Simplifying Administrative Texts for Plain Language using LLM: a Model Comparative Analysis

João Pedro Holanda[1], **Regis Magalhães** [1], **Camilo Almendra** [2], **Luis Gustavo Coutinho do Rêgo**[3]

[1]Universidade Federal do Ceará (UFC) – Campus Quixadá
Quixadá – CE – Brazil

[2]Departamento de Computação
Universidade Federal do Ceará (UFC) – Fortaleza - CE - Brazil

[3]Instituto Federal do Ceará (IFCE) – Campus Quixadá
Quixadá – CE – Brazil

joaopedroph@alu.ufc.br, {regismagalhaes,camilo.almendra}@ufc.br,

gustavo.coutinho@ifce.edu.br

Abstract. Research Context: *Plain language is a growing tendency to create more inclusive and readable documents. Open and closed LLMs can be used to simplify documents, with varying costs and tradeoffs.* ***Scientific and/or Practical Problem:*** *Currently, the UFC Inova institution manually writes simplified versions of their public notices, causing longer publication time and repetitive effort.* ***Proposed Solution and/or Analysis:*** *We design a LLM-based pipeline that generates simplified versions of public notices with zero-shot prompting following plain language directives. We investigate how open LLMs compare to proprietary models. Although proprietary solutions are the cutting-edge models, adoption of open LLMs lowers the cost of ownership and avoids vendor lock-in, making them a sustainable choice for public-sector universities.* ***Related IS Theory:*** *We frame this work as a Socio-Technical Systems intervention to enhance access to public information, grounding it in prior research on text complexity and plain-language communication.* ***Research Method:*** *We evaluated a text simplification pipeline applied against different LLMs. The original and AI-generated versions were compared using statistical readability indexes and morphosyntactic metrics. Afterwards, two pairs of documents were evaluated in a survey with readers' representatives.* ***Summary of Results:*** *The quantitative evaluations indicate that Gemini Flash outperformed the Pro version on both set of metrics, and the Qwen2.5:14b open model was closest to both in the morphosyntactic aspect. Regarding the qualitative evaluation, we observed that automated simplification was well received, but it may better support readers when combined with summarization.* ***Contributions and Impact to IS area:*** *This work provides a process to foster the adoption of plain language in the public sector, along with empirical evidence on the effectiveness of open LLMs compared to proprietary models.*

1. Introduction

Plain language is an international movement started in 1970 with the purpose of formalizing a way of writing accessible texts that transmit information in a simple and objective

way to their targeted audience, and is being adopted by many institutions across the globe, including Brazilian universities [Martins et al. 2023, de Sousa et al. 2024].

Publishing documents in plain language can reduce the need for support, reach more people and improve public time utilization and satisfaction [Cuesta et al. 2019]. But it also requires extra work and resources to train redactors on writing plain-language, write, review and publish these documents. These problems can be tackled by automatizing the process to reduce the burden on human workers.

Text simplification is a well-known NLP task that fits the objectives of plain language by performing lexical, syntactic, and structural changes on the text [Shardlow 2014]. As most of the common NLP tasks, LLMs are capable of generating text simplifications [Färber et al. 2025], although ethical issues must be considered before releasing any LLM-based system for users [Freyer et al. 2024].

However, the evaluation of generative AI-based system is challenging and many metrics and approaches have been recently proposed [Yu et al. 2025]. Three main categories of evaluation approaches emerged in their work: human, automated, and AI-assisted evaluation. In this work, we approached the problem of simplification of public notices using automated and human measurements.

In this work, we propose to perform the automated simplification of administrative texts from UFC Inova using LLMs. We address three research questions that connect text simplification with AI and plain language proposals:

RQ1 To what extent do LLM-based text simplifications improve the readability of the source documents?

RQ2 Do open-source LLMs achieve text-simplification performance comparable to proprietary models?

RQ3 What are readers' perceptions of the benefits and risks associated with the automated simplification of public notices?

The quantitative aspects considered in **RQ1** include the frequency of syntactic patterns that are considered more complex for readers, the overall size of the sentences in the text, and the proportion of complex words present in the text. To address **RQ2**, we evaluated the efficiency of text simplification performed by different LLMs in a zero-shot prompting workflow. To measure the quantitative results, we used consolidated statistical indexes jointly with linguistic metrics from the NILC-Metrix project [Leal et al. 2024]. Lastly, we addressed **RQ3** by means of a qualitative study with members of the Federal University of Ceará (UFC) using an online questionnaire.

1.1. Plain Language

This is a solid movement that involves multiple institutions around the world, such as the International Plain Language Federation [1], the Center for Plain Language [2], Clarity [3], and the Plain Language Network (PLAIN) [4]. There is an published ISO standard that provides governing principles and guidelines for developing plain language documents [International Organization for Standardization 2023], and another

[1] https://www.iplfederation.org/our-work/

[2] https://centerforplainlanguage.org/

[3] https://www.clarity-international.net/

[4] https://plainlanguagenetwork.org/

that add guidance and techniques to help authors of legal and administrative documents [International Organization for Standardization 2025].

Currently, more than 20 countries in Europe, North and South America (including Brazil) already have official or federated regulations for writing public documents in plain language [Martins et al. 2023]. There are several Brazilian institutions that provide guidelines for document writing in plain language [de Sousa et al. 2024, UNICAMP 2024, UEG 2023, IFPE 2023, IFMT 2021, CGE 2021], but the current state of the art still lacks automated tools to simplify existing corpora according to plain language principles.

2. Related Works

Recent research on text simplification spans different domains, such as healthcare [Picton et al. 2025, Swanson et al. 2024] and educational [Hartmann and Aluísio 2020, Day et al. 2025], and also include multi-genre datasets curated for text simplification evaluation [Farajidizaji et al. 2024]. However, these studies either addressed only a single aspect of readability, such as lexical choices, or did not employ morphosyntactic metrics that indicate discourse aspects related to readability. Most of these LLM-based studies also considered only the largest proprietary and open models available at the time.

