Evaluating the Simplification of Brazilian Legal Rulings in LLMs Using Readability Scores as a Target

Antônio Flávio Castro Torres de Paula[1], Celso Gonçalves Camilo[1],

[1]Institute of Informatics,
Federal University of Goiás (UFG),
Goiânia, Brazil, 74690-900
antonio.castro@discente.ufg.br, celso@inf.ufg.br

Abstract

Legal documents are often characterized by complex language, including jargon and technical terms, making them challenging for Natural Language Processing (NLP) applications. We apply the readability-controlled text modification task with an emphasis on legal texts simplification. Additionally, our work explores an evaluation based on the comparison of word complexity in the documents using Zipf scale, demonstrating the models' ability to simplify text according to the target readability scores, while also identifying a limit to this capability. Our results with Llama-3 and Sabiá-2 show that while the complexity score decreases with higher readability targets, there is a trade-off with reduced semantic similarity.

1 Introduction

Legal documents, in their majority, have a complex language, characterized by the use of jargon and words that are infrequently used in common vocabulary, as well as domain-specific technical terms Cemri et al. (2022a), Collantes et al. (2015). These features hinder access to information for the Brazilian population and pose a challenge that must be addressed by the Brazilian justice system.

Most text simplification approaches require a ground truth, typically provided by human experts Huang and Kochmar (2024). However, the availability of resources and techniques for Brazilian Portuguese is limited, and even more so when considering the specific task of text simplification in the legal domain.

The task of automatic text simplification is a natural language processing task whose objective is to modify the text to make it more understandable.

In this work, we evaluate the simplification of Brazilian legal rulings, using the method proposed by Farajidizaji et al. (2024), and propose an evaluation approach that considers complex words specific to the evaluated domain.

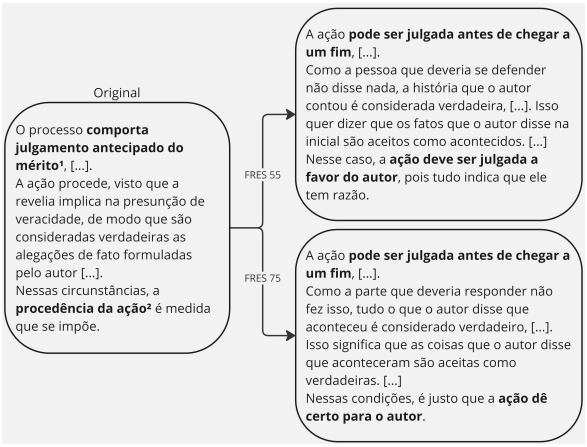

Figure 1: Example of a text simplification selected from the dataset. Highlighted excerpts: **1)** in English, "allows for an early judgment on the merits of the case", simplification could be translated to "can be decided before reaching a conclusion"; **2)** in English, "the claim's success is warranted", simplified to "the case should be decided in favor of the plaintiff" (FRES 55) and "the case goes well for the plaintiff" (FRES 75);

As far as we know, this is the first work evaluating LLMs for text simplification focused on legal documents in Brazilian Portuguese.

2 Related Work

In (Cemri et al., 2022b), the authors present USLT, an unsupervised method that identifies complex words through word frequency and applies measures to quantify complexity. These complex terms are replaced by candidates predicted by a masked language model and ranked based on various word characteristics. Finally, the solution applies sentence splitting, breaking down the original sentence into smaller ones. The results of the study show that the proposed method offers advantages over previous models developed for regular language. Moreover, it demonstrates that using a specific corpus and language models improves text simplification in legal documents.

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 117–125

In (Urchs et al., 2022), the authors describe a study on the automatic simplification of legal texts to make them more accessible to people with low literacy levels. The study focuses on South Korean legislation, comparing the original version with its official simplified version and exploring the differences between them in terms of sentence length, use of passive voice, and modal verbs, among other factors. The first model used is LSBert, specialized in lexical simplification. The second is a combination of ACCESS and MUSS, which paraphrases the original sentences. The authors conclude that while these models can quantitatively reduce complexity, they may struggle to retain all the important information from the original text.

Recently, Farajidizaji et al., 2024 presented a new task of readability-controlled text modification, along with new metrics. The work evaluates that LLM models like ChatGPT and Llama-2 are capable of paraphrasing texts using readability scores as a target, although the final readability remains correlated with the original text. This work applies the methodology proposed by Farajidizaji et al., 2024, but focuses on higher FRES scores, aiming to evaluate only text simplification given a target score.

