A Generative AI approach for creating and validating simplified versions of government documents.

Karine Alves[a*], *Matheus* Silva[b], *Edney* Santos[c], *George* Valença[d], *Kellyton* Brito[e].

[a] Universidade Federal Rural de Pernambuco, karine.vfalves@ufrpe.br, 0009-0004-9429-4317.

[b] Universidade Federal Rural de Pernambuco, matheus.fsilva@ufrpe.br, 0009-0003-8972-0636.

[c] Universidade Federal Rural de Pernambuco, edney.santos@ufrpe.br, 0009-0002-6784-8789.

[d] Universidade Federal Rural de Pernambuco, george.valenca@ufrpe.br, 0000-0001-9375-5354.

[e] Universidade Federal Rural de Pernambuco, kellyton.brito@ufrpe.br, 0000-0002-6883-8657.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 20 May 2025

Abstract. In the context of Brazil's re-democratization and the need for greater transparency in public administration, the 1988 Constitution established the right to access public information. However, the complexity of legal language, particularly in court documents, poses a significant barrier to understanding for the general public, especially given that only about 25% of Brazilians aged 25 or older have completed or are pursuing higher education. This study addresses this issue by leveraging generative AI models to simplify legal texts from the Court of Accounts of Pernambuco into plain language, making them more accessible to individuals with a high school education level. The research evaluates the effectiveness of two Large Language Models (GPT and Gemini) and five prompt techniques (Tree of Thought, COSTAR, Zero Shot, One Shot, and Meta Prompting) in producing simplified versions of 14 preliminary decisions. A total of 140 simplified texts were generated and evaluated using an 18-question questionnaire based on plain language principles, with scores generated by AI models and validated through human review. The results show that Gemini with the Tree of Thought technique achieved the highest average score (67.64), based on responses to the plain language questionnaire, while GPT with the COSTAR technique performed best in preserving essential information and achieving the highest readability scores (Flesch Reading Ease: 55.26). However, omissions of critical information were a common issue across all models, highlighting the need for human oversight. The study also found that GPT outperformed Gemini in evaluation accuracy, with lower Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) compared to human evaluations. Despite this, AI models tended to overestimate readability and comprehension, underscoring the importance of a hybrid approach that combines AI-generated assessments with human review. The findings demonstrate the potential of generative AI to reduce costs and improve accessibility to legal and governmental documents, while also emphasizing the need for further research to address limitations such as omissions, biases, and ethical considerations. This study contributes to the growing body of literature on AI-assisted text simplification and provides a foundation for future work in this area.

Keywords. Generative AI, Large Language Model, Plain Language, E-Government
Research paper, DOI: https://doi.org/10.59490/dgo.2025.968

1. Introduction

Simplifying public documents for better citizen understanding is a global objective (Petelin, 2010). In 1988, the text of the Brazilian Constitution was enacted. In the context of the country's re-democratization, it became necessary to add subparagraph XXXIII to Article 5, aiming to allow any person to access public agency information, with the goal of increasing transparency and citizen participation in public administration—activities that bring various benefits to society(Andersen, 2009). However, it was only in 2011 that the country began to intensify actions within this agenda, joining the Open Government Partnership (OGP) and, in the same year, publishing the Access to Information Law (LAI) — Law No. 12,527 — which regulates the constitutional provision and outlines the procedures and deadlines for accessing information.

Thus, various Brazilian institutions have started publishing their information on open data online portals, making it accessible to the entire society. However, as these institutions operate in different fields, they use linguistic variations with varying levels of complexity. Among those with a high degree of linguistic complexity are the courts, making it difficult for people who are not regularly immersed in the technical context of the judiciary to understand this information (Belém, 2013). Similarly, this challenge affects a large part of the population, given that only about 25% of Brazilians aged 25 or older have completed or are pursuing higher education, while the rest have at most a high school education (Bandeira, 2024).

Therefore, recognizing this issue, some courts have started adopting the use of plain language in legal documents. However, one of the key aspects to consider when drafting a text in plain language is the target audience. If the document is intended for a more technical audience, lexical simplification may not be necessary. On the other hand, if the document is to be used both internally and by different audiences, multiple simplified versions must be created, each tailored to its respective group. As a result, the cost of producing various simplified documents becomes high for institutions.

While efforts to simplify legal texts—through manual rewriting, visual law, or controlled language systems (Belém, 2013; Martínez et al., 2024)—have existed for decades, their scalability remains limited by human resource constraints. Generative AI introduces a paradigm shift: by automating simplification, it enables institutions to produce *multiple tailored versions* of the same document (e.g., for laypersons, students, or professionals) at negligible marginal cost. This scalability aligns with the constitutional mandate of accessibility (*Constituição*, n.d.) while addressing the socioeconomic disparity in education levels (Bandeira, 2024). However, the trade-off between automation and accuracy—particularly omissions and hallucinations (Ray, 2023)—demands rigorous evaluation, as undertaken in this study.

With the launch of ChatGPT in 2022 and the rapid public adoption of the tool, both the public and private sectors saw an opportunity to solve problems using these models. However, generative AIs can produce hallucinations, that is, incorrect information in the final response (Ray, 2023). Therefore, it is necessary for the generated content to go through a revision process. So far, the literature recommends that the public sector perform human reviews before making the content available to the general public (K. Alves et al., 2024) However, the review process is expensive in environments where a lot of content is generated by AI.

In this context, the goal of this work is to simplify and evaluate documents issued by the Court of Accounts of Pernambuco. For this, 2 generative AIs were used with 5 different prompt techniques: Tree of Thought, Zero Shot, One Shot, COSTA and Meta prompting; producing a sample set of simplifications. These simplifications were evaluated by 2 different AI models. Then, the simplifications from the model and prompt combination that received the highest average score were manually reviewed by a researcher to compare with the score given by the AIs. And assess the quality of the evaluations. Thus, both the simplifications and the evaluations were assessed.

The remainder of this paper is organized as follows. Section 2 provides the related works. Section 3 outlines the methodology used to develop the experiment. Section 4 details the results obtained. Section 5 discusses the results, followed by Section 6, presenting concluding remarks and recommendations for future work.

2. Background and related works

For a better contextualization of the study, four important topics will be detailed: Plain Language in Governments, Application of Generative AI in Governments, Prompt Techniques, and finally, a review of related works.

2.1 Plain language in Government

The main element of plain language is the reader, who must be able to use the text to achieve the intended purpose (Petelin, 2010).Therefore, it is necessary for the document to be tailored to the reader, rather than expecting the reader to adapt to the document. In this way, readers can access government information and promote greater transparency in the public sector (Petelin, 2010). Recognizing the benefits, many countries have begun adopting plain language in various public information like Australia (APSC (Australian Public Service Comission), 2023), Belgium (*Heerlijk Helder | Vlaanderen.Be*, n.d.), Canada (*Plain Language, Accessibility, and Inclusive Communications*

- *Privy Council Office - Canada.Ca*, n.d.) and Norway (*Klarspråk - Språkrådet*, n.d.).

Considering the U.S. scenario, in 1978, U.S. President Jimmy Carter issued an executive order to make federal regulations clearer [Document design: a, 1980], which led the Internal Revenue Service to spend considerable time reconstructing the tax form. The movement for the use of plain language also reached other government agencies, such as in Washington State Courts (Dyer et al., 2013), which reports the need for the court to use plain language, as approximately 65% of families who go to court in Washington do not have the assistance of a lawyer.

Furthermore, in Brazil, the Fiscal Responsibility Law (LRF), in Article 48, establishes that fiscal oversight agencies must widely disclose both original and simplified versions of the following documents: plans, budgets, and budgetary guidelines laws; financial statements and their respective prior opinions; and the summarized fiscal management report. However, Article 48 does not provide clear guidelines for the construction of simplified versions of these documents. Additionally, the Brazilian Court of Accounts Members Association (ATRICON) recommended, in mid-2023, the use of plain language and visual law in the creation of their documents (*Atricon Recomenda Que Tribunais de Contas Adotem Linguagem Simples e Direito Visual – Atricon*, n.d.).