Regarding the task of verifying plain language adherence in text content, the work of [Silveira et al. 2024] considered four different readability indexes, that are also used in this work, to classify the complexity of existing texts in the Brazilian educational system, and compared them with the already labeled grade-level of the texts. But only the statistical aspects captured by these indexes, such as number of sentences complex words, etc., were used to measure complexity.

Text simplification has also been addressed along with LLMs and plain language rules in the work of [Färber et al. 2025], where a system for text simplification according to plain language with different outputs for specific audiences is presented. The generation of simplified texts occurred in a zero-shot learning format. Results were evaluated using BLEU, SARI, Flesch Reading Ease, and Flesch-Kincaid Grade level, and the GPT-4o model was compared to Llama 3.1. However, the PLAIN language constraints in the prompt concerned only sentence length, and morphosyntactic aspects of the texts were not considered in their evaluation.

Table 1 summarizes the characteristics of each related work.

Table 1. Comparison of Most Related Works.

Work	Readability Indexes	Semantic Similarity	Human Evaluation	Text Statistics	Adheres to Plain language	Considers small models ($<$ 14b parameters)
[Picton et al. 2025]	✓	✓	×	×	×	×
[Day et al. 2025]	✓	✓	✓	✓	×	×
[Swanson et al. 2024]	✓	×	✓	✓	×	✓
[Färber et al. 2025]	✓	×	×	×	✓	×
[Farajidizaji et al. 2024]	✓	✓	×	✓	×	✓
[Silveira et al. 2024]	✓	×	×	×	✓	–
This work	✓	×	✓	×	✓	✓

3. Methodology

Our methodology consisted of the followed steps, detailed in the following s 1. Extraction of notices published online and conversion from PDF to markdown text; 2. Text simplifi-

cation with generative AI. 3. Evaluation with readability indexes, morphosyntactic metrics and a questionnaire with academics from UFC.

3.1. Dataset

We selected 10 notices published in 2025 by UFC Inova[5] to be used in our evaluation; three of them were a summarized version, and one was a ratification of a previous notices. All are directed towards students of the university.

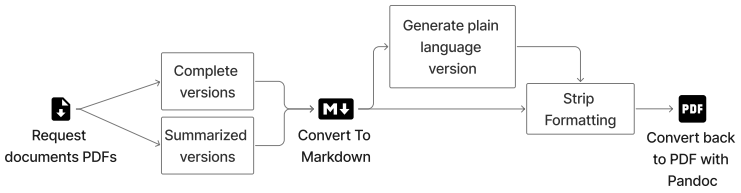

Figure 1. Steps for generating PDFs.

After extracting the PDF files to our dataset, we transformed these documents into Markdown and performed an initial cleaning process as described in [Holanda et al. 2026]. These resulting files were used as input for the prompts sent to the LLMs.

3.2. Text Simplification

All language models received a fixed prompt and a chunk of up to 4,000 characters from the formatted text created with Langchain's default separators with no overlap; for the sake of replicability, we set the response temperature to 0. The resulting chunks were put together in a single Markdown text file. The used prompt is presented in [Holanda et al. 2026].

The models were accessed through Open WebUI and hosted in a server with the following hardware specifications:

- **RAM:** 128 GB DDR4
- **Processing:** Intel Core I9-2900F 12th 24 cores
- **GPU**: Nvidia Geforce RTX 3080 Ti 12228 MB

After generating the AI text simplifications, all files were converted from Markdown to PDF using Pandoc [6] as shown in Figure 1; these resulting PDFs were then presented in our questionnaire.

3.3. Evaluating Complexity

To evaluate how complex a text is, we combined statistical indexes, such as Flesch reading ease [Flesch 1948], with a set of morphosyntactic metrics chosen from the NILC-Metrix project [Leal et al. 2024]. We combined both sets of metrics, since these indexes reveal structural problems, such as long paragraphs or sentences, but do not address the syntactic

[5]https://ufcinova.ufc.br/pt/2025-2/
[6]https://github.com/jgm/pandoc

structure, whereas morphosyntactic metrics are good indicators of complex structures in the text content.

Figure 2 explains the metric extraction process, where the Markdown files for complete and AI-generated texts underwent another text-cleaning process described in [Holanda et al. 2026]. In the next sections, we present all readability indexes and morphosyntactic metrics used in this work.

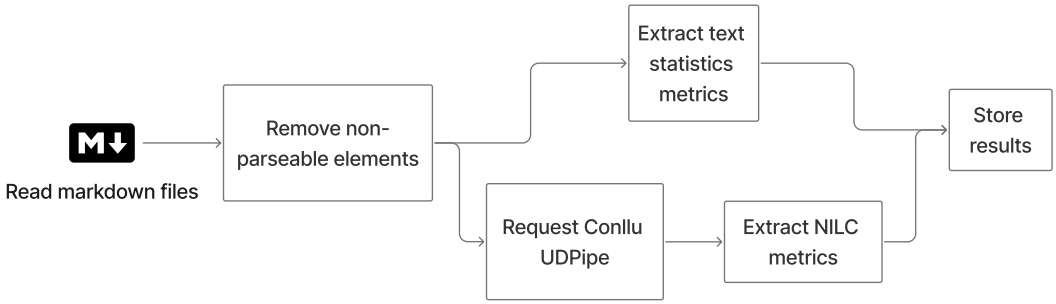

Figure 2. Steps for metric extraction.

3.3.1. Readability Indexes

The most basic metrics used to evaluate text complexity are based on statistical aspects of a text, such as number of words, sentences, syllables, and the number of complex words (with varying definitions of what constitutes a complex word) [Moreno et al. 2023]. They are a reliable source of information about the text structure, despite the limitation of considering only a few quantitative aspects that impact complexity [Leal et al. 2024].

