sensitive to sentence length and, in some cases, provides only minor impacts on other metrics, like

SARI [22]. A sentence with many token deletions can cause a loss of important information and may not be able to represent the real meaning of the text message.

BLEU and SARI are also being widely used to evaluate the quality of TS. BLEU aims to correlate grammatically with human perception, in addition to preserving meaning. SARI is a good option to get an overview of simplicity, but it has better results assessing lexical paraphrasing systems [2].

BERTScore has good results at identifying when the reference sentences are similar to the system outputs. It is also a good option to evaluate meaning preservation and similarities with the reference, but a high score does not mean the output was simplified [2].

The metrics we used for output evaluation are briefly described next.

1. **SARI** — System for Automatic Evaluation is an automatic metric that evaluates how well a system preserves meaning while making text accessible and easier to understand. The metric is calculated through precision, recall, and F1-score for unigrams, bigrams, and trigrams, incorporating semantic similarity [25]. SARI is calculated on the n-grams added and kept, and the proportion of those deleted.
2. **BLEU** — Bilingual Evaluation Understudy is a metric used to evaluate machine translation systems. It measures n-gram overlap between the output and one or more human references. The closer the BLEU score is to 1, the closer the output is to the human reference according to syntax (word choices and their order) [17].
3. **BERTScore** — is a metric for evaluating the quality of text generation by comparing the generated text to a reference text. Unlike traditional metrics like BLEU or ROUGE, which rely on exact n-gram length, BERTScore leverages contextual embeddings from the BERT model. He has a better evaluation of meaning preservation because of the BERT background, computing a similarity score between each token in the candidate and reference texts, capturing semantic nuances and contextual meaning [26].
4. **ROUGE** — Recall-Oriented Understudy for Gisting Evaluation assesses the overlap of n-grams, as well as the overlap of word sequences and word pairs, between the generated text and a reference text. By focusing on recall, ROUGE emphasizes capturing as much of the reference text content as possible.

## 4 Experiments and Results

The experiments carried out in this work aim to answer the research question posed in Section 3.

### 4.1 Quantitative Evaluation

Our quantitative evaluation was based on the metrics described in Section 3.5. This validation was made over the 91 hand-picked instances that are from legal texts (Section 3.1). The average value of the metrics obtained by each model is shown in Table 3 and gives us a brief overview to answer *RQ2*.

**Table 3.** Average values of each evaluation metric on the instances from the test set

| | BLEU | SARI | BERTScore | | | ROUGE | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 |
| FT-PTT5 | .42 | .40 | .89 | .83 | .86 | .74 | .70 | .65 |
| FT-PTT5 + RL | .51 | .36 | .93 | .91 | .92 | .88 | .90 | .84 |
| GPT-3.5-Turbo | .15 | .42 | .84 | .81 | .83 | .74 | .75 | .75 |
| GPT-4o | .19 | **.43** | .83 | .82 | .83 | .69 | .68 | .67 |
| Flan-T5-Large | **.66** | .37 | **.95** | **.96** | **.96** | **.96** | **.97** | **.96** |

SARI is the main metric used to evaluate TS, while BERTScore, BLEU, and ROUGE are used to measure the meaning preservation of the simplified text. GPT models and FT-PTT5 obtained the highest values for SARI, indicating that the sentences are simpler to understand. The winner here is GPT-4o, although when looking at BLEU scores, the two GPT models have lower scores in comparison to all T5 models, which means they did not preserve the syntax of the original sentences.

All the T5 family has high BLEU scores. However, this changes when we look at BERTScore and ROUGE values. The GPT models are still outperformed by the T5 models. According to BERTScore and ROUGE, GPT-4o and GPT-3.5-Turbo are good options for TS, as they have the highest simplicity scores while preserving meaning.

Flan-T5-Large is the winner in meaning preservation; it has the best scores for similarity with reference simplifications, but SARI is the penultimate score. The generated simplifications are really discrete, not making significant changes to the original sentence, as shown in the following example: **Original:** "Nada veda que a declaratória seja ajuizada em conexão com o pedido constitutivo ou condenatório.". **Simplified:** "Nada proíbe que a declaração seja julgada em conexão com o pedido constitutivo ou condenatório."