2.2 Generative AI's applications

Generative Artificial Intelligence (GenAI), particularly large language models (LLMs), has undergone rapid evolution since its inception. The foundation for modern LLMs was laid by early neural language models like Word2Vec (Mikolov et al., 2013) and ELMo (Peters et al., 2018), which introduced contextual word embeddings. A significant leap occurred with the Transformer architecture (Vaswani et al., 2017), enabling models to process sequential data more efficiently through self-attention mechanisms. This breakthrough paved the way for models like GPT (Generative Pre-trained Transformer) (Openai et al., n.d.) and BERT (Devlin et al., 2019), which demonstrated the power of pre-training on large corpora followed by fine-tuning for specific tasks.

The release of ChatGPT in 2022 marked a turning point, showcasing the capabilities of LLMs in generating human-like text and engaging in conversational interactions. Built on the GPT-3.5 and later GPT-4 architectures (OpenAI, n.d.), ChatGPT achieved widespread adoption, reaching 1 million users in just five days. The tool began being used in a virous scenarios, ranging from assisting in literature reviews (Haman & Školník, 2024)to recipe recommendations with faster, more personalized, and diverse meal options (Papastratis et al., 2024). It was also investigated how ChatGPT can contribute to the software requirements engineering process, exploring how it can assist in the activities of elicitation, validation, and documentation of requirements (Marques et al., 2024).

In the public sector, these technologies are being cautiously adopted to bridge the gap between legal systems and citizens. Brazilian courts have pioneered applications ranging from document simplification (Silva et al., 2024) to AI-powered judicial assistants like JuLIA (Araújo, 2024). Brazil's national audit institution (TCU) has implemented AI tools to streamline internal processes, including document analysis and workflow automation (*Uso de Inteligência Artificial Aprimora Processos Internos No Tribunal de Contas Da União – Notícias | Portal TCU*, n.d.), demonstrating institutional commitment to technological innovation. The Brazilian Court of Accounts of Pernambuco' initiative examined in this study represents a critical advancement in this domain, demonstrating how AI can operationalize constitutional transparency mandates at scale. However, the use of these tools must be planned to reduce risks and have mitigation plans in place should these risks occur (K. Alves et al., 2024).

2.3 Prompt Technique

A prompt is the set of instructions sent as input to generative AI models. It is through these prompts that the models generate new data. However, generative AIs are not 100% accurate; they can hallucinate, meaning they may introduce incorrect data in the final responses. However, there are ways to improve the performance of the responses obtained. By constructing the prompt well, it is possible to minimize hallucinations and get responses that are closer to what is expected, without needing to retrain the language model. Thus, in this section, we clarify the prompt techniques used in the experiment conducted in this work. The following techniques were applied: Tree of Thought, Costar, Zero Shot Prompt, One-shot Prompt, and Meta Prompting.

- **Tree of thought**: In this prompt style, the problem is sent to the model without any input examples. However, the prompt explicitly asks the model to divide the problem into subproblems and generate alternative solutions for each of them. The model will then evaluate each possible solution and select the best one (Long, 2023).
- **Costar**: Costar is a framework to support the construction of prompts. In this model, the prompt must be constructed with the following aspects: context, objective, style, tone, audience, and response. In other words, all these aspects are mapped in the prompt to contextualize the LLM and also to determine the expected response format.

- **Zero Shot Prompt**: In this prompt style, the task to be performed is described in detail, without labeled training data (Sahoo et al., 2024).Thus, the model is led to use its pre-existing knowledge through a deeply described activity.
- **One-shot Promp**t: The One-shot is the prompt style that refers to using an input example for model learning (Chen et al., 2024) In other words, the model will receive an input with its respective expected result, and the input will be used for the LLM to work and achieve an output based on the example input for learning.
- **Meta prompting**: This prompt technique focuses on the structure and format of the interaction with the LLM, allowing greater control over how the model processes and presents information (Zhang et al., 2024).Therefore, there is a greater concern in describing which topics should be addressed in the response, for example.

2.3 Related Works

Although generative AI has only recently become widely known, there are already several related works on text simplification using this technology. In Ospina-Henao et al. (Ospina-Henao et al., 2024), generative AI was used to assist in simplifying technical texts for people with cognitive impairments, the elderly, and non-natives. Martinez et al. (Martínez et al., 2024) highlight the existence of guides for creating simple and easy-to-read content in plain language. However, there is no standardization between these guides, and standardization usually needs to be done manually, which is more costly.

A prominent study in the area of automatic text simplification is presented at (A. Alves et al., 2023). A total of 100 documents from the Regional Federal Court of the 5th Region (TRF5) and 100 documents from the Supreme Federal Court (STF), two major Brazilian courts, were used. The simplifications were performed using MUSS(EN), MUSS(PT), Transformers, and NMT + Attention. After simplification, the texts were evaluated using the Flesch Reading Ease (FRE) method. Only the NMT + Attention technique did not improve text readability, while the Transformers model achieved a score of 64.71, the highest readability average compared to the original text in TRF5 documents. MUSS(EN) achieved a readability score of 60, making it the best-performing model for STF documents. The study also used a human evaluation form to assess the simplified texts; however, Plain Language guidelines were not applied to determine whether the texts were effectively simplified.

Another study is being presented by Silva et al. (Silva et al., 2024). In this work, the target audience for simplification was individuals with a high school education, and prior opinions issued by the Court of Accounts of Pernambuco were used. The ChatGPT-4 was used, and the prompt technique applied was the zero-shot prompt. The simplified texts in this work contained little or no technical language while maintaining the main points of the decision. However, the work used only one Large Language Model, one prompt technique, and no simplification process was performed using plain language guides.

3. Methodology

The main objective of this work is to transform government documents, often perceived as complex by the general public, into plain language. To achieve this, we formulated the following research questions:

RQ01 – Which prompt techniques and Large Language Models most effectively translate the State Court of Accounts of Pernambuco’s preliminary decisions into plain language for the audience of at least high school graduates?

RQ02 – Which Large Language Model demonstrates superior performance in evaluating the quality of AI-generated plain language texts, as measured by consistency with human judgments?

To address these questions, we structured our methodology into six complementary steps, aiming to create simplified versions of the preliminary decisions that are both satisfactory and generated automatically. The six steps are: (i) collecting the preliminary decisions; (ii) creating ten simplified versions of the decisions; (iii) automatic evaluation of simplified documents; (iv) selecting the best model; (v) human evaluation; (vi) grade alignment analysis. Figure 1 illustrates the methodology steps.

3.1 Collecting the preliminary decisions

The preliminary decisions were obtained using a system called “Decisões Simplificadas” (Translates to “Simplified Decisions”) (Silva et al., 2024) which already contained the decisions. This system performed preprocessing by dividing the original preliminary decisions into logical sections, which were then stored in a database for easy retrieval and use in generating simplified versions. The scope of the preliminary decisions was limited to a specific state and its capital, as it is a state-level court of accounts. The analysis focused on the most significant decisions for the state and its capital, starting from year 2016, totaling 14 documents.