In this work, we selected two indexes that measure complexity on a scale of 0-100, with 100 being the most readable: **Flesch reading ease (FR)** and **Gulpease index (GI)**, and four that measure grade level, or years of education required to understand a text: **Flesch-Kincaid grade level (FK)**, **Gunning Fog index (GF)**, **Automated Readability Index (ARI)**, and **Coleman-Liau index (CL)**. These indexes used their constant values adapted to Portuguese (except for the Gulpease index, which had not needed adaptation) as described in [Moreno et al. 2023].

In our implementation of Gunning Fog Index, we considered a word to be complex if its lemma was not in the top 5,000 word occurrences (excluding numbers and punctuation) in all of Linguateca's corpora [7]. Syllables were identified using the Pyphen Python library [8].

[7]https://www.linguateca.pt/acesso/tokens/formas.todos.txt
[8]https://doc.courtbouillon.org/pyphen/stable/

3.3.2. Morphosyntactic Metrics

To delve into causes of text complexity that are not covered by the statistical indexes, but highlighted by Plain Language guides, we chose a subset of 7 metrics from the NILC-Metrix project [Leal et al. 2024]. These metrics were selected based on plain language recommendations presented in [UNICAMP 2024, Plain Language Association International (PLAIN) , UEG 2023, IFPE 2023, Almeida et al. 2024, IFMT 2021, ÍRIS 2022] and their names are described alongside their categories and relation with Plain Language in Table 2.

Table 2. Selected metrics with their respective groups.

Category	Metric	Plain Language Principles	Effect on readability
Syntactical Complexity	non_svo_ratio	Direct word order	Negative
Syntactical Complexity	passive_voice_ratio	Active voice instead of passive	Negative
Syntactical Complexity	words_before_main_verb	Short sentences Direct word order	Negative
Morphosyntactic word information	personal_pronoun_ratio	Avoiding Impersonality	Positive
Referential cohesion	coreference_pronoun_ratio demonstrative_pronoun_ratio	One idea per phrase	Negative Negative
Text easability	long_sentence_ratio	Short sentences	Negative
Lexical Diversity [9]	foreign_word_ratio	Avoid foreign words	Negative

We also propose a new metric in this work: the *foreign_word_ratio*, which evaluates the proportion of unique foreign types compared to all types in the text, using the Universal Dependencies (UD) feature "Foreign" [10]. It was created based on the same recommendations as the previous metrics.

long_sentence_ratio describes how many sentences have an excessive number of words, in [Leal et al. 2024] sentences with more than 15 words were considered long, but we chose 20 words as the threshold based in the works of [UNICAMP 2024, Plain Language Association International (PLAIN)].

The metric *personal_pronoun_ratio* counts the ratio of personal pronouns in all words of the text, and is relevant for plain language because texts that have more direct communication with the reader are desirable [Almeida et al. 2024] . We considered that more personal pronouns were a positive factor.

Regarding the order of the phrase elements, *non_svo_ratio* describes the number of sentences that don't follow the subject-verb-object word order used in Portuguese and English, the more sentences that don't follow this order the harder it is to understand a text.

passive_voice_ratio measures the ratio of sentences which used passive voice, these sentences are harder to understand and should be avoided, according to the plain language guides mentioned before.

As for the *coreference_pronoun_ratio* and *demonstrative_pronoun_ratio*, they describe the number of possible references made in a sentence based in the words of the

[9] Since foreign_word_ratio is not included on NILC-Metrix, we selected lexical diversity as the best fitting category.

[10] https://universaldependencies.org/u/feat/Foreign.html

previous one, if a pronoun can refer to many different words from a previous sentence, it gets harder to resolve which one is the correct, thus making the sentence less readable. We used the mean of the possible references found between all pairs of sentences that have subject pronouns, for the first metric, or demonstrative pronouns, for the second.

Lastly, *words_before_main_verb* metric represents the average number of words before the main verb of each sentence in the text, the more words it has, the harder it is to interpret.

The morphosyntactic metrics were evaluated using the UDPipe 2.0 framework [Straka 2018], trained with the Porttinari treebank [Pardo et al. 2021], to perform tokenization, tagging, and dependency parsing on the target texts.

Some of the selected metrics were implemented to use UD's part of speech tags and features. Making the evaluation more tool independent, as the metric implementation can be used with every model compatible with UD.

3.4. Academics Evaluation

We conducted a study to evaluate the automated simplification of documents from the point of view of academics, with our target population consisting mainly of faculty and undergraduate students of the Federal University of Ceará (UFC) who read or submit public notices at least once per year. In this section, we detail the study methodology and the results.

This study with academics is classified as survey research [Easterbrook et al. 2008] based on a self-administered questionnaire [Kitchenham and Pfleeger 2008]. The objective was to gather their perceptions of the suitability of the AI-generated simplified documents regarding readability and usefulness. We defined the following Survey Research Questions (SRQ) to guide the evaluation:

SRQ1 How easy is it to understand the simplified document, compared to the original one?
SRQ2 How useful is the automatically generated version to support a quick comprehension of the main information?
SRQ3 Is the simplified document ready for publication as a complement to the original version?
SRQ4 What aspects of both simplified versions can improve the understanding by their target public?
SRQ5 Have you identified any risk or limitation in the content of these simplified versions?

Our study instrument comprised reading tasks and opinion questions. The reading tasks provide participants with a pair consisting of the original and a simplified version of the same document. Participants are asked to rate the level of readability, usability, and readiness by comparing the simplified version to the original. Then, they shared their judgment of the overall solution in the opinion questions.

