Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights

Anonymous ACL submission

Abstract

Sentence simplification (SS) focuses on adapting sentences to enhance their readability and accessibility. While large language models (LLMs) match task-specific baselines in English SS, their performance in Portuguese remains underexplored. This paper presents a comprehensive performance comparison of 26 state-of-the-art LLMs in Portuguese SS, alongside two simplification models trained explicitly for this task and language. They are evaluated under a one-shot setting across scientific, news, and government datasets. We benchmark the models with our newly introduced Gov-Lang-BR corpus (1,703 complex-simple sentence pairs from Brazilian government agencies) and two established datasets: PorSimplesSent and Museum-PT. Our investigation takes advantage of both automatic metrics and large-scale linguistic analysis to examine the transformations achieved by the LLMs. Furthermore, a qualitative assessment of selected generated outputs provides deeper insights into simplification quality. Our findings reveal that while open-source LLMs have achieved impressive results, closed-source LLMs continue outperforming them in Portuguese SS.

1 Introduction

Sentence simplification aims to make a sentence more straightforward to read and understand without changing its key points (Alva-Manchego et al., 2020). This task offers numerous critical social applications, benefiting a wide range of individuals (Stajner, 2021). For instance, it plays a key role in enhancing accessibility for people with reading difficulties, ensuring that texts are more approachable for those who struggle with complex structures (Aluísio and Gasperin, 2010). It supports individuals with cognitive impairments, such as aphasia (Carroll et al., 1998) and dyslexia (Rello et al., 2013; MADJIDI and CRICK, 2024). Moreover, it proves valuable for non-native speakers, helping them navigate unfamiliar vocabulary and grammatical forms (Paetzold and Specia, 2016).

In addition, text simplification has emerged as an increasingly helpful NLP application to bridge communication gaps in specialized fields, such as medicine and law, where the lexicon is often dominated by technical jargon and complex constructions (Luo et al., 2022; Garimella et al., 2022). Notably, in Brazil's public administration sector, the government is required to adhere to legal principles when carrying out any administrative act, including the principle of transparency.[1] To ensure public acts are as clear and accessible as possible, it is essential to use plain language in communication with all those affected by the actions of public authorities.[2] The wide range of services provided to citizens, such as legal and tax departments, usually hold specific terms. This often forces people to hire third-party services to address simple issues they could resolve independently.

LLMs have shown remarkable performance across a wide range of NLP tasks without requiring task-specific training, leading to the belief that they have the potential to solve virtually any task (Brown et al., 2020; Qin et al., 2024; Yang et al., 2024). This prompts the creation of benchmarks in specific domains and tasks to evaluate the capabilities of LLMs (Wang et al., 2018). Although there are benchmarks in languages other than English (Fenogenova et al., 2024; Thellmann et al., 2024; Liu et al., 2024a), those available in Portuguese are mainly limited to classification tasks. (Pires et al., 2023; Garcia et al., 2024).

Thus, the performance of recent LLMs in the

[1] https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, https://www.camara.leg.br/noticias/1023177-camara-aprova-uso-de-linguagem-simples-na-comunicacao-de-orgaos-publicos/

[2] https://www.gov.br/gestao/pt-br/assuntos/inovacao-governamental/cinco/cinforme/edicao_1-2023/linguagem-simples

task of sentence simplification in Portuguese remains largely unexplored. While some studies have evaluated specific LLMs for this task (Kim, 2022; Liu et al., 2024b; Alves et al., 2023; Scalercio et al., 2024; Shardlow et al., 2024), there is no comprehensive, large-scale analysis that assesses the potential of different LLMs in Portuguese SS.

In this paper, we study the capabilities of LLMs and specific simplification models on three Portuguese datasets: PorSimplesSent (Leal et al., 2018), Museum-PT (Scalercio et al., 2024) and Government Language-BR, our curated dataset containing complex-simple pairs from a diverse set of Brazilian public agencies. The datasets cover a wide variety of domains (science, news, and government) and feature diverse simplification operations.

We adopt in-context learning (ICL) in a one-shot prompting scenario to assess LLM capabilities. We evaluate 26 widely used generative models, including both open and closed-weight models across several dimensions. We employ automatic evaluation metrics commonly used in the SS literature. We also quantify and compare the linguistic transformations the LLMs perform during simplification. We investigate which types of one-shot example produce the best and worst simplifications. Finally, we conduct a qualitative analysis to validate our findings and to gain deeper insights into the quality of the generated simplifications. As expected, the closed-weight models usually outperform their open-weight contenders. However, a family of open-weight LLM has achieved impressive results, even surpassing some closed-weight LLMs. The results from the open-weight models are especially significant because they are quantized to make it possible to run their inference on a single 24GB GPU. Our findings show that Portuguese sentence simplification can be effectively achieved with open-weight LLMs, even in a low-resource regime.

The contributions of this paper are:

1. An evaluation benchmark on the Portuguese sentence simplification task using 26 LLMs in a one-shot scenario.
2. An evaluation framework including automatic and linguistic in-depth simplification metrics.
3. A qualitative analysis of the results, with manual annotation of simplification operations.
4. A newly compiled sentence simplification dataset with $1,703$ complex-simple sentence pairs, the Government Language-BR dataset. We publicly release code, datasets, and generated outputs as a resource for SS research.

2 Related Work

2.1 Sentence Simplification

Most research in sentence simplification usually follows a generative or an edit-based supervised strategy. The first case includes sequence-to-sequence models (Nisioi et al., 2017) using transformer (Vaswani et al., 2017a) architectures and reinforcement learning (Zhang and Lapata, 2017), leveraging external paraphrase datasets (Zhao et al., 2018), and integration of syntactic rules (Maddela et al., 2021). In contrast, edit-based supervised models use parallel complex-simple sentence pairs. Alva-Manchego et al. (2017) learns which operations should be performed to simplify a sentence, and Omelianchuk et al. (2021) predicts token-level operations in a non-autoregressive manner.

Controllable sentence simplification involves fine-grained techniques that guide generation, conditioning simplified sentences on both the input and desired attributes(Nishihara et al., 2019). These attributes include low-level linguistic features, such as dependency tree depth, word rank, number of characters, Levenshtein similarity, and high-level features, like the desired target level of readability (Martin et al., 2020; Ristad and Yianilos, 1996). Target-level simplification refers to the process of generating output tailored to specific readability levels or reader profiles, overcoming the need for specific linguistic knowledge(Kew and Ebling, 2022; Chi et al., 2023; Agrawal et al., 2021; Qiu and Zhang, 2024).

2.2 Simplification in Portuguese

Previous works on sentence simplification in Portuguese that uses machine learning often rely on parallel corpora. Specia (2010) proposed a Statistical Machine Translation (SMT) framework to learn how to convert complex sentences into simpler ones, using a parallel corpus of original and simplified texts. Hartmann and Aluísio (2020) developed a pipeline specifically for the lexical simplification of elementary school text in Brazilian Portuguese. Given the limited resources, zero-shot, few-shot, and unsupervised methods have emerged as promising strategies for simplifying Portuguese texts.

In this context, Martin et al. (2022) introduced a neural model[3] trained on a large corpus of mined Portuguese paraphrases, using control tokens. Scalercio et al., 2024 also trained a neural model using mined paraphrases but adopted a different training procedure, learning a style representation using context and linguistic features.

2.3 LLM-based Simplification

Recent work on text simplification has taken advantage of the new age of foundational LLMs through fine-tuning and prompt engineering to produce simplifications (Cripwell et al., 2023; Farajidizaji et al., 2024). Given LLMs' strong performance, sentences can now be simplified using an off-the-shelf model without domain-specific training. Some specific simplification models compared their simplification capabilities with LLMs to benchmark their performance (Sun et al., 2023; Chi et al., 2023; Ryan et al., 2023; Scalercio et al., 2024).

Feng et al. (2023) analyzed the zero-/few-shot learning ability of LLMs to simplify sentences in several languages, including Portuguese. However, their results only reached a limited number of LLMs and evaluation metrics. Kew et al., 2023 is the most extensive work analyzing LLM on sentence simplification, benchmarking 44 LLMs on English Sentence Simplification. Our work also follows the tendency to benchmark LLMs on sentence simplification. Still, our study focuses on the Portuguese language. It provides an extensive linguistic analysis of the simplification process performed by the LLM, along with an investigation of the best one-shot examples.