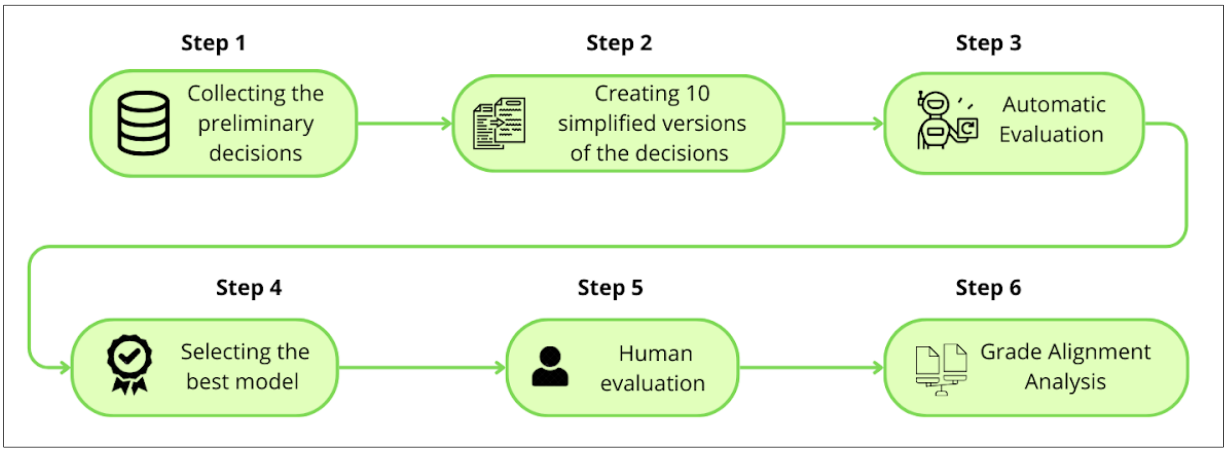
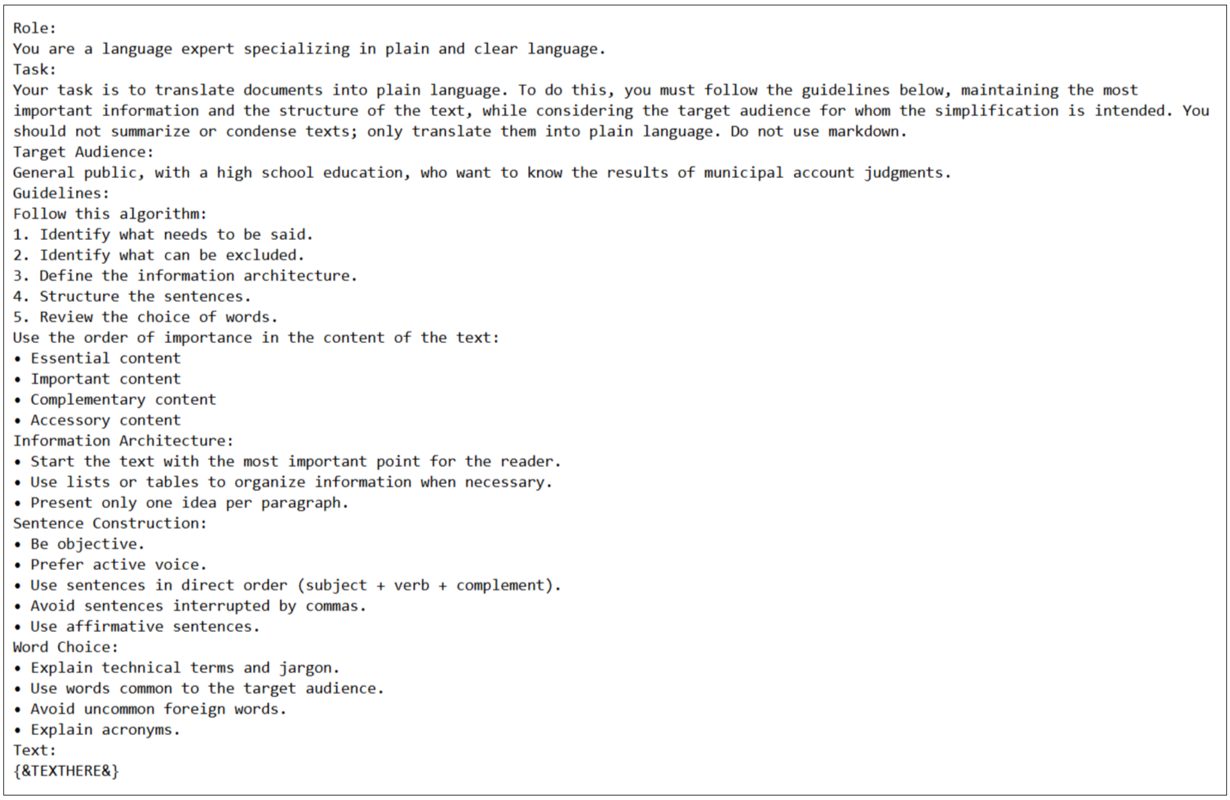


Fig. 1 – Methodology Diagram

3.2 Create ten simplified versions of the decision

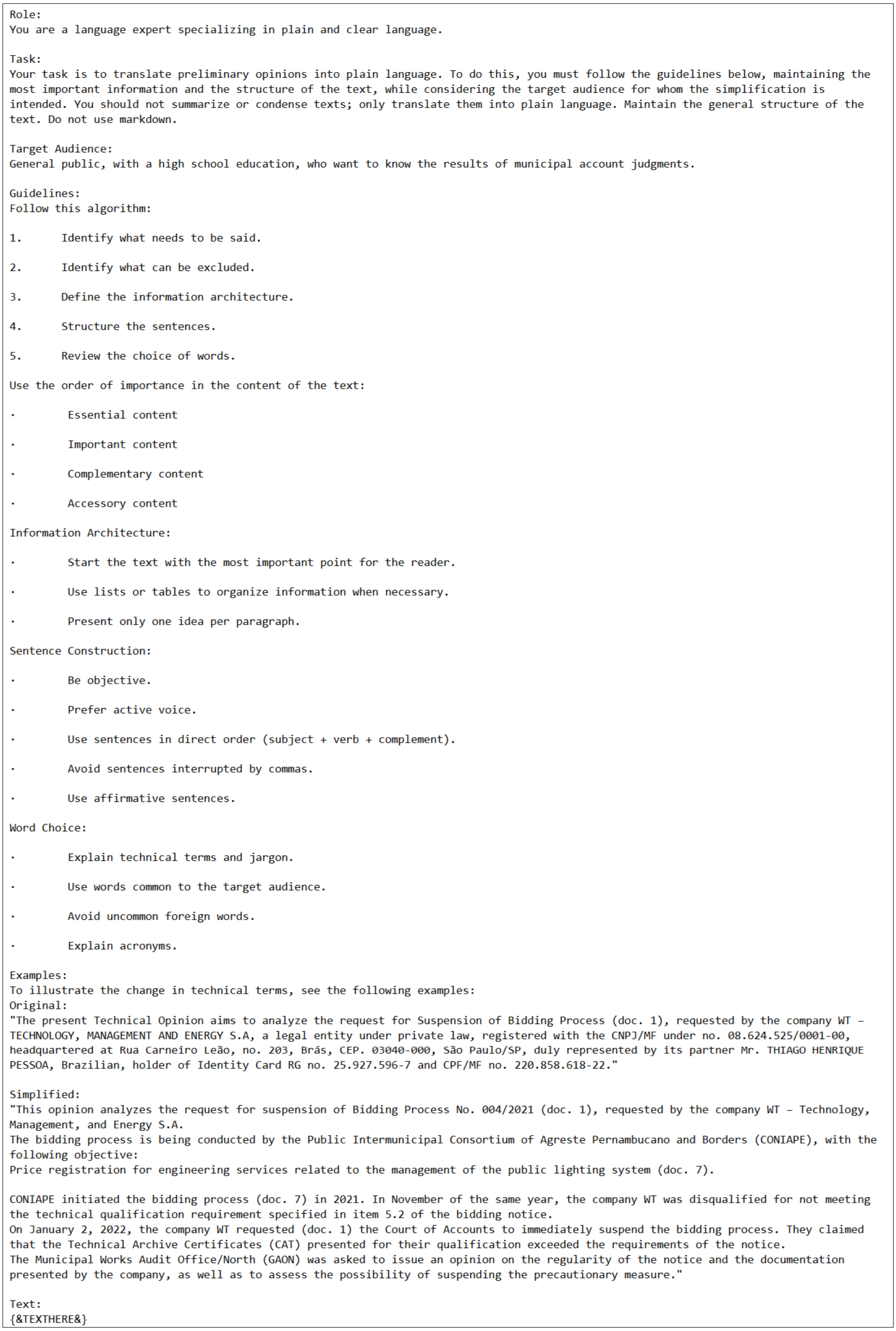
Then, we chose five prompt techniques—zero-shot, one-shot, COSTAR, meta-prompting, and tree of thoughts—were selected for their varying approaches to contextual understanding and generative capabilities. The respective prompts are available in figures 2, 3, 4, 5 and 6. The prompts and all original texts were in Portuguese, the working language of the Court of Accounts.



```
Role:
You are a language expert specializing in plain and clear language.
Task:
Your task is to translate documents into plain language. To do this, you must follow the guidelines below, maintaining the most
important information and the structure of the text, while considering the target audience for whom the simplification is intended. You
should not summarize or condense texts; only translate them into plain language. Do not use markdown.
Target Audience:
General public, with a high school education, who want to know the results of municipal account judgments.
Guidelines:
Follow this algorithm:
1. Identify what needs to be said.
2. Identify what can be excluded.
3. Define the information architecture.
4. Structure the sentences.
5. Review the choice of words.
Use the order of importance in the content of the text:
• Essential content
• Important content
• Complementary content
• Accessory content
Information Architecture:
• Start the text with the most important point for the reader.
• Use lists or tables to organize information when necessary.
• Present only one idea per paragraph.
Sentence Construction:
• Be objective.
• Prefer active voice.
• Use sentences in direct order (subject + verb + complement).
• Avoid sentences interrupted by commas.
• Use affirmative sentences.
Word Choice:
• Explain technical terms and jargon.
• Use words common to the target audience.
• Avoid uncommon foreign words.
• Explain acronyms.
Text:
{&TEXTHERE&}
```