We selected two notices to be used in our evaluation: the longest and the shortest non-ratification for contrast. Each of these document pairs was presented in a different

reading task, where the simplified version came from the open model Phi4, which had the best result on the readability indexes. For each task, we collected the opinion of the participants regarding the questions **SRQ1**, **SRQ2**, and **SRQ3**. **SRQ4** and **SRQ5** came after and asked for opinions considering both notice pairs.

The survey participation was individual and anonymous. We chose a non-probabilistic sampling, which refers to any approach in which participants are not randomly selected [Linäker et al. 2015]. We used convenience sampling as a strategy with our collaboration network. We sent an invitation by instant messaging to our contacts, explaining the evaluation objectives and how to perform it. The complete instrument distributed to participants is available at [Holanda et al. 2026].

4. Results

In this section, we present the results of the statistical indexes, morphosyntactic metrics, and, lastly, the participants' answers to the questionnaire. The evaluation of quantitative results was made with a comparison of the confidence intervals of the results for all documents per model. The raw data is also present in [Holanda et al. 2026].

4.1. RQ1 – To what extent do LLM-based text simplifications improve the readability of the source documents?

Some models hallucinated during the text generation, creating either unintelligible texts or repeating the question prompt inside the response, or containing more than 3 paragraphs in English. We reviewed all generated texts manually and included the number of hallucinated documents per model in the Table 3.

These hallucinations impacted the evaluation as they introduced outliers, in the case of unrelated responses, much shorter than the original text, giving false positives, and responses in English, which negatively affected Gunning Fog and the morphosyntactic metrics.

Table 3. Number of documents with hallucinations per model.

Model	Number of documents with hallucinations
phi3:3.8b	7
granite3-dense:8b	3
granite3-dense:2b	1
qwen2.5:14b	1

4.1.1. Readability Indexes

Table 4 shows the mean values of each metric across all documents in each different model, with a p value of 0.05, where Gemini Flash and Pro had better values than the original measures in all metrics. It is also important to note that, surprisingly, Gemini Flash had better results than Pro in all of the metrics.

The best open models were Phi4 and Qwen2.5, with both outperforming the original results in the Gulpease and Gunning Fog indexes; Phi4 also gave a better result in

Table 4. Model comparison by mean readability values and confidence intervals across all documents.

Model	Flesch Ease↑	Gulpease↑	Flesch-Kincaid↓	ARI↓	Gunning Fog↓	Coleman Liau↓
Original	51.54 (47.43 - 55.66)	49.77 (47.00 - 52.54)	10.54 (9.54 - 11.55)	14.08 (12.98 - 15.18)	10.78 (9.58 - 11.98)	16.39 (15.75 - 17.02)
gemma2_tools	52.54 (47.95 - 57.13)	51.05 (47.98 - 54.13)	10.17 (9.17 - 11.17)	13.59 (12.66 - 14.52)	10.30 (9.45 - 11.15)	16.29 (15.69 - 16.89)
deepseek-r1:14b	56.11 (51.01 - 61.22)	53.91 (50.35 - 57.47)	9.02 (8.01 - 10.03)	13.43 (12.52 - 14.34)	9.31 (8.66 - 9.95)	17.29 (16.44 - 18.13)
gemini-2.5-flash	**74.33 (72.22 - 76.45)**	**62.02** (60.40 - 63.63)	**6.25** (5.89 - 6.61)	**9.71** (9.23 - 10.19)	**7.46** (7.11 - 7.81)	**13.22** (12.74 - 13.71)
gemini-2.5-pro	**68.35 (66.16 - 70.54)**	**60.71** (58.30 - 63.12)	**7.07** (6.68 - 7.46)	**10.48** (9.87 - 11.09)	**7.49** (7.01 - 7.97)	**14.19** (13.66 - 14.71)
gemma3:4b	53.02 (48.05 - 57.99)	51.16 (47.73 - 54.60)	10.32 (9.08 - 11.55)	13.52 (12.12 - 14.92)	10.45 (9.32 - 11.58)	15.75 (14.92 - 16.58)
granite3-dense:2b	50.69 (45.96 - 55.43)	48.93 (46.38 - 51.47)	10.75 (9.68 - 11.82)	14.30 (13.31 - 15.29)	11.17 (9.98 - 12.37)	16.49 (16.00 - 16.99)
granite3-dense:8b	50.86 (47.06 - 54.67)	49.35 (46.81 - 51.89)	10.76 (9.82 - 11.71)	14.13 (13.12 - 15.14)	11.21 (10.04 - 12.38)	16.21 (15.48 - 16.93)
llama3.2:3.2b	51.77 (48.21 - 55.33)	50.68 (48.62 - 52.75)	10.28 (9.49 - 11.06)	13.69 (12.96 - 14.42)	10.25 (9.48 - 11.02)	16.42 (15.85 - 16.99)
phi3:3.8b	**76.85 (57.56 - 96.15)**	57.35 (47.30 - 67.39)	6.32 (2.59 - 10.06)	14.18 (12.36 - 16.00)	12.40 (11.08 - 13.71)	17.17 (16.26 - 18.08)
phi4:13.7b	57.72 (54.25 - 61.18)	**56.60 (53.84 - 59.35)**	**8.65 (7.94 - 9.35)**	12.47 (11.86 - 13.08)	**8.58 (8.10 - 9.06)**	16.42 (15.82 - 17.02)
qwen2.5:14b	56.53 (51.75 - 61.31)	**57.45 (53.62 - 61.27)**	8.72 (7.89 - 9.55)	12.55 (11.87 - 13.23)	**8.75 (8.17 - 9.33)**	16.65 (15.93 - 17.36)

Mean value across all documents for each readability metric with confidence of 95%.
Metrics with better intervals than the original values are marked bold.
The arrow shows whether the more readable results are represented by higher (up arrow) or lower (down arrow) values.

the Flesch-Kincaid metric. Lastly, Phi3 had the best results for the Flesch reading ease, but this can be related to the previously cited hallucinations, since the variation in the text size was not due to an optimal simplification.