3 Experimental Setting

3.1 Datasets

We assess LLMs on Portuguese SS using three datasets spanning different domains and styles.

PorSimplesSent (Leal et al., 2018) was built from the parallel corpus PorSimples (Aluísio and Gasperin, 2010). It features multiple versions, distinguishing whether the complex texts were split during simplification. To allow comparison with previous work, we use the same test set as Scalercio et al., 2024 where the complex sentences remain unsplit. It comprises a total of 606 sentences for the test set.

Museum-PT is a document simplification dataset proposed in Finatto and Tcacenco (2021) with its sentences aligned in Scalercio et al. (2024). The set comprises written texts accompanying experiments and objects from science and technology museums, aimed at a general audience. For benchmarking the models on SS, we selected all aligned sentences, totaling 476 complex-simple pairs.

Both PorSimplesSent and Museum-PT datasets originated from simplifications carried out by linguists, aiming to reduce or eliminate complexity by applying Plain Language[4] techniques and adhering to principles of Textual and Terminological Accessibility (Saggion and Hirst, 2017).

Moreover, we propose and evaluate LLMs on **Brazilian Government Language (Gov-Lang-BR)**, a new dataset containing 1,703 complex-simple pairs. To construct this dataset, we gathered publicly available pairs of texts and their simplified versions from various Brazilian government agency websites, encompassing federal, state, and municipal levels. These sentences are closely aligned with the goals of the respective agency. For instance, some are collected from a municipal planning agency focused on making financial and planning terminology more accessible to the general public. The simplifications were refined with the expertise of domain specialists and plain-language experts . The distribution of the data according to its source together with further statistics are in the Appendix A.

3.2 Large Language Models

We investigate a total of 26 LLMs with different sizes, architectures, and training objectives, including open-weight and closed-weight models. Open-weight models refer to those whose trained weights are accessible, enabling users to host them independently. The open-weight models we consider range from 3 to 72 billion parameters, all based on the transformer architecture (Vaswani et al., 2017b). All have undergone a self-supervised pre-training stage. Some models leverage instruction-tuning, i.e., fine-tuning a pre-trained base model on labeled instruction-response pairs from diverse tasks.

In comparison, closed-weight models refer to those whose weights are kept private and can be queried only through APIs. We included as many as possible the models that perform best according to the open Portuguese LLM leaderboard[5]. The

[3]https://github.com/facebookresearch/muss.git

[4]https://www.iso.org/obp/ui#iso:std:iso:24495:-1:ed-1:v1:en,https://snow.idrc.ocadu.ca/accessible-media-and-documents/text-simplification-guidlines/

[5]https://huggingface.co/spaces/eduagarcia/ope

open-weight models include variants of the Qwen family (Bai et al., 2023a), OLMo (Groeneveld et al., 2024a), LLaMA models (Touvron et al., 2023b), Phi-3 models (Abdin et al., 2024a), and a model from Google Gemma family (Team et al., 2024b). The closed-weight models are developed by OpenAI[6], Cohere[7], and Maritaca AI[8], the first due to the popularity of GPT-based models, the second due to their multilingual training, and the third because it provided the first PT-BR language-based LLM, the Sabiá model (Pires et al., 2023). Details on each family of models and the characteristics of the open-weight LLMs are in the Appendix B.

3.3 Baselines

Our evaluation uses two recent, robust baselines trained for Portuguese SS.

MUSS-Unsupervised (Martin et al., 2022): This is an unsupervised multilingual simplification method that fine-tunes BART (Lewis et al., 2020), leveraging paraphrases and control tokens from ACCESS (Martin et al., 2020) during training.

Enhancing-PT-SS (Scalercio et al., 2024): This is an unsupervised Portuguese-only simplification method that employs a T5 (Raffel et al., 2020) Seq2Seq model enhanced with an extra T5 encoder. The extra encoder learns a style representation that aids the decoder during generation. This model is also fine-tuned in mined paraphrase pairs.

3.4 Inference details

We run inference on local GPUs using the LM Studio[9] framework for open-weight models. Unless otherwise specified, we load the models with 4-bit quantization (Q4_K_M method), which allows us to run inference efficiently on a single RTX4090 24GB GPU. We use the APIs provided by Cohere, OpenAI and Maritaca AI for closed-weight models. Following previous work (Kew et al., 2023), we use Nucleus Sampling with a probability of 0.9, a temperature of 1.0, and a context size of 1024 tokens. For our one-shot exemplars, we selected four different complex-simple pairs, each performing a different type of simplification: syntactic simplifications, changes in phrase order, anaphora resolution, and eliminating redundant information. We perform inferences using each one individually. We also perform each inference run three times to account for the probabilities. Thus, we generate twelve simplifications for each input sentence and aggregate the results for each metric. We adopted a single prompt throughout the experiments. More details about the demonstration examples and prompts are in Appendix E.

3.5 Automatic and Linguistic Metrics

Our evaluation comprises automatic metrics widely used in text simplification task (Sheang and Saggion, 2021; Martin et al., 2022), which are also readily applicable to Portuguese. We measure simplicity using SARI (Xu et al., 2016), meaning preservation using BERTScore (Zhang* et al., 2020) and BLEU (Papineni et al., 2002). These metrics are computed using the EASSE package (Alva-Manchego et al., 2019)[10]. We also report the percentage (%) of unchanged outputs (i.e., exact copies), following Agrawal and Carpuat (2023).

To gain insights into the simplification process performed by LLMs, we devised a morphosyntactic analysis, comparing model-generated to experts-produced sentences (Section 4.2). The 18 linguistic metrics used in this analysis were developed based on linguistic hypotheses about complexity. These hypotheses are derived from descriptive corpus-based studies (Charles, 2013) and psycholinguistic research on language processing complexity (Juola, 1998; Gibson, 1998; Corrêa et al., 2019), adapted to align with the available tagset for automatic morphosyntactic analysis of texts.

From the 18 metrics, we take a closer look at the four that exhibited the most variation when comparing the original sentences with their respective expert-produced references. This analysis focuses exclusively on the Museum-PT and PorSimplesSent datasets, as their references are certainly linguist-produced texts. The four selected metrics are: (1) Lemma/Token Ratio (LTR) that measures lexical diversity; (2) Ratio of passive to active voice verbs (P/A) to measure more direct constructions; (3) Proportion of adverbial clauses preceding the main clause (AdvLeft), capturing sentence structure tendencies; and (4) Ratio of fully developed to reduced relative clauses (D/R), reflecting syntactic simplifications. Appendix C details the 18 metrics and their values across the datasets.

n_pt_llm_leaderboard

[6]https://openai.com/

[7]https://cohere.com/

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

[9]https://lmstudio.ai/

[10]https://github.com/feralvam/easse

		PorSimplesSent		Museum-PT		Gov-Lang-BR	
		SARI	BScore	SARI	BScore	SARI	BScore
Baseline	MUSS	38.30	.8976	39.31	.8534	28.00	**.8221**
	Enhanc-PT-SS	**39.64**	**.9024**	**41.62**	**.8550**	**31.84**	.8129
Open-weight LLM	aya-23-8b	33.87	.8534	43.61	.8269	41.61	.7799
	gemma2-27b-it	30.84	.8352	41.12	.8130	41.13	.7808
	llama-3.1-8b-instruct@q4_k_m	30.17	.8289	40.28	.8101	41.27	.7793
	mistral-7b-instruct-v0.3	33.08	.8465	41.32	.8154	40.07	.7892
	qwen2-7b-instruct@q4_k_m	35.75	.8661	44.54	.8319	41.85	.7969
	qwen2-72b-instruct	34.69	.8576	43.94	.8296	41.19	.7818
	qwen2.5-7b-instruct@q8_0	36.30	.8694	44.51	.8354	43.54	.7998
	qwen2.5-7b-instruct@q4_k_m	**36.61**	**.8701**	44.20	.8347	43.50	.7980
	qwen2.5-14b-instruct	33.96	.8534	43.42	.8183	42.86	.7844
	qwen2.5-32b-instruct	35.81	.8651	**45.74**	**.8369**	**44.05**	**.8021**
	deepseek-r1-distill-qwen-7b	34.95	.8523	39.11	.8120	38.63	.7783
	deepseek-r1-distill-qwen-32b	**36.46**	.8689	44.69	.8352	**43.91**	.8019
Closed-weight LLM	Command-r-08-2024	32.60	.8329	42.79	.8110	44.35	.7924
	GPT3.5-Turbo	38.18	.8805	47.23	.8468	-	-
	GPT4o-mini	**39.75**	**.8838**	**48.92**	**.8508**	**45.14**	.8155
	o1-mini	39.26	.8472	47.26	.8252	**45.24**	.7808
	Sabia-2-small	38.16	.8732	44.44	.8353	44.29	**.8172**
	Sabia-3	35.12	.8546	44.72	.8270	42.56	.7889