Figure 3 reveals that GPT-4o may be the best model, with a slight advantage over GPT-3.5-Turbo. FT-PTT5 has better results for SARI compared to FT-PTT5+RL. RL relied on FKGL as the simplicity metric, and these results reinforce the idea that FKGL is not suitable to measure TS.

The scores of the model with RL and Flan-T5-Large are quite similar, with a slight advantage to Flan-T5-Large. As the results for meaning preservation are a little better than the PTT5 fine-tuned, it seems that Levenshtein Distance is also a good alternative to check the similarity between the sentences.

### 4.2 Qualitative Evaluation

To provide insights into the quality of the generated simplifications, the outputs of the TS methods were analyzed by a domain specialist. The annotator is a

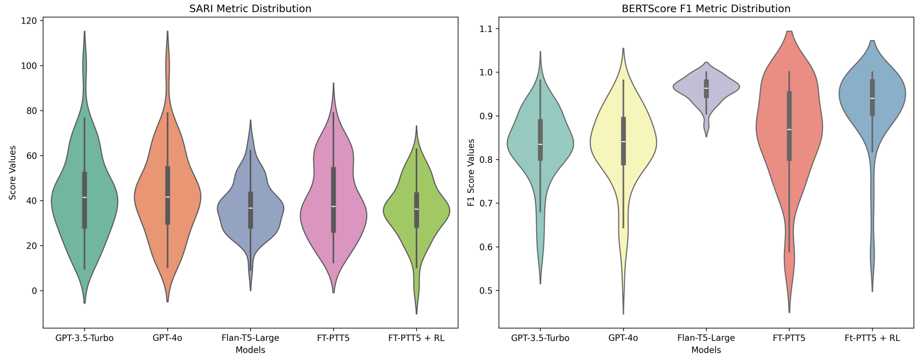

**Fig. 3.** Results of each model on SARI and BERTScore, evaluating simplicity and meaning preservation, respectively.

judicial analyst who takes on the role of editorial reviewer in a Brazilian Court of Justice.

The sample selected for annotation contains 91 pairs for each of the five simplification methods, containing the original sentence and the simplification output of each method adopted. The annotator was not aware of which method produced the simplification. The specialist received the original and the simplified sentences and was asked to rate the quality of the simplification by answering three questions:

- *Q1 — Is the simplification correct?* A correct simplification should be grammatically correct from the point of view of the norms of the Portuguese language and the conceptual legal norm. The possible values answers are: ‘Yes’, ‘No’, or ‘Partially’.
- *Q2 — Is the simplified sentence indeed simpler to understand?* The possible answers are: ‘Yes’, ‘No’, or ‘Same’ (when the output is identical to the input).
- *Q3 — What is the quality of the simplified sentence?* This is based on the previous questions, meaning how can the simplified sentence be evaluated. The values to be annotated are: ‘Good’, ‘Reasonable’, or ’Bad’.

The main guidelines from the Plain Language Book [8], which supported the evaluation of the sentences generated by the models, are: write short sentences with an average length of 15 to 20 words; choose commonly used words; use the direct order of the sentence; prefer direct action verbs; divide the text into topics; avoid gender markers; present the most important information at the beginning; and use appropriate punctuation.

Table 4 provides an overview of the human evaluation of the generated simplifications for each question (answer to *RQ2*). It shows the percentage of the instances that were classified by the expert into each of the possible categories.

GPT-3.5-Turbo is the winner, with the most robust performance across all three evaluative questions. The only output with ‘bad’ quality is in Table 5,

**Table 4.** Results of the qualitative evaluation.

| Models | Q1 | | | Q2 | | | Q3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | Partially | Yes | No | Same | Good | Bad | Reasonable |
| FT-PTT5 | .44 | .31 | .25 | .45 | .31 | .24 | .14 | .30 | .56 |
| FT-PTT5 + RL | .43 | .53 | .04 | .05 | .57 | .38 | .01 | .02 | .97 |
| GPT-3.5-Turbo | **.99** | — | .01 | **.98** | .02 | — | **.84** | .01 | .15 |
| GPT-4o | .89 | .04 | .07 | .81 | .18 | .01 | .57 | .19 | .24 |
| Flan-T5-Large | .17 | .83 | - | .01 | .03 | .96 | .01 | .99 | - |