Fig. 2 – Zero Shot Prompt

Regarding the Large Language Models, Google's Gemini 1.5 flash and OpenAI's GPT 4o mini were selected. The decision to use these tools was based on their popularity and endorsement, which facilitate the process of simplification. To automate the process of generating each text, Python scripts were used to connect with the database containing the original preliminary decision texts. The GPT and Gemini APIs were then used to send the prompts, applying the chosen prompt technique along with the original text. The response from the API, containing the plain language version, was stored in a table within the same database.

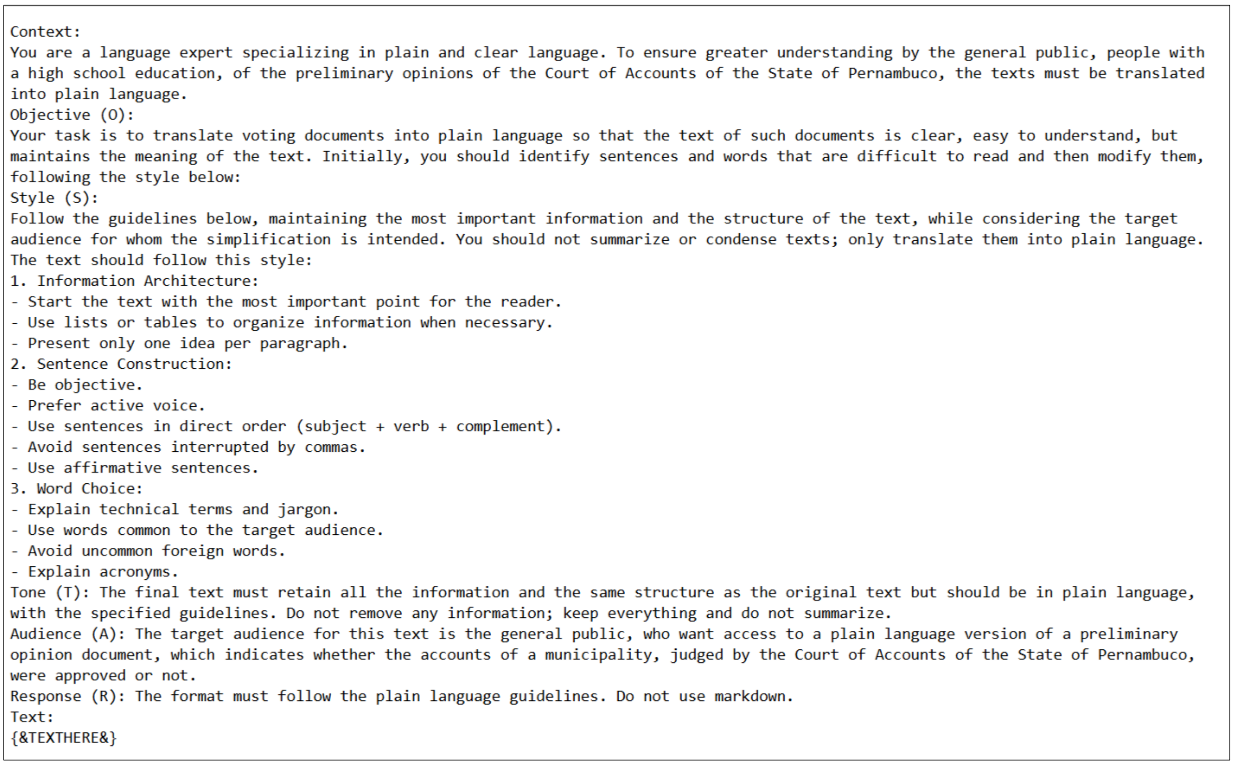


```
Role:
You are a language expert specializing in plain and clear language.

Task:
Your task is to translate preliminary opinions into plain language. To do this, you must follow the guidelines below, maintaining the
most important information and the structure of the text, while considering the target audience for whom the simplification is
intended. You should not summarize or condense texts; only translate them into plain language. Maintain the general structure of the
text. Do not use markdown.

Target Audience:
General public, with a high school education, who want to know the results of municipal account judgments.

Guidelines:
Follow this algorithm:

1.      Identify what needs to be said.

2.      Identify what can be excluded.

3.      Define the information architecture.

4.      Structure the sentences.

5.      Review the choice of words.

Use the order of importance in the content of the text:

·        Essential content

·        Important content

·        Complementary content

·        Accessory content

Information Architecture:

·        Start the text with the most important point for the reader.

·        Use lists or tables to organize information when necessary.

·        Present only one idea per paragraph.

Sentence Construction:

·        Be objective.

·        Prefer active voice.

·        Use sentences in direct order (subject + verb + complement).

·        Avoid sentences interrupted by commas.

·        Use affirmative sentences.

Word Choice:

·        Explain technical terms and jargon.

·        Use words common to the target audience.

·        Avoid uncommon foreign words.

·        Explain acronyms.

Examples:
To illustrate the change in technical terms, see the following examples:
Original:
"The present Technical Opinion aims to analyze the request for Suspension of Bidding Process (doc. 1), requested by the company WT –
TECHNOLOGY, MANAGEMENT AND ENERGY S.A, a legal entity under private law, registered with the CNPJ/MF under no. 08.624.525/0001-00,
headquartered at Rua Carneiro Leão, no. 203, Brás, CEP. 03040-000, São Paulo/SP, duly represented by its partner Mr. THIAGO HENRIQUE
PESSOA, Brazilian, holder of Identity Card RG no. 25.927.596-7 and CPF/MF no. 220.858.618-22."

Simplified:
"This opinion analyzes the request for suspension of Bidding Process No. 004/2021 (doc. 1), requested by the company WT – Technology,
Management, and Energy S.A.
The bidding process is being conducted by the Public Intermunicipal Consortium of Agreste Pernambucano and Borders (CONIAPE), with the
following objective:
Price registration for engineering services related to the management of the public lighting system (doc. 7).

CONIAPE initiated the bidding process (doc. 7) in 2021. In November of the same year, the company WT was disqualified for not meeting
the technical qualification requirement specified in item 5.2 of the bidding notice.
On January 2, 2022, the company WT requested (doc. 1) the Court of Accounts to immediately suspend the bidding process. They claimed
that the Technical Archive Certificates (CAT) presented for their qualification exceeded the requirements of the notice.
The Municipal Works Audit Office/North (GAON) was asked to issue an opinion on the regularity of the notice and the documentation
presented by the company, as well as to assess the possibility of suspending the precautionary measure."

Text:
{&TEXTHERE&}
```

Fig. 3 – One Shot Prompt

3.3 Automatic Evaluation:

In this step, the quality of the generated simplified versions was evaluated using two complementary approaches. First, an 18-question questionnaire, primarily based on Patricia Roedel's plain language guidebook (Roedel, 2024), was developed to assess adherence to plain language principles. This questionnaire included specific guidelines for simplification as well as broader questions about readability and comprehension. The original text, simplified text, and questionnaire were sent to the same large language models, which generated a score between 0 and 100 for each simplified version based on the plain language criteria. The questionnaire and the questions' weights is available at table 1.