Table 1: Simplification (SARI) and Meaning Preservation (BERTScore) metrics for the best-performing LLMs and baselines. The best SARI and BERTScore results for Baselines, and open- and closed-weight LLMs are in bold.

4 Quantitative Results

4.1 Automatic Evaluation

We evaluate all LLMs and baselines automatically on the three datasets. Table 1 reports the SARI and BERTScore results of the best-performing LLMs and baselines. The complete results for the 26 LLMs are in Appendix D. We observe that the closed-weight gpt4o-mini achieved the best results overall. However, the qwen2.5-7b-instruct, qwen2.5-32b-instruct and Sabia-2-small models also performed well across all datasets, staying close to GPT models. Scaling the size of the LLM did not improve performance for all models. For example, qwen2.5-14b-instruct and qwen2-72b-instruct models were outperformed by smaller versions in all datasets. Sabia-2-small also outperformed Sabia-3 in two of the three datasets. Quantization using 4 bits achieved similar or better results than with 8 bits. Notably, the reasoning model o1-mini achieved decent simplification but lost significant meaning. Designed to break down complex problems step by step, they often introduce excessive explanations and additional context instead of condensing information (Cuadron et al., 2025).

Many top-ranked models performed poorly, likely because the leaderboard evaluates only classification tasks, excluding generation. Given the small test sets, we used the Paired Bootstrap Resampling test (Koehn, 2004) to assess the statistical significance of the SARI scores. More than one bolded LLM in the same column indicates no statistical superiority among them, with a 95% significance level.

In **PorSimplesSent**, OpenAI's GPT4o-mini outperforms all other tested LLMs according to SARI, with GPT3.5-Turbo, Sabia-2-small and both baselines very close to it. Meanwhile, we can see that only qwen2.5-7b-instruct and r1-distill-qwen-32b are competitive for open-weight contenders, achieving the best balance between simplicity and meaning preservation according to automatic metrics. In this dataset, both baselines achieved the highest meaning preservation metric. This can be explained by the fact that the reference sentences are not very different from the input sentences, indicating a non-aggressive simplification process. This favors baselines that make fewer changes to the input. This can be confirmed by their high value of the % of unchanged outputs metric (Appendix D).

In **Museum-PT**, we observe a decrease in meaning preservation compared to the PorSimplesSent dataset. This can be explained by the particular domain, with many words and phrases coming from the subject of physics. In terms of simplicity, GPT models outperform all LLMs and baselines by a reasonable margin. This superiority might indicate a higher and broader level of training data than the other LLMs. qwen2.5-32b-instruct and Sabiá models also achieved good results, with a good balance between content preservation and simplicity.

In **Gov-Lang-BR**, GPT4o-mini and Sabia-2-small achieved the highest values for both metrics, with very similar values. Although Sabia-2-small achieved the highest value for content preservation, GPT4o-mini achieved the highest simplicity metric. The optimal result of the Brazilian language model is probably because this dataset is the most specific to Brazilian Portuguese, containing many legal terms and terminology from the Brazilian public administration. qwen2.5-32b-instruct and r1-distill-qwen-32b are competitive, achieving the best balance between simplicity and meaning preservation according to automatic metrics, and having a SARI score next to GPT4o-mini. Since GPT4o-mini outperforms GPT3.5T and is cheaper, the latter was not evaluated on the Gov-Lang-BR.

4.2 Morphosyntactic analysis of the Sentence Simplification task

We perform a large-scale linguistic analysis of the transformations performed during the simplification by the GPT3.5-Turbo and GPT-4o-mini. For the PorSimplesSent and Museum-PT datasets, we analyze the simplifications of GPT-3.5 Turbo LLM and not GPT-4o-mini, as the latter was not yet available at the time of the analysis. To interpret the simplification process carried out by LLMs and determine what they are doing or failing to do, we performed a morphosyntactic analysis of simplifications generated by both humans and LLMs.

As Section 3.4 explains, twelve simplifications are generated for each input sentence during inference. Two sets of simplified sentences were created to perform a linguistic analysis of the LLM's simplifications. One set always contains the best-generated sentence among the twelve, and the other contains the worst, both according to the SARI metric. For each dataset, we morphosyntactically annotated four sets of data: the original complex sentences, their respective human simplification references, the best simplifications generated by the LLM, and the worst ones. 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 trained on a scientific treebank (Straka et al., 2016; de Souza et al., 2021). Then, we calculate the linguistic metrics and choose the most impacted simultaneously in PorSimplesSent and Museum-PT datasets (Section 3.5).

Dataset	Linguistic Metrics			
	LTR	P/A	AdvLeft	D/R
PorSimplesSent				
Complex	.224	.010	.49	.81
Simple	.198	.009	.26	1.03
BestGPT3.5T	.216	.012	.26	.99
WorstGPT3.5T	.227	.012	.26	.93
Museum-PT				
Complex	.147	.016	.33	.91
Simple	.128	.005	.54	2.56
BestGPT3.5T	.159	.012	.35	1.34
WorstGPT3.5T	.165	.018	.30	1.09
Gov-Lang-BR				
Complex	.050	.011	.071	.59
Simple	.062	.014	.051	.67
BestGPT4o-m	.052	.013	.054	1.28
WorstGPT4o-m	.052	.013	.058	1.21

Table 2: Linguistic Metrics for three datasets

The results in Table 2 point out to what extent the language models followed or diverged from the human simplification trends. The PorSimplesSent and Museum-PT datasets show that the best simplification set metrics are always closer to the reference metrics than the worst simplification set. It indicates that our chosen linguistic metrics indeed correlate with the SARI metric.

Moreover, despite the high SARI metric obtained by the best set, there is still room for improvement in the simplifications compared to the linguistic metrics of the reference set. In particular, the passive-to-active voice, the developed-to-reduced relative clauses, and the LTR metrics can be significantly improved in both the Museum-PT and PorSimplesSent to achieve reference standards.

Regarding the Gov-Lang-BR dataset, we observe that its reference sentences do not follow two of the three trends observed in the other two datasets. We see an increase in the lexical diversity, indicated by the LTR metric, and in the passive-to-

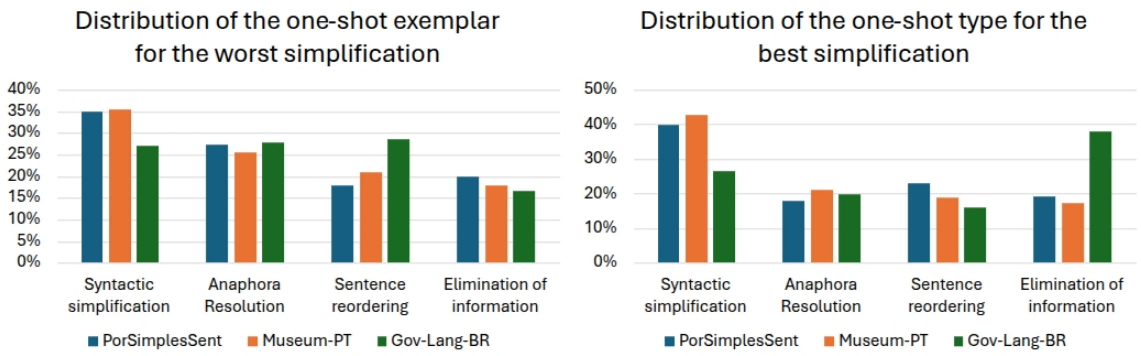

Figure 1: Distribution of the one-shot type for the worst and best simplifications generated by the GPT4o-mini.