The remaining models results overlapped the original values in all the metrics, but most results had a slight improvement in the mean value. The only degradations in each metric occurred in these models:

- **Flesch Ease**: Granite3-dense (2b and 8b).
- **Gulpease Index**: Granite3-dense (2b and 8b).
- **Flesch-Kincaid**: Granite3-dense (2b and 8b).
- **ARI**: Granite3-dense (2b and 8b).
- **Gunning Fog**: Granite3-dense (2b and 8b) and Phi3.
- **Coleman-Liau**: Deepseek-r1, Granite3-dense:2b, Llama3.2, Phi3, Phi4 and Qwen2.5.

It's clear from this listing that both Granite3 models had the worst results overall, with Granite3-dense:2b being worse than the original in all metrics and Granite3-dense:8b performing worse in 5. It is also worth remark that even the best performing models showed a slight degradation in the Coleman-Liau Index.

4.1.2. Morphosyntactic metrics

Table 5 shows the mean value across all documents for each model and the original measures, with a p value of 0.05. The Gemini models were the best performing models in 4

out of the 8 metrics, but this time, Gemini Pro surpassed Gemini Flash in two metrics, passive voice ratio and long sentence ratio.

Although the personal pronoun ratio was considered as directly related to a higher text complexity on NILC-Metrix, some plain language guides consider a text with more personal communication elements to be more accessible to readers [UNICAMP 2024]. Thus, we considered it a positive metric that improved readability. shows that the Gemini models had the highest introduction of personal pronouns, followed by Phi3, which hallucinated in at least half of the documents.

It is also worth noting that the changes in the *foreign_word_ratio* were the lowest across all metrics (0.004 was the mode), since the original documents did not have a large number of distinct foreign words from the start. Most models slightly reduced the ratio, but Phi3 had a large degradation, with a mean over 50 times bigger than the mode. This occurred primarily because of the hallucinations described in 4.1. Overall, no model was significantly better than the original.

Table 5. Model comparison by mean morphosyntactic proportions and confidence intervals across all documents.

Model	CP ↓	DP ↓	NS ↓	PV ↓	WB ↓	LS ↓	FW ↓	PP ↑
Original	.700 (-.471 – 1.871)	1.049 (.679 – 1.420)	.052 (.034 – .070)	.213 (.158 – .268)	5.118 (4.104 – 6.132)	.283 (.190 – .376)	.004 (.001 – .006)	.003 (.002 – .004)
gemma2 _tools	1.400 (.085 – 2.715)	1.179 (.713 – 1.645)	.040 (.021 – .059)	.200 (.150 – .251)	4.809 (3.928 – 5.690)	.263 (.195 – .330)	.003 (.000 – .006)	.004 (.003 – .005)
deepseek-r1:14b	.633 (.021 – 1.246)	.766 (.226 – 1.306)	.027 (.005 – .050)	.153 (.104 – .201)	**3.288 (2.498 – 4.079)**	**.130 (.078 – .182)**	.001 (.000 – .003)	.002 (.001 – .003)
gemini-2.5-flash	1.116 (.861 – 1.372)	.925 (.663 – 1.188)	.071 (.054 – .089)	**.089 (.071 – .107)**	**3.227 (2.837 – 3.618)**	**.091 (.071 – .112)**	.002 (.001 – .002)	**.019 (.015 – .022)**
gemini-2.5-pro	1.350 (1.146 – 1.554)	1.027 (.861 – 1.193)	.041 (.030 – .052)	**.082 (.070 – .093)**	**3.250 (2.945 – 3.555)**	**.078 (.058 – .098)**	.002 (.000 – .004)	**.012 (.008 – .016)**
gemma3:4b	1.125 (.268 – 1.982)	1.126 (.747 – 1.504)	.077 (.053 – .102)	.216 (.169 – .263)	5.327 (4.257 – 6.397)	.278 (.173 – .383)	.002 (.000 – .003)	.004 (.002 – .006)
granite3-dense:2b	.242 (-.043 – .526)	1.417 (.863 – 1.970)	.045 (.030 – .060)	.228 (.178 – .278)	4.553 (3.354 – 5.753)	.301 (.232 – .370)	.002 (.000 – .005)	.002 (.001 – .003)
granite3-dense:8b	1.150 (-.135 – 2.435)	1.255 (.838 – 1.673)	.058 (.035 – .080)	.224 (.168 – .281)	4.435 (3.415 – 5.455)	.291 (.218 – .364)	.007 (-.002 – .015)	.002 (.001 – .003)
llama3.2:3.2b	.350 (-.185 – .885)	1.135 (.615 – 1.655)	.049 (.036 – .061)	.204 (.168 – .240)	4.650 (3.654 – 5.647)	.261 (.198 – .325)	.002 (.000 – .004)	.004 (.002 – .006)
phi3:3.8b	.425 (-.281 – 1.131)	.471 (.030 – .911)	.036 (.008 – .064)	.108 (.010 – .205)	3.315 (1.768 – 4.862)	.254 (.081 – .427)	.293 (.147 – .439)	.005 (.002 – .007)
phi4:13.7b	.902 (.162 – 1.643)	1.046 (.684 – 1.407)	.033 (.021 – .045)	.120 (.091 – .150)	3.300 (2.380 – 4.220)	**.129 (.085 – .172)**	.002 (.001 – .003)	.004 (.003 – .006)
qwen2.5:14b	.325 (-.209 – .859)	.468 (.087 – .849)	.028 (.013 – .044)	.139 (.096 – .183)	**2.708 (2.163 – 3.253)**	**.097 (.065 – .128)**	.002 (.001 – .003)	.003 (.002 – .005)

Mean value across all documents for each proportional morphosyntactic metric with confidence of 95%.
Bold values represent intervals that are better than the original.
Arrows show whether the less complex results are associated with higher (up arrow) or lower (down arrow) values.
The red text highlights the hallucination effect in Phi3, responsible for a big anomaly in the FW metric.
CP = Coreference Pronoun Ratio; DP = Demonstrative Pronoun Ratio; NS = Non SVO Ratio; PV = Passive Voice Ratio; WB = Words before main verb; LS = Long Sentence Ratio; FW = Foreign Word Ratio; PP = Personal Pronoun Ratio.