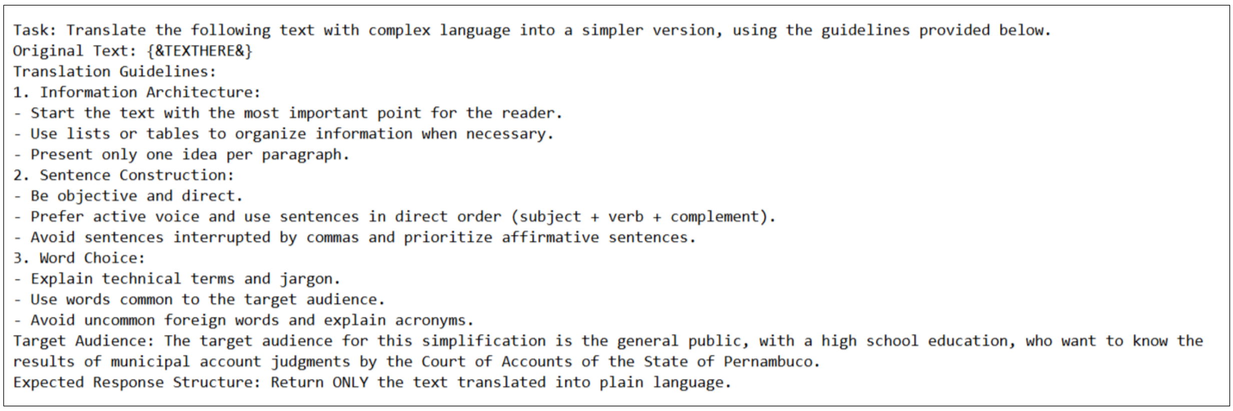


```
Context:
You are a language expert specializing in plain and clear language. To ensure greater understanding by the general public, people with a high school education, of the preliminary opinions of the Court of Accounts of the State of Pernambuco, the texts must be translated into plain language.
Objective (O):
Your task is to translate voting documents into plain language so that the text of such documents is clear, easy to understand, but maintains the meaning of the text. Initially, you should identify sentences and words that are difficult to read and then modify them, following the style below:
Style (S):
Follow the guidelines below, maintaining the most important information and the structure of the text, while considering the target audience for whom the simplification is intended. You should not summarize or condense texts; only translate them into plain language.
The text should follow this style:
1. Information Architecture:
- Start the text with the most important point for the reader.
- Use lists or tables to organize information when necessary.
- Present only one idea per paragraph.
2. Sentence Construction:
- Be objective.
- Prefer active voice.
- Use sentences in direct order (subject + verb + complement).
- Avoid sentences interrupted by commas.
- Use affirmative sentences.
3. Word Choice:
- Explain technical terms and jargon.
- Use words common to the target audience.
- Avoid uncommon foreign words.
- Explain acronyms.
Tone (T): The final text must retain all the information and the same structure as the original text but should be in plain language, with the specified guidelines. Do not remove any information; keep everything and do not summarize.
Audience (A): The target audience for this text is the general public, who want access to a plain language version of a preliminary opinion document, which indicates whether the accounts of a municipality, judged by the Court of Accounts of the State of Pernambuco, were approved or not.
Response (R): The format must follow the plain language guidelines. Do not use markdown.
Text:
{&TEXTHERE&}
```

Fig. 4 – CoStar Prompt

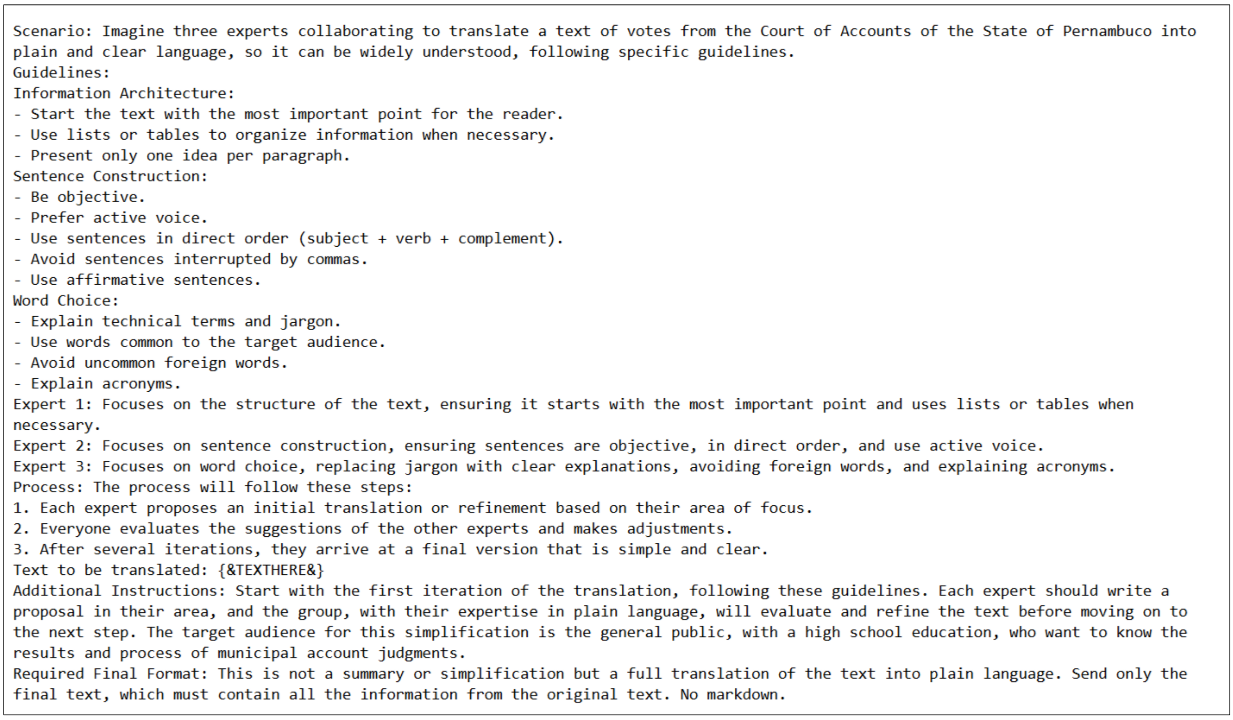
In addition to the questionnaire-based evaluation, we assessed the readability of each simplified text using two standardized metrics: the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL).

The Flesch Reading Ease score, calculated using Portuguese-adapted formulas that account for the language's specific phonetic and syntactic characteristics, measures text accessibility on a scale from 0 to 100. Higher scores indicate greater readability, with scores above 50 generally considered appropriate for audiences with basic literacy skills. This metric considers both average sentence length and average syllables per word.



```
Task: Translate the following text with complex language into a simpler version, using the guidelines provided below.
Original Text: {&TEXTHERE&}
Translation Guidelines:
1. Information Architecture:
- Start the text with the most important point for the reader.
- Use lists or tables to organize information when necessary.
- Present only one idea per paragraph.
2. Sentence Construction:
- Be objective and direct.
- Prefer active voice and use sentences in direct order (subject + verb + complement).
- Avoid sentences interrupted by commas and prioritize affirmative sentences.
3. Word Choice:
- Explain technical terms and jargon.
- Use words common to the target audience.
- Avoid uncommon foreign words and explain acronyms.
Target Audience: The target audience for this simplification is the general public, with a high school education, who want to know the results of municipal account judgments by the Court of Accounts of the State of Pernambuco.
Expected Response Structure: Return ONLY the text translated into plain language.
```

Fig. 5 – Meta Prompting



```
Scenario: Imagine three experts collaborating to translate a text of votes from the Court of Accounts of the State of Pernambuco into
plain and clear language, so it can be widely understood, following specific guidelines.
Guidelines:
Information Architecture:
- Start the text with the most important point for the reader.
- Use lists or tables to organize information when necessary.
- Present only one idea per paragraph.
Sentence Construction:
- Be objective.
- Prefer active voice.
- Use sentences in direct order (subject + verb + complement).
- Avoid sentences interrupted by commas.
- Use affirmative sentences.
Word Choice:
- Explain technical terms and jargon.
- Use words common to the target audience.
- Avoid uncommon foreign words.
- Explain acronyms.
Expert 1: Focuses on the structure of the text, ensuring it starts with the most important point and uses lists or tables when
necessary.
Expert 2: Focuses on sentence construction, ensuring sentences are objective, in direct order, and use active voice.
Expert 3: Focuses on word choice, replacing jargon with clear explanations, avoiding foreign words, and explaining acronyms.
Process: The process will follow these steps:
1. Each expert proposes an initial translation or refinement based on their area of focus.
2. Everyone evaluates the suggestions of the other experts and makes adjustments.
3. After several iterations, they arrive at a final version that is simple and clear.
Text to be translated: {&TEXTHERE&}
Additional Instructions: Start with the first iteration of the translation, following these guidelines. Each expert should write a
proposal in their area, and the group, with their expertise in plain language, will evaluate and refine the text before moving on to
the next step. The target audience for this simplification is the general public, with a high school education, who want to know the
results and process of municipal account judgments.
Required Final Format: This is not a summary or simplification but a full translation of the text into plain language. Send only the
final text, which must contain all the information from the original text. No markdown.
```

Fig. 6 – Tree of Thought

The Flesch-Kincaid Grade Level metric converts the FRE score into the corresponding U.S. educational grade level needed to comprehend the text. Lower scores represent simpler texts that require less formal education to understand. For our target audience of high school graduates (typically corresponding to grade 12), we aimed for FKGL scores at or below this threshold. These Portuguese-adapted versions of the tests maintain the original metrics' validity while accounting for linguistic features specific to Portuguese, such as its typically longer words and more complex verb conjugations compared to English.