Dataset	Worst SARI	Average SARI	Best SARI
PorSimplesSent	31.37	39.75	49.27
Museum-PT	40.62	48.92	57.47
Gov-Lang-BR	37.69	45.14	53.53

Table 3: Range of SARI values reached by GPT4o-mini LLM

active voice ratio. This is likely due to the fact that there is no guarantee that linguists specialized in plain language were involved in its creation. A fact that supports this hypothesis is that the developed-to-reduced relative clause metric obtained by the LLM, both for the best and worst sets, was higher than that of the reference set.

4.3 One-shot Exemplars Analysis

For each dataset, if we generate two sets of sentence simplifications – one by consistently selecting the simplification with the lowest SARI score among the twelve generated by the LLM, and the other by selecting the one with the highest SARI score – we can establish the minimum and maximum performance extremes of the LLMs according to the SARI metric. Looking at these values in Table 3 for the GPT4o-mini model, we can see that this range can vary significantly. This variance comes from the stochastic nature of the LLM and the type of one-shot exemplar provided to the LLM during inference. While making it deterministic would compromise its behavior, the one-shot example can be selected to optimize the results.

Here, we investigate whether exemplar type impacts simplification performance or if the choice is negligible. To this end, we identify which exemplar type yields the best and worst simplifications for each complex sentence.

Figure 1 shows the distribution of the best and worst one-shot simplification types. As we can see, the anaphora resolution and sentence reordering examples are rarely the best simplification types in all datasets and are the worst with a higher frequency. The elimination of redundant information was the most successful exemplar in the Gov-Lang-BR dataset, being the best almost 40% of the time and the worst only about 15% of the time. In both PorSimplesSent and Museum-PT datasets, the syntactic simplification type produces the best simplifications more than 40% of the time and the worst about 30% of the time.

The overall results indicate that exemplars with syntax and lexical edits are more likely to impact the simplification process. Public language tends to be bureaucratic, with technical jargon, and often verbose, making using examples with lexical changes and eliminations sensible. On the other hand, examples simplifying structure seem to aid LLMs more in journalistic and scientific styles.

5 Qualitative Analysis

Automatic metrics are recognized for having limitations and are not always entirely reliable (He et al., 2023). We perform a human qualitative analysis on 180 system outputs to alleviate this issue. We follow a mix of bottom-up and top-down strategies for conducting the manual analysis (van Miltenburg et al., 2021). The bottom-up refers to selecting the three LLMs with the best-observed results following the SARI metrics. Then, for each dataset, we randomly select 20 simplifications from each one of them for annotation[11]. Next, the top-down component of the strategy involves defining eight key questions related to the simplification

[11] Annotations were answered by one of the authors and reviewed by another.

Model-Dataset	%S	%MP	%L	%S	%D	%Sp	%R	%H
Qwen2.5-7B-PorSimplesSent	75.0	65.0	80.0	60.0	55.0	0.0	25.0	0.0
Qwen2.5-7B-Museum-PT	85.0	80.0	70.0	65.0	45.0	0.0	25.0	0.0
Qwen2.5-7B-Gov-Lang-BR	65.0	65.0	85.0	60.0	75.0	5.0	10.0	5.0
Qwen2.5-7B	75.0	70.0	78.3	61.7	58.3	1.7	20.0	1.7
Sabia-2-S-PorSimplesSent	65.0	70.0	65.0	65.0	45.0	0.0	30.0	10.0
Sabia-2-S-Museum-PT	80.0	65.0	55.0	50.0	55.0	0.0	20.0	5.0
Sabia-2-S-Gov-Lang-BR	50.0	40.0	65.0	35.0	85.0	0.0	5.0	5.0
Sabia-2-S	65.0	58.3	61.7	50.0	61.7	0.0	18.3	6.7
GPT4o-m-PorSimplesSent	85.0	85.0	60.0	55.0	50.0	5.0	25.0	5.0
GPT4o-m-Museum-PT	100	85.0	85.0	65.0	60.0	0.0	25.0	0.0
GPT4o-m-Gov-Lang-BR	90.0	70.0	85.0	75.0	75.0	0.0	0.0	0.0
GPT4o-m	91.7	80.0	76.7	65.0	61.7	1.7	16.7	1.7

Table 4: Results of our qualitative analysis. The questions are S: accepted simplification, MP: meaning preserved, L: lexical edit, S: syntactic edit, R: reordering, D: deletion, Sp: sentence splitting, H: hallucination.

process. These questions aim to assess different aspects of the generated simplifications: *Is the output a valid simplification?*, *Is the meaning preserved?*, *Was there a lexical change?*, *Was there a syntactic change?*, *Was there a deletion operation?*, *Was there a sentence splitting?*, *Was there a sentence reordering?*, *Is the output a hallucination?*.

We followed the types of edit operations described in Heineman et al., 2023, but we assumed it was unnecessary to annotate whether there was a lexical insertion specifically. The question regarding content preservation already addresses the cases of added information, making further consideration redundant. We consider a simplification valid if it is simpler than the input and has no inappropriate changes to the original text's meaning and ungrammatical outputs. The meaning is preserved if the general information remains in the simplified sentence. We annotate a simplification as a hallucination if the generation possesses information that is not in the input and cannot be directly inferred. Table 4 shows the results of this analysis.

Similarly to the findings of automatic results, GPT4o-mini is the best model, considering both simplification and meaning preservation capabilities. However, the superiority of Sabia-2-small compared with Qwen2.5-7B was not observed in terms of both simplicity and meaning preservation. This poor result came mainly from the negative analysis of the Gov-Lang-BR dataset, which contains many long sentences that make the simplification process quite difficult, misleading the automatic metrics. Since only 20 sentences from this model were evaluated in this dataset, randomness may have contributed to this poor result. We also observed that Qwen2.5-7B and GPT4o-m have very similar distributions of operations, with high values of lexical and syntactic operations. On the other hand, Sabia-2-small has fewer lexical and syntactic operations and much more hallucinations.

6 Conclusion

This paper evaluated how recent LLMs perform in Portuguese SS in the one-shot in-context learning scenario. We found that the best LLMs outperform baselines trained specifically for the task, while also producing a more diverse set of simplifications. We also established that closed-weight models perform better than open-weight ones. However, the best open-weight LLM achieved very competitive results. Our qualitative analysis endorsed the results of the automatic metrics in this regard. We demonstrated that 7B and 32B LLMs can achieve good results on a single 24GB GPU using modern quantization techniques.

The linguistic metrics extracted from the best performing LLMs showed that LLMs still have a gap to fill when comparing their simplifications to those generated by humans. Our analyses of the one-shot exemplars revealed that syntactic and lexical simplification examples are more suitable for prompting the LLM, being the most likely examples to generate the best simplification. This benchmark has established a solid base to guide future Portuguese SS research. Future research could investigate alternative document-level simplification methods and incorporate pre-trained LLMs in fine-tuning or retrieval-based scenarios.

Limitations

While our study provides valuable benchmark results for the sentence simplification task in Portuguese, there are some limitations that should be acknowledged. First, we cannot guarantee that the simplified sentences in Gov-Lang-BR were subjected to linguistic validation by experts. We could not acquire this information from the administrative sectors that make the sentences available on their web page. This way, although the data reflects real-world usage, the lack of formal validation may introduce noise, particularly in the case of regional and colloquial variations in Portuguese, or lack of a unified guide of simplification. Moreover, while we motivate our work by analyzing a language less explored than English, for example, our findings cannot generalize to other languages or even to other variations of Portuguese spoken in less represented countries like Mozambique.