3 Methodology

3.1 Readability-controlled text modification task

The readability-controlled text modification task, presented in (Farajidizaji et al., 2024), defines that for each text, 8 variations are generated based on a target readability score. The function chosen to calculate the target score was the Flesch Reading Ease (FRES) index. Each range of the FRES index results in a text variation.

In this work, our goal is to evaluate the simplification capability, i.e., higher scores of FRES. Therefore, we will generate 5 variations of the original text, considering the following target readability scores: $r1$ = 55, $r2$ = 65, $r3$ = 75, $r4$ = 85, and $r5$ = 95. Each value of r represents half of the FRES score range.

3.2 Flesch reading Ease Portuguese Adaptation

The FRES score was originally developed for English and indicates that the higher the score, the easier the text is to read. The score takes into account the number of words, the number of sentences, and the number of syllables.

US	Brazil
5th grade	*5º ano do Ensino Fundamental I*
6th grade	*6º ano do Ensino Fundamental II*
7th grade	*7º ano do Ensino Fundamental II*
8-9th grade	*8º ano ao 9º ano do Ensino Fundamental II*
10-12th grade	*1º ao 3º ano do Ensino Médio*

Table 1: Proposed correspondence between education levels in the US and Brazil for the interpretability of the FRES index.

In this work, we will use an adaptation of the Flesch score for Brazilian Portuguese, presented by (Scarton and Aluísio, 2010).

The formula of the proposed adaptation is indicated by Equation 1.

$$248.835 - (1.015 ASL) - (84.6 * ASW) \quad (1)$$

where ASL = average sentence length (the number of words divided by the number of sentences) and ASW = average number of syllables per word (the number of syllables divided by the number of words).

Originally, each FRES score range can be interpreted as an education level and is accompanied by a description that details the meaning of each range. However, these levels and descriptions are applicable to the English language and the education system in the United States. The education level and description will be used in the experiments as text input, which is why we also adapted this information.

Similar to the United States, Brazil also has a 12-year educational system, and the age at each educational level is the same. For this reason, we adapted the corresponding education level for each FRES score range. The level and description will be used as input in the experiments. Table 1 shows the proposed correspondences.

3.3 Evaluation

Originally in (Farajidizaji et al., 2024), three levels of evaluations are performed.

First, at the **individual level**, for each example in the dataset, the model is evaluated on three aspects

concerning the expected readability score: ranking, regression, and classification.

In ranking, Spearman's correlation is calculated to measure whether the ranking of the generated rewrites is maintained relative to the target scores. In regression, the root mean square error (RMSE) between the target score and the actual score of the generated text is calculated. The formula is given by Equation 2.

$$rmse = \left[\frac{1}{5}\sum_{r \in R}(F(y_{(r)}) - r)^2\right]^{1/2} \quad (2)$$

where F represent FRES funcion from Equation 1, $y(r)$ is the text generated with target score r. R is the list of target scores to evaluation. Finally, in classification, the accuracy (Equation 3) of the calculated score is checked within the expected FRES score complexity range, given the target.

$$acc = \frac{1}{5}\sum_{r \in R}\mathbf{1}_{A(r)}(F(y_{(r)})) \quad (3)$$

For the three aspects at the individual level, the mean is reported across the dataset.

The second level of evaluation is called **Population-scale readability control**, where a decorrelation between the generated text and its source is expected. However, since this work aims to evaluate text simplification, a dependency on the source text is expected. This level will not be evaluated.

In the third and final level, **paraphrase metrics** are evaluated. The word error rate (WER [1]) is calculated to measure the lexical divergence between the original and generated texts. Another metric evaluated is the BERTScore [2], which assesses the semantic similarity between the generated and original texts, using cosine similarity between the embeddings of each text.

Additionally, we will evaluate the number of complex words in the original text and in the rewritten texts for each target score.

3.4 Using Complex Word Identification as Score

In (Cemri et al., 2022b), part of the proposed lexical simplification system is the identification of complex words (CWI), which is performed automatically without requiring a predetermined list of labeled complex words. The work uses word frequency across two different corpora to determine the list of complex words.

According to Zipf's law, words with lower frequency tend to be longer and more complex than more frequent and shorter words (Quijada and Medero, 2016). Word frequency was also highlighted as the most effective way to determine word complexity in the study conducted in (Paetzold and Specia, 2016).

To allow comparison between two corpora of different sizes, word frequency was calculated based on the Zipf scale (van Heuven et al., 2014). The Zipf scale is logarithmic, ranging from very low (Zipf value 1) to very high (Zipf value 7) frequency words, and can be represented by Equation 4.

$$Zipf = log\left(\left(\frac{wf * 1000000}{corpus\_size}\right) + 3\right) \quad (4)$$

where wf is the word frequency.