The combination of these two approaches—questionnaire scores and readability metrics—allowed for a comprehensive evaluation of the simplified texts. The questionnaire scores were used to determine how well the simplified versions adhered to plain language principles, while the readability metrics provided objective measures of text complexity. Together, these evaluations helped identify the best-performing model.

3.4 Selecting the best model:

Scores were calculated using a weighted average, with each question assigned a specific weight, based on the importance of each question. Particular emphasis was placed on assessing whether the large language model omitted information or generated hallucinations, ensuring the output was a true simplification rather than a summary (questions q1 and q2). Besides, higher weights were also given on questions 17 and 18, Using the LLM-generated scores for each simplified version, the next step was to determine the best-performing model by identifying the combination of prompt technique and Large Language Model that achieved the highest average grade. Alongside this, the Flesch Reading Ease and Flesch-Kincaid Grade Level metrics were calculated to verify whether the simplified versions were indeed easier to read compared to the original texts. While these readability metrics were not used to select the best model, they provided additional validation that the simplified texts were more accessible to a broader audience.

3.5 Human evaluation:

The original version of the document, the simplified version, and the questionnaire were sent to human evaluators, who assessed the 14 text samples from the prompt-model combinations that achieved the best average performance. Although the human evaluators were not language specialists, they received instructions on how to conduct the evaluations, including guidelines on the principles of plain language, its objectives, and its definition. These evaluations were limited to the simplified versions produced by the selected best-performing model. To quantify the differences between human and LLM grades, tests on Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), metrics widely used in the evaluation of AI models were conducted. These metrics were chosen as they provide clear insights into the consistency and accuracy of the LLM evaluations compared to human assessments.

Tab. 1 – Evaluation Questionnaire

No.	question	weight
q1	Does the simplified text omit any information from the original text?	5
q2	Does the simplified text include any information that is not in the original text?	5
q3	Does the text present the most important information first?	1
q4	Does the text structure the information using lists, tables, graphs, etc.?	1
q5	Does the text convey only one idea per paragraph?	1
q6	Does the text use the minimum number of words?	1
q7	Does the text use the active voice?	1
q8	Does the text use direct word order?	1
q9	Does the text avoid sentences with intercalated clauses separated by commas?	1
q10	Does the text use words that are generally familiar to the target audience?	1
q11	Does the text translate technical terms and jargon?	1
q12	Does the text use precise words?	1
q13	Does the text spell out acronyms in full?	1
q14	Does the text avoid foreign words without common usage?	1
q15	Does the text avoid nominalizing verbs?	1
q16	Does the text use any pejorative terms?	1
q17	Has the simplification compromised the full understanding of the content?	3
q18	Is the simplified text appropriate for the target audience's level of knowledge?	3

4. Results

In the following section, we present the results achieved, discussing the most relevant aspects of the implemented methodology and the findings obtained during its development.

4.1 Research Question 01

For Research Question 01, "Which prompt techniques and Large Language Models most effectively translate the State Court of Accounts of Pernambuco's preliminary decisions into plain language for the audience of at least high school graduates?", it was crucial to identify the best-performing prompt technique and large language model for translating text into a plain language version. To develop this, it was necessary to understand the overall performance of each combination of prompt techniques and large language model. Each preliminary decision was sent to GPT 4o mini and Gemini 1.5 flash, using five different prompts for evaluation. The evaluation process then generated automatic scores, as described in the methodology section.

With the automatic scoring system, we evaluated different combinations of prompt techniques and large language models (LLMs) to determine the most effective approaches, as summarized in Table 2. The results show that Gemini achieved the highest score (67.64) using the "Tree of Thought" technique, followed closely by its performance with "COSTAR" (67.36). GPT's strongest results came from the same "Tree of Thought" approach (67.21), slightly outperforming its "COSTAR" score (66.07). Notably, Gemini consistently scored higher than GPT across most techniques, with the exception of "Tree of Thought," where GPT was nearly on par. The weakest performance came from GPT using "Meta Prompting" (62.79), which scored lower than even GPT's "Zero Shot" baseline (63.71). Overall, structured techniques like "Tree of Thought" and "COSTAR" delivered the best results for both models, suggesting that methodical, multi-step prompting strategies yield superior performance compared to simpler approaches like zero-shot or one-shot prompting.

Table 3 illustrates the omission rate, additional info rate, comprehension compromised rate, and audience suitability rate from each combination of prompt technique and LLM. These rates were based on the questionnaire responses to Questions 1, 2, 17, and 18, where the values are expressed as percentages. As shown, GPT with the

COSTAR and Zero Shot (90%) technique had the lowest omission rate, even though this means that from all of the documents, 90% had something missing. This was closely followed by Gemini with the Tree of Thought technique, which also performed well. Most models struggled to follow the instructions to avoid omitting information, with most exhibiting an omission rate of 100%, indicating either missing information or significant data loss. On a positive note, none of the models added information that was not present in the original text, indicating they did not hallucinate.

Tab. 2 – Average Score of each model

Average Score	LLM	Prompt Technique
67,64	gemini	Tree Of Thought
67,36	gemini	COSTAR
67,21	gpt	Tree Of Thought
66,71	gemini	One Shot
66,07	gemini	Meta Prompting
66,07	gpt	COSTAR
65,71	gemini	Zero Shot
63,71	gpt	Zero Shot
62,79	gpt	Meta Prompting
60,71	gpt	One Shot

Regarding the comprehension compromised rate, GPT led with the Tree of Thought at 37%, followed by Gemini with One Shot at 45%, and Gemini with Tree of Thought at 50%, maintaining a good average across the rates. For audience suitability, all models performed well, except for Gemini using the COSTAR prompt.

Tab. 3 – Omission, additional info, comprehension compromised and audience suitability rate by model

LLM	Prompt Technique	Omission Rate	Additional Info Rate	Comprehension Compromised Rate	Audience Suitability Rate
gemini	Tree Of Thought	91%	0%	50%	100%
gpt	Tree Of Thought	100%	0%	37%	100%
gemini	Meta Prompting	100%	0%	54%	100%
gpt	Meta Prompting	100%	0%	59%	100%
gpt	COSTAR	90%	0%	63%	100%
gemini	Zero shot	100%	0%	50%	100%
gpt	Zero shot	90%	0%	59%	100%
gemini	One Shot	100%	0%	45%	100%
gpt	One Shot	100%	0%	59%	100%
gemini	COSTAR	100%	0%	50%	90%

Table 4 presents readability scores for different prompt techniques and the two different models utilized, compared with the original texts. The readability metrics were the Flesch Reading Ease (FRE), which has higher scores for easier to read texts, and Flesch-Kincaid Grade Level (FKGL), that has lower scores for easier to read texts. The original text has very low readability (FRE: 11.13) and a high FKGL score (22.71), indicating that it is highly complex and likely inaccessible to a general audience, which confirmed the need for simplification to enhance accessibility. GPT + COSTAR achieved the best readability score (FRE: 55.26, FKGL: 8.82), which was approximately five times the original FRE metric, going from 'very difficult' score to just below the plain language standard, making it the most effective at simplifying the text. Tree of Thought, by the metrics, underperformed compared to other techniques, especially with Gemini, which was contrary to the other conclusions.