Second, our approach relied on one-shot and in-context learning, rather than fine-tuning LLMs. While this choice was made to test the general adaptability of LLMs without additional training, it limits the depth of model optimization that could have been achieved through more focused fine-tuning. In practice, fine-tuning a specific Portuguese dataset could yield better performance and more precise handling of linguistic nuances.

Finally, due to resource constraints, we could not conduct as many experiments as would have been ideal for a thorough exploration of the model's capabilities. Given infinite resources, additional experiments—including hyperparameter tuning and fine-tuning large and small language models could have provided more comprehensive insights.

Ethics Statement

In the context of sentence simplification, it is essential to acknowledge the ethical considerations related to simplifying texts without taking into account the specific needs or abilities of the individuals receiving the simplified content. Simplification without understanding the unique challenges of the target audience – whether related to cognitive disabilities, language proficiency, or educational background – risks reducing the accessibility of the text. This one-size-fits-all approach may oversimplify content, stripping it of important nuance, context, or meaning. Moreover, by not regarding the level of simplification to the individual's needs, we may unintentionally disempower users who require different levels of complexity in the text. Some users might benefit from simplified language, while others might need different types of assistance, such as more detailed explanations or visual aids, to better understand complex ideas. Failing to account for these factors could perpetuate inequities in access to information, particularly for marginalized groups or individuals with specific learning or language challenges. In light of these concerns, future work on sentence simplification should consider a more inclusive approach that accounts for individual differences in language processing and comprehension.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024a. Phi-3 technical report: A highly capable language model locally on your phone.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024b. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Conference on Empirical Methods in Natural Language Processing*.

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3757–3769, Online. Association for Computational Linguistics.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

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

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Alexandre Alves, Péricles Miranda, Rafael Mello, and André Nascimento. 2023. Automatic simplification of legal texts in portuguese using machine learning. In *Legal Knowledge and Information Systems*, pages 281–286. IOS Press.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023b. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.

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

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

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

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks.

Elvis de Souza, Aline Silveira, Tatiana Cavalcanti, Maria Clara Castro, and Cláudia Freitas. 2021. Petrogold–corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38. SBC.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *ArXiv*, abs/2302.11957.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Maria José Bocorny Finatto and Lucas Meireles Tcacenco. 2021. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. *Tradterm*, 37(1):30–63.

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

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh

Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024a. Olmo: Accelerating the science of language models.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024b. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Nathan Siegle Hartmann and Sandra Maria Aluísio. 2020. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. *Linguamática*, 12(2):3–27.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Maliheh Izadi, Jonathan Katzy, Tim Van Dam, Marc Otten, Razvan Mihai Popescu, and Arie Van Deursen. 2024. Language models for code completion: A practical evaluation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Taeuk Kim. 2022. Revisiting the practical effectiveness of constituency parse extraction from pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5398–5408, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Liutao Liutao, Jinwang Song, Hongying Zan, Sun Li, and Deyi Xiong. 2024a. OpenEval: Benchmarking Chinese LLMs across capability, alignment and safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–210, Bangkok, Thailand. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024b. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. 2022. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3550–3562, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

ELHAM MADJIDI and CHRISTOPHER CRICK. 2024. Towards inclusive reading: A neural text generation framework for dyslexia accessibility.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Ramon Pires, Hugo Abonizio, Thales Rog'erio, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Xinying Qiu and Jingshen Zhang. 2024. Label confidence weighted learning for target-level sentence simplification.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Eric Sven Ristad and Peter N. Yianilos. 1996. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:522–532.

Michael Joseph Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-english text simplification: A unified multilingual benchmark. In *Annual Meeting of the Association for Computational Linguistics*.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Arthur Scalercio, Maria Finatto, and Aline Paes. 2024. Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15076–15091, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP* *for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*, pages 30–39. Springer.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024b. Gemma: Open models based on gemini research and technology.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024c. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards multilingual llm evaluation for european languages.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169.

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

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

A Gov-Lang-BR Information

Table 5 displays the distribution of sentences in the dataset according to their originating government agency.

As can be observed, most of the data came from the executive branch, but there are also 52 examples originating from judicial branch courts. The language originating from the judiciary is more focused on legal terms. On the other hand, texts from the executive branch, sourced from departments of finance, planning, and regulatory agencies, focus on administrative terms specific to the tax and financial areas. In the case of the regulatory agency INMETRO (National Institute of Metrology, Quality, and Technology), the texts describe technical terms outlining inspection procedures.

Table 6 displays some surface statistics of the three corpora used.

Agency	**Level**	**Branch**	**#Pairs**
INMETRO	Federal	Executive	63
Secretaria de Planejamento – Niterói	City	Executive	1487
Secretaria de Fazenda – Mato Grosso	State	Executive	101
Tribunal de Justiça – Rio de Janeiro	State	Judicial	4
Tribunal de Justiça – Rio Grande do Sul	State	Judicial	40
Tribunal Regional Eleitoral – Paraná	Federal	Judicial	8
Total			**1703**

Table 5: Distribution of Sentence Pairs by Government Agency

Dataset	**Style**	**# Sentences**	**Tokens per Sentence**
PorSimplesSent	Complex	606	22.52
	Simple	606	21.88
Museum-PT	Complex	476	21.44
	Simple	476	15.60
Gov-lang-BR	Complex	1703	33.49
	Simple	1703	21.15

Table 6: Statistics of Different Datasets

B LLM Details

Table 7 presents the characteristics of the 20 selected open-weight LLMs, including quantization type, number of parameters, and Hugging Face model name.

Below, we briefly describe some information related to each LLM considered in this paper.

1. GPT (Generative Pre-trained Transformer) (Brown et al., 2020) is one of the most widely recognized large language models. We considered the 3.5, 4o-mini

and o1-mini versions of GPT, all of them were trained by OpenAI along the years with increasing data and larger architecture. They are pre-trained on vast amounts of text data from the internet. o1-mini is a more affordable reasoning model from openAI. These models excel at a wide range of tasks, including text generation, translation, summarization, and code completion (Basyal and Sanghvi, 2023; Wu and Hu, 2023; Li et al., 2024; Izadi et al., 2024). GPT models are known for their general-purpose capabilities. However, GPT is a closed-weight model, accessible only via API or downloadable software, with its architecture and training details unavailable to the public.

2. Qwen (Bai et al., 2023b)[12], created by Alibaba, is an advanced LLM stably pretrained for up to 3 trillion tokens of multilingual data (with a focus on Chinese and English) with a wide coverage of domains (Bai et al., 2023b). It includes models designed for various tasks such as text creation, translation, dialogue simulation, and even multimodal tasks involving audio, vision, and structured data. The Qwen series includes models with 7, 14, and up to 72 billion parameters, with instruction-tuned versions for better alignment with user needs. A notable feature of Qwen is its use of a technique called Group Query Attention (Ainslie et al., 2023), which optimizes performance by improving both speed and memory efficiency during inference. We also evaluated dense models based on the Qwen architecture, distilled from DeepSeek-R1 (DeepSeek-AI et al., 2025), a reasoning model that has achieved strong performance across multiple LLM benchmarks.

3. LLaMA (Touvron et al., 2023a), developed by Meta AI, is a family of open-source LLMs that has evolved through several iterations, with the latest being Llama 3, is an open-source model under Meta's licensing designed for efficiency and accessibility. The models are pre-trained on an extensive dataset of approximately 15 trillion tokens, providing them with a broad knowledge base for tasks such as text generation, multilingual translation, and more. LLama 3 includes a more efficient tokenizer, group Query Attention, extended context window, and multimodal capabilities. Llama 3 is designed to be a competitive open-source alternative to proprietary models like GPT-4, with a strong focus on multilingual capabilities and computational efficiency.

4. Command-R [13] is part of Cohere's series of enterprise-grade language models designed specifically for retrieval-augmented generation (RAG) and tool use at a production scale. This model has a 128K token context limit, allowing it to handle long, complex conversations and detailed queries accurately. Command-R integrates with other Cohere tools, such as Embed and Rerank, further enhancing its ability to retrieve and optimize relevant information for end-users. The latest version, Command-R+ (released in 2024), offers efficiency, latency, and performance improvements while maintaining a lower computational cost than models like GPT-4. It is well-optimized for multilingual tasks, handling over 10 languages (including Portuguese). Aya-23 (Aryabumi et al., 2024), also developed by Cohere, is an open weights research release of an instruction fine-tuned model with highly advanced multilingual capabilities. It covers 23 languages, including Portuguese.