**Table 5.** Sentences simplified by the winning model, GPT-3.5-Turbo

| **Original** ($O$) | **Simplified** | **Q1** | **Q2** | **Q3** |
|---|---|---|---|---|
| Este é um direito básico que deve ser garantido a todo cidadão. | Todo cidadão tem o direito básico de manter o mesmo plano de saúde coletivo após se aposentar. | No | No | Bad |
| Oficie-se como requer para que remeta a este juízo, no prazo de 15 dias, cópia do documento X. Anexar ao ofício cópia da inicial e desta decisão. | Envie um ofício pedindo uma cópia do documento X no prazo de 15 dias. Anexe ao ofício uma cópia da petição inicial e desta decisão. | Yes | Yes | Good |

along with an example of 'good' quality. The first does not correspond to the content of the reference sentence, adding information. In the second output, the order of the sentences is changed to allow a fluid reading, and the vocabulary is simplified using more frequent words.

GPT-4o also demonstrates competitive results, with high counts in 'Yes' responses for *Q1* and *Q2*. Both models present very low BLEU scores but high scores in BERTScore. BLEU uses the exact n-gram size, while BERTScore considers the context of the sentence. By applying TS, syntax and lexical changes can bring differences in terms of n-grams but keep the meaning. This shows the importance of using different metrics in the evaluation step.

Flan-T5-Large presented high scores for meaning preservation but low scores for simplicity. As we can see in Table 4, 95.6% of sentences were not modified from their original forms, agreeing with the results presented in the quantitative evaluation in Table 3. FT-PTT5 and FT-PTT5 + RL have similar results in quantitative evaluation, with the first one winning in simplicity, but in qualitative evaluation, FT-PTT5 + RL have less bad quality sentences than FT-PTT5 only.

The FT-PTT5 and FT-PTT5+RL provide better results compared to Flan-T5-Large, considering both qualitative and quantitative evaluations. Both PTT5-based models have problems in *Q1*, not preserving the legal meaning or with grammar problems. Since the fine-tuning does not have training examples from the domain, this can be the main reason for not being able to generalize for the domain (answer to *RQ1*).

DRESS [27] is the RL state-of-the-art with the best results until now, but the change of metrics used in this work did not yield good results in both evaluations.

Fine-tuning models have shown to be promising for TS, and RL is able to increase the performance too (answer to *RQ3*), as only 2% of the outputs were considered bad against 30% of FT-PTT5.

## 5 Conclusion

This work evaluated LLM for TS in legal texts written in Portuguese. We fine-tuned the pre-trained PTT5 model and developed an RL algorithm based on DRESS [27] but with PTT5 as a base for text-to-text generation. These models were compared with two GPT-3.5-Turbo, GPT4o, and FLAN-T5-Large.

There are many types of legal documents in Portuguese available, but without their ground truth simplifications, they cannot be readily used to train or evaluate TS. We relied on a dataset of legal case updates with their plain language explanations obtained from GPT-based models as training data. Although this data is in the domain of interest, it is not a TS dataset in the strict sense. In addition, our qualitative evaluation was made by only one specialist.

As future work, experiments using other models, such as BARD[8] and Maritaca,[9] can be performed. Fine-tuning the larger versions of LLMs can be very costly, but it has been shown to improve TS results. The use of RL had similar results to the fine-tuned model. However, the use of other metrics in the reward functions could improve performance, closing the gap in comparison to the GPT models. Other methods, such as performing TS according to readability levels, could be explored. The legal domain has different sub-fields, which makes it difficult to generalize a model for all sectors. An expression may have different meanings across sub-fields of the legal domain, or even across Brazilian estates. Thus, the question of whether a model can be robust to such variations is still unanswered.

**Acknowledgments**: This work has been partially funded by CNPq-Brazil, and Capes Finance Code 001. The authors are thankful to Edleno Silva de Moura for sharing the legal case updates created by JusBrasil and to INOVAJUS for sharing examples of simplification.