Tab. 4 – Flesch and Flesch-Kincaid for each model

Prompt technique	LLM	FRE average	FKGL Average
Original text		11.13	22.71
COSTAR	gpt	55.26	8.82
Meta Prompting	gpt	54.97	8.40
Zero shot	gpt	53.62	8.75
One Shot	gpt	51.86	9.22
Tree Of Thought	gpt	50.79	9.39
COSTAR	gemini	44.66	11.43
One Shot	gemini	40.51	11.11
Meta Prompting	gemini	40.40	11.43
Tree Of Thought	gemini	39.40	13.10
Zero shot	gemini	36.29	12.68

Overall, based on all metrics from the tables 2, 3, and 4 the best-performing combination for simplified text generation was Gemini with Tree of Thought, which had the highest average score, created the best simplified version for four of the preliminary decisions, kept a good average across the rates, even though it had a non-satisfactory result in the readability metrics. This model was followed by GPT with COSTAR, which created the best scored simplified version, lowest omission rate, and the second highest average score.

4.2 Research Question 02

To address Research Question 02—"Which Large Language Model demonstrates superior performance in evaluating the quality of AI-generated plain language texts, as measured by consistency with human judgments?"—we conducted a comparative analysis between AI-generated evaluations and human assessments. This evaluation aimed to determine the reliability of large language models in assessing the quality of simplified preliminary decisions.

As described in the methodology section, the evaluation process involved human raters grading the simplified versions produced by the best-performing prompt technique and LLM combination (Gemini with Tree of Thought). The human evaluations were then compared with the AI-generated scores from both Gemini and GPT to quantify discrepancies. To measure the degree of deviation between human and AI-generated scores, we employed Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

Tab. 5 – MAE and MAPE for each model

LLM	MAE	MAPE
Gemini	22.14	32.34%
GPT	16.79	25.76%
Average	19.03	28.5%

The results indicate that GPT provided scores that were closer to human evaluations than Gemini. GPT's lower MAE and MAPE suggest a more reliable alignment with human judgment, making it a stronger candidate for evaluating plain language texts. In contrast, Gemini's higher MAE and MAPE suggest a greater deviation from human ratings, meaning its evaluation system might not be as precise for this specific task.

A notable finding is that the AI-generated scores consistently overestimated readability and clarity, suggesting that large language models may havea bias toward assuming their outputs are clearer than they actually are for human readers. Additionally, while Gemini outperformed GPT in text simplification, it performed worse in evaluation, raising concerns about its effectiveness as an assessment tool.

Another important observation is the variability of AI evaluations across different texts. While some simplified versions received similar scores from both AI and human evaluators, others showed significant discrepancies. This suggests that AI evaluations may struggle with certain linguistic features, such as legal terminology, sentence complexity, or implicit contextual meaning.

5. Discussion

This study demonstrates that Large Language Models can *operationalize* plain language principles at scale, a task previously hindered by labor-intensive manual processes (Petelin, 2010). Where traditional simplification relies on linguists or legal experts to draft single versions, AI models like Gemini and GPT generate context-aware variants in seconds, as evidenced by our 140 simplified texts. This efficiency could transform public institutions' capacity to comply with transparency laws (e.g., Brazil's LAI) while reducing costs. However, our findings reveal a critical gap: AI's tendency to omit details (Table 3) mirrors the 'summary effect' observed in early machine translation (A. Alves et al., 2023), underscoring that *scalability does not equate to reliability*. Hybrid workflows—combining AI's speed with human oversight for legal precision—emerge as the optimal path forward.

Haman & Školník (2024) demonstrated GPT's potential for complex texts, which we extend to legal simplification. Alves et al. (2023) improved readability using models like Transformers and MUSS, but our study goes further by incorporating plain language principles, human evaluation, and a hybrid approach to ensure both readability and information retention. These comparisons highlight our study's novelty in enhancing public access to legal information while addressing prior limitations.

As for ethical and legal considerations, the deployment of AI for legal text simplification raises critical challenges regarding accountability (e.g., determining liability for errors in AI-generated documents), risks of misinformation from model hallucinations or omissions, and potential erosion of public trust in official records. These concerns are particularly acute in legal contexts where precision is paramount, as highlighted by AI safety literature (Ray, 2023) and emerging regulations like the EU AI Act's high-risk classification for public-sector AI. While Brazil's evolving AI governance framework acknowledges transparency needs (K. Alves et al., 2024), specific safeguards for legal applications—such as mandatory human review of AI outputs, audit trails, and public education about AI-processed documents—must be prioritized to balance accessibility with legal integrity.

However, despite promising results, some challenges can be highlighted. GPT outperformed Gemini in accuracy, suggesting its reliability for preliminary assessments. However, discrepancies between human and AI scores indicate that a hybrid approach—combining AI with human review—is most effective. Despite structured prompts, omissions were common, showing that critical information loss remains a risk. Another limitation noted is the small sample size, which may be sufficient for a preliminary analysis, however may not capture the full complexity of legal languages across diverse contexts.

Moreover, human evaluation was limited to 14 simplified versions due to resource constraints and conducted by non-specialists, potentially introducing bias. While evaluators were trained in plain language principles, their lack of legal expertise could affect judgments about information retention and clarity. Incorporating feedback from both legal professionals and target audience representatives (e.g., high school graduates) would strengthen future evaluations. Additionally, the cost of proprietary models like Gemini and GPT is a limitation, and future studies could explore open-source models for cost-effective solutions. Lastly, models evolve and new models are being created constantly, as in the case of the very recent DeepSeek (DeepSeek-AI et al., 2025), launched on the week of the submission of this paper. Thus, methodologies like the one presented in this study, mostly based on the prompt rather than exactly on the basis model, are important to be pursued.

While this study focused on proprietary models (GPT-4 and Gemini), open-source alternatives like Mistral and LLaMA offer potential cost-effective solutions, though their performance on legal text simplification remains untested. Domain-specific improvements could be achieved through fine-tuning on Brazilian legal corpora or Retrieval-Augmented Generation (RAG) architectures to reduce omissions of critical legal concepts. The superior evaluation accuracy of GPT over Gemini (Table 5) may stem from architectural differences—GPT's reinforcement learning from human feedback (RLHF) appears better optimized for consistency with human judgments, whereas Gemini's strength in creative generation may compromise evaluation precision. Future work should investigate whether these performance gaps persist when models are specifically adapted to legal Portuguese and constrained by RAG systems.

Finally, the study focused on Portuguese-to-plain-Portuguese simplification, which poses unique challenges due to grammatical structures (e.g., complex verb conjugations) and untranslatable legal terms. While readability metrics were adapted for Portuguese, further research is needed to address cultural and linguistic nuances that AI models may overlook. These limitations highlight the need for hybrid human-AI workflows to ensure both accessibility and precision.

Future research should investigate AI consistency over time, incorporate qualitative methods like expert interviews, and develop hybrid evaluation frameworks combining AI and human judgment. Exploring domain-specific AI models, prompt engineering, and ethical implications, such as biases in multilingual contexts, could further refine AI's role in text evaluation. This study lays the groundwork for developing robust, fair, and adaptable AI-assisted evaluation frameworks.

6. Concluding Remarks

Enhancing the clarity of public documents to improve citizen comprehension is a global goa., moreover in the context of Brazil's redemocratization and the need for greater transparency in public administration, the 1988 Constitution established the right to access public information. However, the complexity of legal language, particularly in court documents, poses a significant barrier to understanding for the general public. This study aimed to address this issue by leveraging large language models to simplify legal texts from the Court of Accounts of Pernambuco into plain language, making them more accessible to individuals with a high school education level. The research sought to identify the most effective combination of prompt techniques and AI models for this task, while also evaluating the reliability of AI-based assessments of simplified texts.

The methodology involved generating 140 simplified versions of 14 preliminary decisions using two large language models (GPT and Gemini) and five prompt techniques (Tree of Thought, COSTAR, Zero Shot, One Shot, and Meta Prompting). These simplifications were evaluated using an 18-question questionnaire based on plain language principles, with scores generated by AI models and validated through human review. The study also analyzed readability metrics, omission rates, and audience suitability to assess the quality of the simplified texts.