5. Mistral 7b[14], trained by the AI French startup of the same name, is an open-weight LLM released in September 2023. Mistral uses Grouped-query attention for faster inference and Sliding Window Attention to handle longer sequences at smaller cost. It supports multiple languages, including Portuguese, along with 80+ coding languages. The model is accessible under both non-commercial and commercial licenses.

6. OLMo (Groeneveld et al., 2024b) developed by AI2, is designed to accelerate research and development in language modeling by providing a fully transparent framework. Unlike most language models that only release weights and inference code, OLMo offers open access to training data, training

[12] https://github.com/QwenLM/Qwen2.5

[13] https://docs.cohere.com/docs/command-r

[14] https://mistral.ai/news/announcing-mistral-7b/

code, evaluation code, and intermediate checkpoints, allowing researchers to thoroughly study the impact of pretraining and architecture decisions. This transparency supports a deeper understanding of LLMs' behavior, biases, and performance. OLMo has been trained on the Dolma dataset, composed of 3 trillion tokens from various data sources, including web content, books, code repositories, and academic publications. This open dataset is structured to allow researchers to experiment with and reproduce the effects of different data curation and filtering techniques on model performance. OLMo currently comes in models with 1B and 7B parameters. It has demonstrated competitive performance across a range of NLP benchmarks.

7. The Phi (Li et al., 2023; Abdin et al., 2024b) family of models, developed by Microsoft, represents a series of small language models (SLMs) designed to offer impressive performance with fewer parameters. The Phi-3 series, introduced in 2024, includes models ranging from 3.8 billion to 14 billion parameters, and despite their smaller size, these models achieve results comparable to much larger models like GPT-3.5. Phi-3-mini is a 3.8 billion parameter model capable of handling up to 128K tokens. Phi-3-medium has 14 billion parameters and was trained on 4.8 trillion tokens. Microsoft's focus is on optimizing datasets—using high-quality, filtered web data and synthetic data. Phi models are also available for use and further development on models hub platforms.

8. Gemma[15] (Team et al., 2024a,c) is a family of lightweight, open-source language models developed by Google DeepMind, based on the technology behind the Gemini models. It includes models with 2 billion and 7 billion parameters, optimized for processing up to 8192 tokens at once. Gemma's key architectural features include GeGLU activation functions and multi-query attention for the 2B model, which helps with efficiency. In comparison, the 7B model uses multi-head attention for richer representations. Gemma's large vocabulary size (256,000 tokens) allows it to handle diverse inputs, including multilingual text.

9. Sabiá (Pires et al., 2023) is a family of LLMs designed explicitly for Portuguese, developed by Maritaca AI. These models were built upon popular architectures like LLaMA and GPT-J but are fine-tuned on a vast corpus of Portuguese text. This specialization allows Sabiá to outperform many English-centric or multilingual models on tasks involving the Portuguese language. The models were evaluated using the Poeta benchmark, consisting of 14 Portuguese datasets spanning different NLP tasks such as text classification, natural language inference, etc. Results show that by focusing solely on Portuguese allows Sabiá models to capture linguistic nuances specific to the language better, giving them an edge in understanding and generating Portuguese text. The model is open-source and available for further experimentation via platforms like Hugging Face. Since its first version, Sabiá has evolved to models trained with larger architecture and corpora.

C Linguistic Metrics Selection

18 morphosyntactic characteristics have been considered to compare the original sentences, references simplified by humans, and simplified texts by GPT3.5-Turbo. Table 8 presents their values along with the number of tokens, sentences, and entries for each dataset. We selected only four of them to compose the model's simplifications comparison because, in only four of them, the human simplifications were consistent across datasets. Here, we explain each of the tested metrics:

Number of tokens per sentence: higher numbers indicate longer sentences.

Type/Token Ratio (TTR): higher numbers indicate greater lexical diversity (considering the form of words). The calculation is made by dividing the number of unique tokens by the total number of tokens in the corpus.

Lemma/Token Ratio (LTR): higher numbers indicate greater lexical diversity (considering the uninflected form – the lemma – of words). The calculation is made by dividing the number of unique lemmas by the total number of tokens in the corpus.

Comma to token ratio: a higher number of commas may indicate a greater number of syntactic shifts.

[15] https://developers.googleblog.com/en/gemma-explained-overview-gemma-model-family-architectures/

Clause to sentence ratio: a higher number of clauses indicates a greater number of verb heads.

Sentence to entry ratio: higher numbers indicate more segmentation of original texts into multiple sentences.

In the example below, the simplified entry (2) consists of 3 sentences, while the original entry (1) consists of only one sentence. In this case, the sentence-to-entry ratio is 1:1 for the original corpus and 3:1 for the simplified one.

Original Museum-PT: (1) *Aperte o botão para ligar o equipamento e gire o disco óptico.*[16]

Simplified Museum-PT: (2) *Aperte o botão para ligar o equipamento. Depois, gire o disco óptico. Você conseguirá produzir alguns feixes de luz, ou seja, pequenos raios.*[17]

Verb to noun ratio: the higher the number, the greater the number of verbs, possibly indicative of actions, as opposed to nouns, possibly indicative of concepts and abstractions.

Adjective to noun ratio: higher numbers indicate a more detailed description, as more adjectives are applied to the relevant nouns.

Adverb to verb ratio: higher numbers indicate a more detailed description of verbal actions (which can occur, for example, "quickly" or "slowly").

Postverbal to preverbal subject ratio: higher numbers indicate a greater number of subjects following the verb they refer to, which characterizes an inversion of the standard syntactic order of Portuguese.

In the example below, the original entry (3) has a postverbal subject (*caverns/cavernas* to the right of the verb *evolve/evolvem*). In the simplified entry (4), the structure is changed so that *caverns/cavernas* is the object of the verb *have/temos*, where it is expected that the object appears to the right of the verb, as the subject of *have/temos* is elliptical (we/nós).

Original Museum-PT: (3) *De sua ampliação e interligação evoluem as cavernas propriamente ditas.*[18]

Simplified Museum-PT: (4) *Quando os espaços por onde a água passa aumentam de tamanho e se ligam a outros espaços, temos as cavernas propriamente ditas.*[19]

Passive to active voice ratio: higher numbers indicate a greater amount of passive voice, when the position of the object and the subject are inverted.

In the example below, the original sentence (5) has the verb in the passive voice, where *equipment/equipamento* functions as the patient subject of a passive clause. In the simplified sentence (6), the structure of the sentence is in the active voice, where the subject is simple, *you/você*, and the verb *will need/precisará* is in the active voice.

Original Museum-PT: (5) *Esse equipamento deve ser utilizado por duas pessoas.*[20]

Simplified Museum-PT: (6) *Para utilizar este equipamento, você precisará de outra pessoa.*[21]

Proportion of verbal periphrases: higher numbers indicate a greater number of complex verb heads composed of more than one verb.

Still using examples (5) and (6), we see that in the original sentence there is a verbal periphrasis (*should be used/deve ser utilizado*), while in the simplified sentence there is only one simple verb, *will need/precisará*.

Proportion of adverbial subordinate clauses: higher numbers indicate a greater number of adverbial clauses.

In sentence (6), we can see the use of an adverbial clause that did not exist in the original sentence: *to use this equipment/para utilizar este equipamento*, indicating the purpose of the main clause verb *will need/precisará*.

Proportion of adverbial subordinate clauses to the left of the head: higher numbers indicate more adverbial clauses to the left of the main clause, an inversion of the standard syntactic order.

Still, in sentence (6), we can see that the adverbial clause is to the left of the main clause, thus requiring a comma to mark the syntactic shift since, in the natural syntactic order of the Portuguese language, adverbial adjuncts come to the right of the verb they modify.

Proportion of developed to reduced relative clauses: higher numbers indicate a greater amount of noun modification by means of relative clauses.