4.1.3. Performance of Foreign Word Ratio

Regarding the performance of the proposed metric, UDPipe showed limitations in the detection of foreign expressions inside the text, for example, the original document with the highest foreign word ratio (0.0106) had 28 foreign types, but only 8 were recognized.

Some common loan words from English were rarely detected. “Link” appeared in five documents, but was recognized in only one of them. The occurrences of foreign words were annotated manually by the authors and considered only words that were either

present in the Cambridge Dictionary[11] or loaned from Latin, ignoring neologisms and entity names.

Table 6 presents the quality of UDPipe predictions for each original document in the considered dataset. No foreign word was recognized in three out of 10 documents, which reflects a recall, precision, and F1 score of 0. In the case of the document "Prointer/UFC Nº 03/2025 - PIBI", it did not have any foreign words, so the zeroed F1 score is justified by the 100% accuracy.

Overall, the low recall values ($<= 33,33\%$) indicate that many foreign words occurred as false negatives. The precision measures were better in the non-zero cases, with only one value lower than 100%. We can conclude that the identification of foreign words was incomplete, but in almost all documents was not incorrect.

Table 6. F1 score, precision, recall and accuracy of the foreign word ratio measured with UDPipe on each document.

Document	F1 Score (%)	Precision (%)	Accuracy (%)	Recall (%)
"Edital Prointer/UFC Nº 09/2025 - PROMISAES"	66,67	100,00	99,56	33,33
"Hackathon Inovan'do UFC"	54,55	100,00	99,07	25,00
"Edital Prointer PIBI Nº 04/2025 - Empreende UFC - Simplified"	44,44	100,00	96,81	25,00
"Edital Prointer PIBI Nº 04/2025 - Empreende UFC"	42,86	66,67	98,74	28,57
"Conjunto UFC - TRE/CE Nº 01/2025" - Simplified	36,36	100,00	96,57	21,05
"Conjunto UFC - TRE/CE Nº 01/2025"	9,52	100,00	98,77	5,00
"Prointer/UFC Nº 03/2025 - PIBI"	0,00	0,00	100,00	0,00
"Prointer/UFC Nº 08/2025 - Bolsas de Internacionalização"	0,00	0,00	99,17	0,00
"Edital Prointer/UFC Nº 02/2025 - UFC Mundo" - Ratification	0,00	0,00	99,38	0,00
"Edital Prointer/UFC Nº 02/2025 - UFC Mundo"	0,00	0,00	98,90	0,00
Mean	25,44	56,67	98,967	13,79

Indefinite values were taken as 0.
The "Prointer/UFC Nº 03/2025 - PIBI" notice did not have any foreign word.

4.2. RQ2 – Do open-source LLMs achieve text-simplification performance comparable to proprietary models?

Considering solely the readability indexes, Gemini 2.5 Flash and Pro were the models with the best performance overall, measuring values better than the original, followed by Phi4, Qwen2.5:14b, and Phi3. The Phi3 model achieved better results than both Gemini models in FR, and better than Pro in FK, but was inferior in the remaining.

The remaining models had smaller variations in their intervals, with six of them having overlaps with the original measured interval in all readability indexes and five of

[11] https://dictionary.cambridge.org/us/dictionary/

them being in the same interval as the original values for all morphosyntactic metrics, indicating that the text simplification did not achieve big changes in the text content.

On the morphosyntactic aspect, Gemini also got the best results, with Qwen2.5 and Deepseek coming after. The correrefence metrics (CP and DP) along with Non-SVO did not have any significant change in the interval range in all models, but the mean of the CP was higher in the two Gemini models, which can be related to the increase of the personal pronouns, where these models had the best results, showing that exists some conflict in these two metrics: a more personal communication with the reader makes the text more accessible, but can also makes the text more complex if there is the need to use references in multiple places of the text.

Regardless of the interval overlaps, most models had some improvement on the mean value; the only models that presented the worst mean values for each metric were:

- **CP**: Gemma2_tools, Gemini 2.5 Flash and Pro (as discussed on the previous paragraph), Gemma3, Granite3-dense:8b and Phi3.
- **DP**: Gemma2_tools, Gemma3, Granite3-dense (2b and 8b) and Lamma3.2.
- **NS**: Gemini 2.5 Flash, Gemma3, Granite3-dense:8b.
- **PV**: Gemma3 and Granite3-dense (2b and 8b).
- **WB**: Gemma3.
- **LS**: Granite3-dense (2b and 8b).
- **FW**: Granite3-dense:8b and Phi3.
- **PP**: Deepseek-r1, Granite3-dense:8b and Phi3.

So the metric with the most negatives changes where CP, degrading the observed values in five out of 11 models. And the models that had the most deterioration were Gemma3 and the two Granite3-dense models.

4.3. RQ3 – What are readers' perceptions of the benefits and risks associated with the automated simplification of public notices?

In this section, we present each of the five participants' profiles and discuss the results of the procedure steps, grouped by research questions. Their profiles are depicted in Table 7. Although the sample is not statistically relevant, the opinion of professors that have experience writing public notices and students, who are the target audience, gives a good direction of what the public perceives as a good simplification.

Table 7. Participants' profile.