For the complex word score, we are interested in words that are more frequent in the domain corpus and less frequent in the Portuguese corpus. For the Brazilian Portuguese corpus, we considered BrWaC Wagner Filho et al. (2018). BrWaC is a large Brazilian corpus created from web pages, containing 2.7 billion tokens.

Based on the normalized frequency (Zipf value) of the two corpora, we can propose a simple metric that creates a complexity ranking of the words in the domain corpus. The score for each word is given by Equation 5.

$$cws = (1 + zipf\_domain) * r\_zipf\_brwac) \quad (5)$$

where $zipf\_domain$ is the word's frequency in Zipf value in the domain corpus, and r_zipf_brwac is the word's ranking index in BrWaC.

As we are only interested in the rarest words in the domain corpus, we consider as complex words only those with a frequency higher than the average frequency in the domain corpus.

The complex score evaluated in the results is the sum of the scores of the complex words identified in the evaluated text. This metrics allow us evaluate the generated text automatically, without data annotation.

[1] Implementation available on https://github.com/jitsi/jiwer.

[2] Implementation available on https://huggingface.co/spaces/evaluate-metric/bertscore.

# examples	# words	# sentences	# paragraphs
10000	216.2 ± 18.6	9.9 ± 4.2	5.8 ± 2.5

Table 2: Statistics of the legal text dataset used in the experiments.

$$complexity\_score = \sum_{x \in CWL} (cws) \quad (6)$$

where x is each word in the text that belongs to the list of complex words in the domain (CWL), and cws is the complexity score of each word.

4 Experiments

4.1 Data

In order to conduct experiments in the context of legal sentence simplification, we prepared a specific dataset for evaluation.

The data used in the experiments are a subset of sentence documents downloaded from the São Paulo State Court website. A total of 80,000 public court sentence documents were downloaded from the period of 2021-06-01 to 2024-06-30, from 1,856 different judges. For this work, only judges with 30 or more sentences were considered, resulting in 195 judges.

The documents were pre-processed, segmenting and extracting the reasoning section of the legal sentences. The final dataset consists of 10,000 documents, with each document being either the entire reasoning or a part of it. Table 2 describe the stats.

4.2 Zero-shot

For the evaluations, we used the Llama-3 (Dubey et al., 2024) (`llama3-8b-8192`) and Sabiá-2 (Almeida et al., 2024) (`sabia-2-small`) models. Inferences for both models were performed via API. The Llama-3 model was accessed through the Groq platform [3] (with free credits available until the publication of this work), and the Sabiá-2 model was accessed through the Maritaca AI platform [4], with our own credits. The cost per million tokens is currently `R$` 1.00 for input tokens and `R$` 3.00 for output tokens.

The input prompts were based on the education level and description of the FRES score interpretation adapted for Portuguese, as shown in Table 1. The prompts used are described in Table 3. Appendix A describes original prompts

In addition to the text to be evaluated and the simplification instructions for each FRES score range, a supplementary prompt was added to prevent model hallucination, different language output and unnecessary structure formatting. The supplementary prompt includes:

- **Do not add facts that do not exist in the original text**: In some cases, the model generated facts that could be inferred from the original text but were not explicitly mentioned.
- **Generate only the rewritten text and in Portuguese**: In some cases, the Llama-3 model generated part of the output in English or described what had been done. For example: "Here is the rewritten document..."
- **Do not segment or separate the text**: Since these are parts of a sentence document, in some cases the model generated headers that structured the output into sections commonly found in such documents, such as Reasoning and Decision.

To demonstrate, without considering the complementary prompt, the Llama-3 model generated the following excerpts as part of the text simplification output for some documents in the dataset:

- "*Espero que isso ajude!*" (I hope this helps!);
- "*Espero que isso seja fácil de entender!*" (I hope this is easy to understand!);
- "*Lembre-se de que o texto original é um trecho de um julgamento e foi escrito em um estilo jurídico, então foi necessário adaptá-lo para que fosse mais fácil de entender para um estudante do 7º ano do Ensino Fundamental II.*" (Keep in mind that the original text is an excerpt from a legal ruling and was written in a legal style, so it had to be adapted to make it easier to understand for a 7th-grade student.);

[3] https://groq.com.

[4] API documentation `https://docs.maritaca.ai/pt/modelos`.