In the example below, we see that a simplification solution (8) was to transform what originally (7) were nouns, *production/produção* and *confinement/confinamento*, into reduced relative clauses,

[16] Press the button to turn on the equipment and rotate the optical disc.

[17] Press the button to turn on the equipment. Then, rotate the optical disc. You will be able to produce some light beams, i.e., small rays.

[18] From their expansion and interconnection, the caverns themselves evolve.

[19] When the spaces through which the water passes expand and connect to other spaces, we have the caverns themselves.

[20] This equipment should be used by two people.

[21] To use this equipment, you will need another person.

to produce/produzir and *to isolate/isolar*. Another option could have been the use of developed relative clauses, where the verb is in a finite form and the subordinating conjunction is explicit, for example: *developed a powerful machine that produces and isolates plasma/desenvolveram uma máquina poderosa que produz e isola plasma*.

Original Museum-PT: (7) *Na Rússia foi desenvolvida uma potente máquina para produção e confinamento de plasma, o Tokamak, em 1960, com a finalidade de gerar energia elétrica.*[22]

Simplified Museum-PT: (8) *Em 1960, na Rússia, os cientistas desenvolveram uma potente máquina para produzir e isolar plasma: o Tokamak. Essa máquina serviria para gerar energia elétrica.*[23]

Proportion of objective noun clauses: higher numbers indicate a greater number of objects (verbal complements) in the form of clauses.

In the example below, the original sentence (9) has a direct objective subordinate noun clause, whose head is *have/têm* and whose main clause is *observe*. In the human simplification (10), the two clauses gave way to only one sentence, whose head is *have/têm*.

Original Museum-PT: (9) *Observe que os dois objetos têm a mesma massa, pois a balança encontra-se em equilíbrio.*[24]

Simplified Museum-PT: (10) *Os dois objetos têm a mesma massa, pois a balança está equilibrada.*[25]

Proportion of coordinated clauses: higher numbers indicate a greater number of coordinated clauses (verbs).

Proportion of coordinated nominals: higher numbers indicate a greater number of coordinated nominals (nouns, adjectives, pronouns, etc.).

D Additional Results

Tables 9, 10, and 11 show full simplification results on PorSimplesSent, Museum-PT and Gov-Lang-BR, respectively.

[22]In Russia, a powerful machine for the production and confinement of plasma, the Tokamak, was developed in 1960, with the purpose of generating electricity.

[23]In 1960, in Russia, scientists developed a powerful machine to produce and isolate plasma: the Tokamak. This machine would serve to generate electricity.

[24]Observe that the two objects have the same mass, as the scale is in balance.

[25]The two objects have the same mass, as the scale is balanced.

E Prompts and Demonstration Examples

We followed recent Portuguese sentence simplification work (Scalercio et al., 2024) for preparing our prompt and selecting demonstration examples. As there, the instruction follows Feng et al. (2023): "*Substitua a frase complexa por uma frase simples. Mantenha o mesmo significado, mas torne-a mais simples.*
Frase complexa: {original}
Frase Simples: "[26].

And the one-shot exemplars are disposed in Table 12. Here, we add the simplification category that guided the selection of exemplars.

[26]In English: *"Replace the complex sentence with a simple sentence. Keep the same meaning but make it simpler.*
Complex sentence: {original}
Simple Sentence: "

Table 7: Characteristics of selected open-weight LLMs

Arch	Param	Model	Quantiz	Hugging Face Model Name
command-r	8B	aya-23-8b	Q4_K_M	bartowski/aya-23-8B-GGUF
gemma2	27B	gemma-2-27b-it	Q4_K_M	bartowski/gemma-2-27b-it-GGUF
llama	8B	meta-llama-3.1-8b-instruct	Q8_0	lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF
llama	8B	meta-llama-3.1-8b-instruct	Q4_K_M	lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF
llama	3B	llama-3.2-3b-instruct	Q4_K_M	lmstudio-community/Llama-3.2-3B-Instruct-GGUF
llama	8B	llama-2-7b-chat	Q4_K_M	TheBloke/Llama-2-7B-Chat-GGUF
llama	8B	meta-llama-3-8b	Q4_K_M	QuantFactory/Meta-Llama-3-8B-GGUF
llama	7B	mistral-7b-instruct-v0.3	Q4_K_M	MaziyarPanahi/Mistral-7B-Instruct-v0.3-GGUF
olmo	7B	olmo-7b-instruct	Q4_K_M	ssec-uw/OLMo-7B-Instruct-GGUF
phi3	14B	phi-3-medium-128k-instruct	Q4_K_M	bartowski/Phi-3-medium-128k-instruct-GGUF
phi3	3B	phi-3.5-mini-instruct	Q4_K_M	bartowski/Phi-3.5-mini-instruct_Uncensored-GGUF
qwen2	7B	qwen2-7b-instruct@q4_k_m	Q4_K_M	Qwen/Qwen2-7B-Instruct-GGUF
qwen2	70B	qwen2-72b-instruct	Q4_K_M	Qwen/Qwen2-72B-Instruct-GGUF
qwen2	7B	qwen2.5-7b-instruct@q8_0	Q8_0	lmstudio-community/Qwen2.5-7B-Instruct-GGUF
qwen2	7B	qwen2.5-7b-instruct@q4_k_m	Q4_K_M	lmstudio-community/Qwen2.5-7B-Instruct-GGUF
qwen2	14B	qwen2.5-14b-instruct	Q4_K_M	lmstudio-community/Qwen2.5-14B-Instruct-GGUF
qwen2	32B	qwen2.5-32b-instruct	Q4_K_M	lmstudio-community/Qwen2.5-32B-Instruct-GGUF
qwen2	7B	deepseek-r1-distill-qwen-7b	Q4_K_M	lmstudio-community/DeepSeek-R1-Distill-Qwen-7B-GGUF
qwen2	14B	deepseek-r1-distill-qwen-14b	Q4_K_M	lmstudio-community/DeepSeek-R1-Distill-Qwen-14B-GGUF
qwen2	32B	deepseek-r1-distill-qwen-32b	Q4_K_M	lmstudio-community/DeepSeek-R1-Distill-Qwen-32B-GGUF

Table 8: Linguistic Metrics across datasets

Metric	Museum-PT		Porsimplessent		Gov-Lang-BR	
	Complex	Simple	Complex	Simple	Complex	Simple
Number of tokens	10676	11016	14322	13961	70199	40034
Number of sentences	498	706	636	638	2096	1893
Number of entries	476	476	606	606	1703	1703
Number of tokens per sentence	21.44	15.60	22.52	21.88	33.49	21.15
Type/Token Ratio (TTR)	0.19	0.17	0.28	0.26	0.07	0.09
Lemma/Token Ratio (LTR)	0.15	0.13	0.22	0.20	0.05	0.06
Comma to token ratio	0.05	0.04	0.05	0.04	0.06	0.04
Clause to sentence ratio	2.67	2.03	2.46	2.47	2.82	2.08
Sentence to entry ratio	1.05	1.48	1.05	1.05	1.23	1.11
Verb to noun ratio	0.51	0.52	0.48	0.48	0.27	0.30
Ajective to noun ratio	0.291	0.223	0.260	0.232	0.299	0.244
Adverb to verb ratio	0.328	0.338	0.388	0.360	0.213	0.179
Postverbal to preverbal subject ratio	0.031	0.038	0.074	0.059	0.038	0.018
Passive to active voice ratio (P/A)	0.016	0.005	0.010	0.009	0.011	0.014
Proportion of verbal periphrases	0.115	0.108	0.153	0.159	0.127	0.097
Proportion of adverbial subordinate clauses	0.214	0.158	0.143	0.123	0.124	0.132
Proportion of adverbial subordinate clauses to the left of the head (AdvLeft)	0.326	0.537	0.493	0.260	0.071	0.051
Proportion of developed to reduced relative clauses (D/R)	0.915	2.56	0.815	1.03	0.594	0.668
Proportion of objective noun clauses	0.030	0.045	0.064	0.072	0.018	0.033
Proportion of coordinated clauses:	0.092	0.068	0.051	0.056	0.097	0.098
Proportion of coordinated nominals	0.154	0.150	0.146	0.140	0.845	0.545