## References

1. de Almeida Guimarães, L.H.P.: A simplificação da linguagem jurídica como instrumento fundamental de acesso à justiça. Publicatio UEPG: Ciências Humanas, Linguistica, Letras e Artes **20**(2), 173–184 (2012)
2. Alva-Manchego, F., Scarton, C., Specia, L.: The (un) suitability of automatic evaluation metrics for text simplification. Computational Linguistics **47**(4), 861–889 (2021)
3. Alves, A., Miranda, P., Mello, R., Nascimento, A.: Automatic simplification of legal texts in portuguese using machine learning. In: Legal Knowledge and Information Systems, pp. 281–286. IOS Press (2023)

[8] https://gemini.google.com/

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

4. Candido Jr, A., Maziero, E.G., Specia, L., Gasperin, C., Pardo, T., Aluisio, S.: Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In: Workshop on Innovative Use of NLP for Building Educational Applications. pp. 34–42 (2009)
5. Carmo, D., Piau, M., Campiotti, I., Nogueira, R., Lotufo, R.: Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144 (2020)
6. Cemri, M., Çukur, T., Koç, A.: Unsupervised simplification of legal texts. arXiv preprint arXiv:2209.00557 (2022)
7. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024)
8. Cutts, M.: Oxford Guide to Plain English. Oxford University Press (2013)
9. Ferrés, D., Saggion, H., Guinovart, X.G.: An adaptable lexical simplification architecture for major ibero-romance languages. In: Workshop on Building Linguistically Generalizable NLP Systems. pp. 40–47 (2017)
10. Finatto, M.J.B., Macohin, A.: Pln no direito: Perspectivas e desafios com textos jurídicos e legais. In: Caseli, H.M., Nunes, M.G.V. (eds.) Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português, book chapter 26. BPLN, 2 edn. (2024)
11. Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., Kambhatla, N.: Text simplification for legal domain: Insights and challenges. In: Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., Preoțiuc-Pietro, D. (eds.) Natural Legal Language Processing Workshop 2022. pp. 296–304 (Dec 2022)
12. Hartmann, N.S., Aluísio, S.M.: Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. Linguamática **12**(2), 3–27 (2020)
13. Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., Shardlow, M.: Bless: Benchmarking large language models on sentence simplification. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
14. Leal, S.E., Duran, M.S., Aluísio, S.: A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In: International Conference on Computational Linguistics. pp. 401–413 (2018)
15. Martin, L., Fan, A., de la Clergerie, É., Bordes, A., Sagot, B.: MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Language Resources and Evaluation Conference. pp. 1651–1664 (Jun 2022)
16. North, K., Zampieri, M., Ranasinghe, T.: Alexsis-pt: A new resource for portuguese lexical simplification. arXiv preprint arXiv:2209.09034 (2022)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
18. Pena, T.M.G.: A simplificação da linguagem jurídica como fator de democratização do direito e inclusão social. Revista do Tribunal Regional do Trabalho da 24ª Região (5), 109–129 (2020)
19. Polo, F.M., Mendonça, G.C.F., Parreira, K.C.J., Gianvechio, L., Cordeiro, P., Ferreira, J.B., de Lima, L.M.P., do Amaral Maia, A.C., Vicente, R.: LegalNLP-Natural Language Processing methods for the Brazilian Legal Language. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 763–774 (2021)

20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
21. Sulem, E., Abend, O., Rappoport, A.: Semantic structural evaluation for text simplification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 685–696 (2018)
22. Tanprasert, T., Kauchak, D.: Flesch-kincaid is not a text simplification evaluation metric. In: Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021). pp. 1–14 (2021)
23. Viegas, C.F., Costa, B.C., Ishii, R.P.: Jurisbert: a new approach that converts a classification corpus into an sts one. In: International Conference on Computational Science and Its Applications. pp. 349–365. Springer (2023)
24. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brwac corpus: A new open resource for brazilian portuguese. In: International conference on language resources and evaluation (LREC) (2018)
25. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics **4**, 401–415 (2016)
26. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
27. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: Conference on Empirical Methods in Natural Language Processing (2017)