Target score	Prompt
55	*Reescreva este documento para o nível escolar do 1º ao 3º ano do Ensino Médio (Brasil). Deve ser relativamente difícil de ler.*
65	*Reescreva este documento para o nível escolar do 8º ano ao 9º ano do Ensino Fundamental II (Brasil). Deve ser em português claro e facilmente compreendido por estudantes de 13 a 15 anos.*
75	*Reescreva este documento para o nível escolar do 7º ano do Ensino Fundamental II (Brasil). Deve ser relativamente fácil de ler.*
85	*Reescreva este documento para o nível escolar do 6º ano do Ensino Fundamental II (Brasil). Deve ser fácil de ler e em portugûes coloquial, adequado para o público em geral.*
95	*Reescreva este documento para o nível escolar do 5º ano do Ensino Fundamental I (Brasil). Deve ser muito fácil de ler e de fácil entendimento para estudantes com média de 11 anos de idade.*

Table 3: Prompts considering each target score, in Brazilian Portuguese, translated from English and aligned with the educational level mentioned in Table 1. Appendix A describes original prompts in English.

Model	p(↑)	rmse(↓)	acc(↑)
Original data	0.0	$44.35_{\pm 14.63}$	$2.24_{\pm 6.52}$
Llama-3	$61.05_{\pm 37.38}$	$24.46_{\pm 10.13}$	$10.89_{\pm 15.32}$
Sabiá-2	$22.76_{\pm 49.64}$	$20.41_{\pm 7.29}$	$16.51_{\pm 15.14}$

Table 4: Mean of the individual-level metrics: p value (%) is the Spearman's rank correlation coefficient, rmse measures regression ability, and accuracy of the generated scores classification.

Target	WER	BERTScore (F1)
55	73,6 ± 34.0	78,1 ± 6.1
65	84,1 ± 39.8	74,4 ± 5.1
75	82,9 ± 11.1	74,4 ± 5.1
85	87,4 ± 10.2	72,1 ± 4.8
95	87,2 ± 33.2	72.0 ± 4.7

Table 5: Lexical divergence metrics (WER) and semantic similarity (BERTScore) between the original and generated texts, to model Llama-3. The mean of all examples with one standard deviation.

Target	WER	BERTScore (F1)
55	90,9 ± 25.0	72,8 ± 7.8
65	102,8 ± 19.2	69,9 ± 5.5
75	98,9 ± 18.2	70,3 ± 5.6
85	102.0 ± 17.6	68,3 ± 5.1
95	101,8 ± 15.4	67,8 ± 4.5

Table 6: Lexical divergence metrics (WER) and semantic similarity (BERTScore) between the original and generated texts, to model Sabiá-2. The mean of all examples with one standard deviation.

5 Results and Discussion

Table 4 presents the evaluation metrics at the individual level. The item described as "Original data" refers to the original text, which was also evaluated in some of the metrics based on the target scores.

When analyzing the correlation coefficient applied to the generated ranking, we observe that in the Llama-3 model, despite the zero-shot implementation not achieving the target scores exactly, it has a moderate correlation of 61.05% with the expected scores. On the other hand, the correlation of the Sabiá-2 model is weak (20.76 %). It is also interesting to note that, despite the Sabiá-2 model being trained in Brazilian Portuguese, Llama-3 achieves 38 points higher based on the target score ranking. On the other hand, when evaluating the mean squared error of the proposed models, we find that both models achieve a lower error than the original data. However, a high error is expected in relation to the original data, as it is repeated when measured against the expected scores.

Tables 5 and 6 present the paraphrase metrics for Llama-3 and Sabiá-2, respectively. It is possible to see that in both models, the word error rate increases with higher target readability scores, in-

Target	Sabiá-2 (%)	Llama-3 (%)
55	0,336 (71,1%)	0,571 (50,9%)
65	0,301 (74,2%)	0,457 (60,8%)
75	0,288 (75,3%)	0,401 (65,5%)
85	0,242 (79,2%)	0,355 (69,5%)
95	0,238 (79,5%)	0,342 (70,6%)

Table 7: The mean of complexity score achieved at each readability target score. The presented percentage represents the reduction compared to the original text score. For example, at the target score of 85, the Sabiá-2 model shows a 79.2% reduction in the complexity score.

dicating that the generated texts have a high degree of lexical divergence. This behavior is expected when simplifying a text. It is also observed that the Sabiá-2 model has an average advantage of 16.24% over Llama-3, considering all scores. On the other hand, it is also expected that the generated text remains semantically similar to the original text. In this case, for all target readability scores, the Llama-3 model outperformed Sabiá-2, achieving a mean F1 score of 74.2%, while the mean F1 score for Sabiá-2 was 69.82%.