For RQ01, which focused on identifying the most effective prompt techniques and AI models for simplifying legal texts, the results showed that Gemini with the Tree of Thought technique achieved the highest average score (67.64) based on responses to the plain language questionnaire, and produced the best simplified versions for four of the 14 decisions. However, GPT with the COSTAR technique also performed well, particularly in preserving essential information and achieving the highest readability scores (Flesch Reading Ease: 55.26), which was just below the plain language standard. Despite these successes, omissions of critical information were a common issue across all models, indicating the need for human oversight to ensure the integrity of the simplified texts.

For RQ02, which examined the effectiveness of AI models in evaluating plain language texts, the findings revealed that GPT outperformed Gemini in evaluation accuracy, with lower Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) compared to human evaluations. However, AI models tended to overestimate readability and comprehension, suggesting that while they can provide rapid preliminary assessments, human review remains essential for ensuring the quality and accuracy of simplified texts.

The study also highlighted the practical implications of using Large Language Models for text simplification in the public sector. By automating parts of the simplification process, governments can reduce costs and improve accessibility to legal and administrative documents. However, the findings emphasize the importance of a hybrid approach, combining AI-generated simplifications with human review to address limitations such as omissions and readability overestimations. This approach can enhance transparency and citizen participation in public administration, aligning with the goals of the Access to Information Law (LAI) and the Open Government Partnership (OGP).

The main contributions of this work include: (i) identifying the most effective prompt techniques and AI models for simplifying legal texts, (ii) demonstrating the potential and limitations of AI-based evaluations for plain language transformations, and (iii) proposing a hybrid evaluation framework that integrates AI-generated assessments with human oversight. These findings provide a foundation for future research on AI-assisted text simplification and evaluation in the public sector.

For future work, we recommend: (i) exploring domain-specific fine-tuning of open-source AI models to improve performance and reduce costs, (ii) conducting longitudinal studies to assess the consistency of AI evaluations over time, (iii) incorporating qualitative research methods to gain deeper insights into the strengths and limitations of AI-driven evaluations, and (iv) investigating the ethical and cultural implications of AI-based text simplification to ensure inclusivity and fairness. By addressing these areas, future research can develop more robust and reliable tools for plain language transformation, ultimately enhancing public access to legal and administrative information.

References

Alves, A., Miranda, P., Mello, R., & Nascimento, A. (2023). *Automatic Simplification of Legal Texts in Portuguese Using Machine Learning*. https://doi.org/10.3233/FAIA230975

Alves, K., Santos, E., Silva, M. F., Chaves, A. C., Fernandes, J. A., Valenca, G., & Brito, K. (2024). Towards the regulation of Large Language Models (LLMs) and Generative AI use in the Brazilian Government: the case of a State Court of Accounts. *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance*, 28–35. https://doi.org/10.1145/3680127.3680219

Andersen, T. B. (2009). E-Government as an anti-corruption strategy. *Information Economics and Policy*, *21*(3), 201–210. https://doi.org/10.1016/j.infoecopol.2008.11.003

APSC (Australian Public Service Comission). (2023). *Australian Government Style Manual*. https://www.stylemanual.gov.au/

Araújo, R. (2024, November 18). *JuLIA Explica: novo módulo da IA do TJ-PI simplifica o acesso a informações processuais | Tribunal de Justiça do Piauí*. https://www.tjpi.jus.br/portaltjpi/tjpi/noticias-tjpi/julia-explica-novo-modulo-da-ia-do-tj-pi-simplifica-o-acesso-a-informacoes-processuais/

Atricon recomenda que Tribunais de Contas adotem linguagem simples e direito visual – Atricon. (n.d.). Retrieved April 5, 2025, from https://atricon.org.br/atricon-recomenda-que-tribunais-de-contas-adotem-linguagem-simples-e-direito-visual/

Bandeira, G. (2024, March 22). *54,5% dos brasileiros têm formação básica completa, diz IBGE*. https://www.poder360.com.br/educacao/545-dos-brasileiros-tem-formacao-basica-completa-diz-pnad/

Belém, M. (2013). A simplificação da linguagem jur\'\idica como meio de aproximação do cidadão à justiça. *Revista Jur\'\idica Da Seção Judiciária de Pernambuco*, *6*, 313–320.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2024). *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*. https://arxiv.org/abs/2310.14735

Constituição. (n.d.). Retrieved April 15, 2025, from https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. https://arxiv.org/abs/2501.12948

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/V1/N19-1423

Dyer, C., Fairbanks, J., Greiner, M., Barron, K., Skreen, J., Cerrillo-Ramirez, J., Lee, A., & Hinsee, B. (2013). Improving Access to Justice: Plain Language Family Law Court Forms in Washington State. *Seattle Journal for Social Justice*, *11*(3). https://digitalcommons.law.seattleu.edu/sjsj/vol11/iss3/10

Haman, M., & Školník, M. (2024). Using ChatGPT to conduct a literature review. *Accountability in Research*, *31*(8), 1244–1246. https://doi.org/10.1080/08989621.2023.2185514

Heerlijk Helder | Vlaanderen.be. (n.d.). Retrieved January 30, 2025, from https://www.vlaanderen.be/intern/werkplek/ondersteuning/heerlijk-helder

Klarspråk - Språkrådet. (n.d.). Retrieved January 30, 2025, from https://sprakradet.no/klarsprak/

Long, J. (2023). *Large Language Model Guided Tree-of-Thought*. https://arxiv.org/abs/2305.08291

Marques, N., Silva, R. R., & Bernardino, J. (2024). Using ChatGPT in Software Requirements Engineering: A Comprehensive Review. *Future Internet*, *16*(6), 180. https://doi.org/10.3390/fi16060180

Martínez, P., Ramos, A., & Moreno, L. (2024). Exploring Large Language Models to generate Easy to Read content. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1394705

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in

Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3

OpenAI. (n.d.). *GPT-4 Technical Report*.

Openai, A. R., Openai, K. N., Openai, T. S., & Openai, I. S. (n.d.). *Improving Language Understanding by Generative Pre-Training*. Retrieved April 15, 2025, from https://gluebenchmark.com/leaderboard

Ospina-Henao, V., Flórez, S. L., Núñez, V. J. M., Lamas, Ó. L., & De la Prieta, F. (2024). *Generative AI: Simplifying Text for Cognitive Impairments and Non-native Speakers* (pp. 33–44). https://doi.org/10.1007/978-3-031-73538-7_4

Papastratis, I., Konstantinidis, D., Daras, P., & Dimitropoulos, K. (2024). AI nutrition recommendation using a deep generative model and ChatGPT. *Scientific Reports*, *14*(1), 14620. https://doi.org/10.1038/s41598-024-65438-x

Petelin, R. (2010). Considering plain language: issues and initiatives. *Corporate Communications: An International Journal*, *15*(2), 205–216. https://doi.org/10.1108/13563281011037964

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Plain language, accessibility, and inclusive communications - Privy Council Office - Canada.ca. (n.d.). Retrieved January 30, 2025, from https://www.canada.ca/en/treasury-board-secretariat/topics/government-communications/communications-community-office/communications-101-boot-camp-canadian-public-servants/plain-language-accessibility-inclusive-communications.html

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Roedel, P. (2024). *Manual de Linguagem Simples* (Edições Câmara, Ed.).

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. https://arxiv.org/abs/2402.07927

Silva, M., Santos, E., Alves, K., Silva, H., Pedrosa, F., Valença, G., & Brito, K. (2024). Using Generative AI for Simplifying Official Documents in the Public Accounts Domain. *Anais Do XII Workshop de Computação Aplicada Em Governo Eletrônico (WCGE 2024)*, 246–253. https://doi.org/10.5753/wcge.2024.2915

Uso de inteligência artificial aprimora processos internos no Tribunal de Contas da União – Notícias | Portal TCU. (n.d.). Retrieved April 16, 2025, from https://portal.tcu.gov.br/imprensa/noticias/uso-de-inteligencia-artificial-aprimora-processos-internos-no-tribunal-de-contas-da-uniao

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://arxiv.org/abs/1706.03762v7

Zhang, Y., Yuan, Y., & Yao, A. C.-C. (2024). *Meta Prompting for AI Systems*. https://arxiv.org/abs/2311.11482