Model	SARI	BertS	Bleu	% U
Baselines				
MUSS	38.30	.8976	51.38	3.46
Enh-PT-SS	39.64	.9024	48.2	3.79
Open-weight LLMs				
Aya23-8B	33.87	.8534	26.54	1.66
Gemma2-27B	30.83	.8352	17.08	0
Llama2-7B	27.25	.7993	16.48	2.54
Llama3-8B	31.60	.7658	21.77	5.69
Llama3.1-8B	30.17	.8289	16.31	0.11
Llama3.1-8B-q8	29.55	.8257	15.12	0.07
Llama3.2-3B	30.24	.8104	19.53	3.95
Mistral-7B	33.08	.8465	24.46	0.03
OLMo-7B	27.96	.7864	15.54	0.37
Phi-3-medium	29.06	.8230	15.18	0
Phi3.5-mini	28.97	.7442	13.30	1.24
Qwen2-7B	35.75	.8661	28.84	0.25
Qwen2-72B	34.69	.8576	24.67	0
Qwen2.5-7B	**36.61**	**.8701**	31.19	0.77
Qwen2.5-7B-Q8	36.30	.8694	29.92	0.11
Qwen2.5-14B	33.96	.8534	23.86	0.04
Qwen2.5-32B	35.81	.8651	26.97	0
r1-distill-7b	34.95	.8523	33.59	6.26
r1-distill-14b	29.47	.7373	16.28	0.92
r1-distill-32b	36.46	.8689	29.51	0.50
Closed-weight LLMs				
Command-R	32.60	.8329	21.97	0
Gpt-3.5-T	39.18	.8805	38.01	0.26
Gpt-4o-m	**39.75**	**.8838**	35.17	0
o1-mini	39.26	.8472	35.06	0.04
Sabia-2-S	38.16	.8732	35.46	0.85
Sabia-3	35.12	.8546	26.33	0.26

Table 9: Simplification Results on PorSimplesSent

Model	SARI	BertS	Bleu	% U
Baselines				
MUSS	39.31	.8534	32.12	3.99
Enh-PT-SS	41.62	.8550	32.36	5.46
Open-weight LLMs				
Aya23-8B	43.61	.8269	19.82	1.59
Gemma2-27B	41.12	.8130	12.55	0.05
Llama2-7B	34.52	.7577	9.72	3.12
Llama3-8B	35.45	.7428	14.50	8.54
Llama3.1-8B	40.28	.8101	12.39	0.14
Llama3.1-8B-q8	39.65	.8070	11.45	0.03
Llama3.2-3B	38.56	.7897	13.18	4.35
Mistral-7B	41.32	.8154	16.20	0.04
OLMo-7B	34.81	.7592	8.31	0.68
Phi-3-medium	38.56	.8002	10.48	0
Phi3.5-mini	35.24	.7279	8.36	1.61
Qwen2-7B	44.54	.8319	20.18	0.17
Qwen2-72B	43.94	.8296	17.22	0.07
Qwen2.5-7B	44.20	.8347	21.43	0.50
Qwen2.5-7B-Q8	44.51	.8354	21.37	0.25
Qwen2.5-14B	43.42	.8183	17.86	0.17
Qwen2.5-32B	**45.74**	**.8369**	19.93	0.33
r1-distill-7b	39.11	.8120	19.98	6.81
r1-distill-14b	38.65	.7270	11.40	1.22
r1-distill-32b	44.69	.8352	20.13	0.88
Closed-weight LLMs				
Command-r	42.79	.8110	16.88	0
Gpt-3.5-T	47.23	.8468	26.27	0.63
Gpt-4o-m	**48.92**	**.8508**	25.84	0.14
o1-mini	47.26	.8252	24.23	0.07
Sabia-2-S	44.44	.8353	23.70	0.71
Sabia-3	44.72	.8270	19.17	0.16

Table 10: Simplification Results on Museum-PT

Model	SARI	BertS	Bleu	% U
Baselines				
MUSS	28.00	.8221	19.48	6.98
Enh-PT-SS	31.84	.8129	17.47	3.98
Open-weight LLMs				
aya23-8b	41.61	.7799	12.37	0.09
gemma2-27b	41.13	.7808	9.25	0
Llama2-7B	36.22	.7282	9.00	3.62
Llama3-8B	34.00	.6989	8.40	5.72
Llama3.1-8B	41.27	.7793	10.29	0.01
Llama3.1-8B-q8	40.60	.7759	8.72	0.00
Llama3.2-3B	37.76	.7501	7.61	0.84
Mistral-7B	40.07	.7892	12.71	0.01
OLMo-7B	38.71	.7630	11.54	1.17
Phi-3-medium	39.22	.7693	8.30	0
Phi3.5-mini	37.25	.7133	4.77	0.41
Qwen2-7B	41.85	.7969	13.85	0.01
Qwen2-72B	41.19	.7818	9.34	0
Qwen2.5-7B	43.50	.7980	16.34	0.15
Qwen2.5-7B-Q8	43.54	.7998	15.98	0.09
Qwen2.5-14B	42.86	.7844	13.72	0
Qwen2.5-32B	**44.05**	**.8021**	14.98	0
r1-distill-7b	38.63	.7783	13.60	2.15
r1-distill-14b	40.28	.6958	10.64	0.27
r1-distill-32b	43.91	.8019	15.37	0.04
Closed-weight LLMs				
Command-R	44.35	.7924	11.77	0
Gpt-4o-m	45.14	.8155	17.44	0.01
o1-mini	**45.24**	.7808	17.91	0
Sabia-2-S	44.29	**.8172**	17.40	0.31
Sabia-3	42.56	.7889	11.99	0.01

Table 11: Simplification Results on Gov-Lang-BR

Category	Style	Simplification
Syntactic	Original	Conforme moradores do bairro, a expressão identificaria um grupo de pichadores.
	Simplified	Os moradores do bairro dizem que a frase identificaria um grupo de pichadores.
	Original	According to neighborhood residents, the expression would identify a group of graffiti taggers.
	Simplified	The neighborhood residents say that the phrase would identify a group of graffiti taggers.
Order	Original	Entre os motivos da liderança gaúcha, estão a tradição no cultivo da soja, que hoje representa a maior parte da matéria-prima do biodiesel, e a predominância da agricultura familiar, condição para concessão do selo social.
	Simplified	A tradição na cultura da soja, que hoje representa a maior parte da matéria-prima do biodiesel, e o predomínio da agricultura familiar, condição para conceder o selo social, estão entre os motivos da posição gaúcha de líder.
	Original	Among the reasons for the leadership of Rio Grande do Sul are the tradition in soybean cultivation, which today represents the majority of the raw material for biodiesel, and the predominance of family agriculture, a condition for obtaining the social seal.
	Simplified	The tradition in soybean cultivation, which today represents the majority of the raw material for biodiesel, and the predominance of family agriculture, a condition for granting the social seal, are among the reasons for Rio Grande do Sul's leadership position.
Anaphora	Original	E com eles amarrados a coleiras, do alto de uma duna a cerca de 50 metros do mar, tomava chimarrão às 19h de ontem.
	Simplified	Pandolfo tomava chimarrão às 19h de ontem, no alto de um monte de areia, com os poodles amarrados a coleiras.
	Original	And with them tied to leashes, from the top of a dune about 50 meters from the sea, he drank mate at 7 p.m. yesterday.
	Simplified	Pandolfo was drinking mate at 7 p.m. yesterday, atop a sand dune, with the poodles tied to leashes.
Lexical redundancy	Original	Numa entrevista coletiva conduzida ontem à noite, os gerentes da Nasa deram o veredicto.
	Simplified	Numa entrevista coletiva ontem à noite, os gerentes da Nasa decidiram.
	Original	In a press conference conducted last night, NASA managers delivered the verdict.
	Simplified	In a press conference last night, NASA managers made a decision.

Table 12: Selected simplifications used as exemplars, one for each one-shot demonstration, together with their English versions. Note that the translations might not fully express the simplification if they were done in the original translated sentence.