Question	P1	P2	P3	P4	P5
Affiliation with the university	Faculty	Faculty	Student	Faculty	Student
How often do you read public notices?	Always	Always	Rarely	Occasionally	Rarely
Experience with producing, reviewing, or disseminating public notices?	Frequently	Very Frequently	Rarely Participate	Never participated	Participate occasionally

The answers for **SRQ1** about the improvement in readability of the simplified version were mostly positive. On the first notice, four of the five participants marked

'Slightly clearer' and one marked 'Much clearer'. Whereas the second notice had two 'Much clearer' answers, two 'Slightly clearer', and one 'No significant difference'. So, we observed that the main goal of the automated simplification was perceived by the participants in both notices, although one opinion diverged in the second notice.

On **SRQ2**, regarding the usefulness of quickly comprehending the main information, P3 agreed totally, P1 and P4 agreed partially, and P5 kept a neutral opinion for both notices. P2 disagreed partially on the first notice and was neutral about the second.

Regarding the readiness of the notice at **SRQ3**: P3 and P5 considered both notices as ready. P4 answered that the first notice was good as a complement, but not as a replacement for the original, and the second version simplified the most important information, but the complete version would still be necessary for applicants. P1 disagreed on both notices: the first could be even simpler and visually attractive; the second also, but it additionally had one sentence that contained wrong information. Lastly, P2 also disagreed with the statement for both notices, commenting that the first notice had no graphic elements for the most important information and lacked a fitting topic structure. Moreover, this participant could not differentiate between the simplified and original versions of the second notice.

For **SRQ4**, P3, P4, and P5 highlighted these points about the simplified versions, respectively: quicker understanding; more objective text, and thus tends to be smaller and easier to read; smaller texts have less chance of misinterpretation. P1 answered that a simplification was perceived, but it could be better. P2 suggested structures for a better simplification (beyond the discourse level we are evaluating): usage of graphical elements, clickable links, topic division in the text, and formatting responsiveness to allow sharing in multiple media.

The responses for **SRQ5** about the risks and limitations of simplified versions varied. P3 and P4 answered that they found no risk, but P4 pointed out that some sections had a better simplification than others; P5 answered that certain specific details could be excluded in the AI-generated document; finally, P1 identified an error in the original document after reviewing the simplified version, and P2 pointed out that the first notice had a chance of not being more readable, and the second simplified notice could give the feeling that the original document was not necessary.

Discussion. We observed that although the participants were instructed that the evaluation was focused on the simplification of the documents, there was an expectation that a summarization should also be carried out. This suggests that simplification and summarization should be considered together, as complementary strategies to enhance document clarity and usability. Also, participants raised concerns that the use of plain language does not replace the original public notices. Thus, the target audience should be compelled to read the full version after grasping a general understanding through the simplified version. Finally, the perceived lack of consistency throughout the document may be addressed by refining prompts.

5. Conclusions

We presented an evaluation on text simplification with LLMs in a zero-shot implementation. The acquired results indicate that most open models did not achieve a performance

similar to Gemini Flash and Pro for the readability indexes, but the Qwen2.5:14b model achieved better results over the morphosyntactic metrics, implying that the readability increased, even if the text structure didn't change as much as in Gemini. The high number of hallucinated documents also showed the limitations of certain models in the task of text simplification.

The research survey points out that, even if the metrics results had slight improvements in the mean for six out of 11 models, a summarized version would be more appreciated by the participants, as it acts as a shortcut with the most relevant information instead of replacing the original version, and the improvements on readability were not significant for most of the open models.

Although the foreign word ratio was useful for identifying hallucinations in the Phi3 model, UDPipe had a low precision and recall of foreign words in the documents.

For future work, our pipeline will be modified to perform simplification jointly with summarization. We also aim to explore the combination of prompting strategies in the simplification process, such as few-shot prompting, directional-stimulus, and chain-of-thought, as described in [Day et al. 2025]. We also seek to consider psycholinguistic metrics, such as familiarity and imageability of the words, and plain language directives not covered in the present work, such as the usage of nouns in place of verbs. Lastly, to add the perspective of AI-assisted evaluation for our generative AI pipeline [Yu et al. 2025], we will explore semantic evaluation through the use of embedding cosine similarity between original and AI-generated texts, and LLM-as-a-judge to assess the quality of simplification in a larger number of documents.

References

Almeida, P. C. ., Pozzobon, L. C. ., Figueiredo, J. C. ., Righini, J. C. ., Roedel, P. C. ., Duarte, A. C. ., Costa, L. S. C. ., Quental, C. C. ., Tabak, S. C. ., and Cruz, F. O. (2024). *Simples assim: comunique com todo mundo*. Accessed: Jul 02, 2025.

CGE (2021). Cartilha Como Usar a Linguagem Simples – tornando as comunicações internas e com a sociedade mais fáceis de entender. https://www.cge.ce.gov.br/cartilha-como-usar-a-linguagem-simples/. Accessed: Aug 08, 2025.

Cuesta, A. M., Reyes, A., and Roseth, B. (2019). The Importance of Clarity: Impacts of Colombia's 'Lenguaje Claro' Program on Reducing Administrative Burdens. *IDB Publications*. Publisher: Inter-American Development Bank.

Day, S. L., Cirica, J., Clapp, S. R., Penkova, V., Giroux, A. E., Banta, A., Bordeau, C., Mutteneni, P., and Sawyer, B. D. (2025). Evaluating GenAI for Simplifying Texts for Education: Improving Accuracy and Consistency for Enhanced Readability. http://arxiv.org/abs/2501.09158. Accessed: Dec 04, 2025.

de Sousa, C. M. A. d. O. A., Cardoso, E., and de Andrade, F. D. (2024). DIRETRIZES PARA O USO DE LINGUAGEM SIMPLES: PESQUISA E DESENVOLVIMENTO NO BRASIL E EM PORTUGAL. *Revista da Associação Brasileira de Atividade Motora Adaptada*, 25(2):407–422.