Finally, we have the evaluation of complex words presented in Table 7. Both models show a reduction in the complexity of the words used, considering the original complexity of 1.1635. The Sabiá-2 model has a significant advantage, with an average reduction of 75.84% in the complexity score, while the reduction is 63.44% in Llama-3. It can also be observed that the difference between the target scores of 85 and 95 shows no significant reduction in complexity, which can be interpreted as a simplification limit reached by the models.

6 Conclusions

This work applies the task of readability-controlled text modification, focusing on the simplification of legal texts. We explore an approach based on complex word identification to evaluate the a text based on word complexity, indicating that the evaluated models have simplification capabilities and that there is a limit to this capacity, considering the proposed target scores.

In both evaluated models, Llama-3 and Sabiá-2, we observed that the complexity score decreases with higher readability scores, but with a reduction in the semantic similarity metric, highlighting the challenge of balancing simplification while preserving the main points of the original text.

7 Ethics Statement

This work does not raise any ethical concerns.

8 Limitations

We believe that the score based on complex word identification can be improved, as there is improvements for enhancement in the preprocessing of domain-specific texts. Additionally, the creation of a unified metric that considers various aspects of the generated text could simplify the evaluation of results, instead of assessing each metric in isolation.

Finally, adapting the steps into a framework that can be applied to domains beyond justice and legal texts.

9 Lay Summary

This research focuses on evaluating the simplification of Brazilian legal rulings using large language models (LLMs). Legal documents are often complex, making it difficult for the general public to understand their content. The study examines whether modern language models can simplify legal texts while preserving their original meaning, aiming to improve accessibility to legal information. The main question addressed by the study is whether large language models can automatically simplify Brazilian legal texts. The main question addressed by the study is whether large language models can automatically simplify Brazilian legal texts. Most simplification methods are validated by comparing them to human-made simplifications. However, such resources are limited for Brazilian Portuguese, particularly in the legal domain, making this task both challenging and significant for advancing language technologies in this field.

The findings show that the models are capable of simplifying legal sentences, but there is a trade-off. While both models reduce the complexity of the language used, they also decrease the semantic similarity with the original text, highlighting the challenge of simplifying text while maintaining its core meaning.

This work can benefit Brazilian society by making legal documents more accessible, potentially improving public understanding and compliance with legal decisions. Further research and advancements in this field are needed to enhance the balance between simplification and the preservation of original meaning.

References

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-2: A new generation of portuguese large language models. *Preprint*, arXiv:2403.09887.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022a. Unsupervised simplification of legal texts. *arXiv preprint*.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022b. Unsupervised simplification of legal texts. *arXiv preprint*.

M. Collantes, Maureen Hipe, Juan Lorenzo Sorilla, Laurenz Tolentino, and Briane Paul V. Samson. 2015. Simpatico: A text simplification system for senate and house bills.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres,

Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. *Preprint*, arXiv:2309.12551.

Yichen Huang and Ekaterina Kochmar. 2024. Referee: A reference-free model-based metric for text simplification. *Preprint*, arXiv:2403.17640.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Maury Quijada and Julie Medero. 2016. HMC at SemEval-2016 task 11: Identifying complex words using depth-limited decision trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1034–1037, San Diego, California. Association for Computational Linguistics.

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Stefanie Urchs, Akshaya Muralidharan, and Florian Matthes. 2022. How to simplify law automatically? a study on south korean legislation and its simplified version. In *ICAART: PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON AGENTS AND ARTIFICIAL INTELLIGENCE - VOL 3*, ICAART, pages 697–704. 14th International Conference on Agents and Artificial Intelligence (ICAART), ELECTR NETWORK, FEB 03-05, 2022.

Walter van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: a new and improved word frequency database for british english. *Quarterly journal of experimental psychology (2006)*, 67.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Appendix: Original prompts

This appendix provides the original list of prompts in English for each target score used. The prompts below were translated into Brazilian Portuguese, and the school grade levels were adapted to match the Brazilian education system.

A.1 FRES Target 55

Paraphrase this document for 10th-12th grade school level (US). It should be fairly difficult to read.

A.2 FRES Target 65

Paraphrase this document for 8th/9th grade school level (US). It should be plain English and easily understood by 13- to 15-year-old students.

A.3 FRES Target 75

Paraphrase this document for 7th grade school level (US). It should be fairly easy to read.

A.4 FRES Target 85

Paraphrase this document for 6th grade school level (US). It should be easy to read and conversational English for consumers.

A.5 FRES Target 95

Paraphrase this document for 5th grade school level (US). It should be very easy to read and easily understood by an average 11-year old student.