Easterbrook, S., Singer, J., Storey, M., and Damian, D. (2008). *Selecting Empirical Methods for Software Engineering Research*, pages 285–311. Springer, London.

Farajidizaji, A., Raina, V., and Gales, M. (2024). Is It Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *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.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Freyer, N., Kempt, H., and Klöser, L. (2024). Easy-read and large language models: on the ethical dimensions of llm-based text simplification. *Ethics and Information Technology*, 26(3):50.

Färber, M., Aghdam, P., Im, K., Tawfelis, M., and Ghoshal, H. (2025). SimplifyMyText: An LLM-Based System for Inclusive Plain Language Text Simplification. In Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F. M., Pinelli, F., Silvestri, F., and Tonellotto, N., editors, *Advances in Information Retrieval*, pages 418–424, Cham. Springer Nature Switzerland.

Hartmann, N. S. and Aluísio, S. M. (2020). Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental. *Linguamática*, 12(2):3–27. Number: 2.

Holanda, J. P. P., Magalhães, R. P., Almendra, C. C., and do Rêgo, L. G. C. (2026). Simplifying Administrative Texts for Plain Language using LLM: a Comparative Analysis: Results for Morphosyntactic and Readability Indexes on Public Notices and their AI Generated Plain Language Versions. doi.org/10.6084/m9.figshare.29376692.

IFMT (2021). Cartilha Orientativa sobre o Uso de Linguagem Simples no Contexto do Instituto Federal de Mato Grosso. https://ifmt.edu.br/media/filer_public/38/12/38122512-8c1d-43c4-9044-7c99c23d79ff/cartilha_orientativa_eu_uso_versao_final.pdf. Accessed: Aug 08, 2025.

IFPE (2023). Guia para Comunicação Interna e Externa do IFPE. https://drive.google.com/file/d/1OgpnRQaioywzc00NQ1_Yh6iPitXsB88m/view?usp=drive_link&usp=embed_facebook. Accessed: Aug 08, 2025.

International Organization for Standardization (2023). ISO 24495-1 Plain language Part 1: Governing principles and guidelines. Standard ISO 24495-1:2023, International Organization for Standardization (ISO).

International Organization for Standardization (2025). ISO/PRF 24495-2 Plain language Part 2: Legal communication. Standard, International Organization for Standardization (ISO).

Kitchenham, B. A. and Pfleeger, S. L. (2008). Personal Opinion Surveys. In *Guide to Advanced Empirical Software Engineering*, pages 63–92. Springer London, London.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Linäker, J., Sulaman, S. M., Maiani de Mello, R., and Höst, M. (2015). Guidelines for conducting surveys in software engineering. Technical Report 8ac54dbe-b7ac-4244-9c43-0f0d157efa26, Lund University.

Martins, H. T., da Silva, A. R., and Cavalcanti, M. T. (2023). Linguagem Simples: um movimento social por transparência, cidadania e acessibilidade. *Cadernos do Desenvolvimento Fluminense*, (25).

Moreno, G. C. d. L., de Souza, M. P. M., Hein, N., and Hein, A. K. (2023). ALT: UM SOFTWARE PARA ANÁLISE DE LEGIBILIDADE DE TEXTOS EM LÍ NGUA PORTUGUESA. *Policromias - Revista de Estudos do Discurso, Imagem e Som*, 8(1):91–128.

Pardo, T. A. S., Duran, M. S., Lopes, L., Felippo, A. D., Roman, N. T., and Nunes, M. d. G. V. (2021). Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 1–10. SBC.

Picton, B., Andalib, S., Spina, A., Camp, B., Solomon, S. S., Liang, J., Chen, P. M., Chen, J. W., Hsu, F. P., and Oh, M. Y. (2025). Assessing AI Simplification of Medical Texts: Readability and Content Fidelity. *International Journal of Medical Informatics*, 195:105743.

Plain Language Association International (PLAIN). What is plain language? https://plainlanguagenetwork.org/plain-language/what-is-plain-language/. Accessed: May 05, 2025.

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Silveira, V. I. S., Menezes, P. H. C., Silva, M. S., Carmo, F. A., and Lobato, F. M. F. (2024). Classificação de Linguagem Simples: uma abordagem baseada em Leiturabilidade e Legibilidade. In *Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*, pages 99–110. SBC. ISSN: 2763-8723.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Zeman, D. and Hajič, J., editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Swanson, K., He, S., Calvano, J., Chen, D., Telvizian, T., Jiang, L., Chong, P., Schwell, J., Mak, G., and Lee, J. (2024). Biomedical text readability after hypernym substitution with fine-tuned large language models. PLOS Digital Health, 3(4):e0000489. Publisher: Public Library of Science.

UEG (2023). Uso da linguagem simples no âmbito da Universidade Estadual de Goiás. http://www.ueg.br/legislacao/referencia/13259. Acessed: Jun 16, 2025.

UNICAMP (2024). Guia de Linguagem Simples. https://linguagemsimples.unicamp.br/wp-content/uploads/sites/49/2024/08/guia-de-linguagem-simples.pdf. Accessed: May 14, 2025.

Yu, L., Alégroth, E., Chatzipetrou, P., and Gorschek, T. (2025). Measuring the quality of generative ai systems: Mapping metrics to quality characteristics — snowballing literature review. *Information and Software Technology*, 186:107802.

ÍRIS (2022). Guia Íris de Simplificação: Linguagem Simples e Direito Visual. https://irislab.ce.gov.br/wp-content/uploads/2022/03/Guia-%C3%8DRIS-de-Simplifica%C3%A7%C3%A3o_-Linguagem-Simples-e-Direito-Visual.pdf. Accessed: Aug 08, 2025.