Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5814

Exploring Brazil's LLM Fauna: Investigating the Generative Performance of Large Language Models in Portuguese

Gabriel Assis [**Universidade Federal Fluminense** | ***assisgabriel@id.uff.br***]
Cláudia Freitas [**Universidade de São Paulo** | ***claudiafreitas@usp.br***]
Aline Paes [**Universidade Federal Fluminense** | ***alinepaes@ic.uff.br***]

Institute of Computing, Universidade Federal Fluminense, Av. Gal. Milton Tavares de Souza, s/n, São Domingos, Niterói, RJ, 24210-590, Brazil.

Received: 01 April 2025 • **Accepted:** 25 May 2025 • **Published:** 08 October 2025

Abstract. Large Language Models (LLMs) are now embedded in widely used applications worldwide, yet their evaluation still centers on narrow, discriminative benchmarks. These pipelines often overlook key generative aspects such as discourse coherence, linguistic transformations, and adequacy, which are crucial for real-world applications. In addition, most large-scale evaluations remain heavily biased toward English, limiting our understanding of LLM performance in other languages. This research addresses these gaps by presenting a comprehensive analysis of Brazilian Portuguese LLMs across three core Natural Language Generation tasks: summarization, simplification, and generative question answering. We evaluate six Brazilian models and compare them to the widely used GPT-4o. Our findings, supported by diverse automatic metrics, an LLM-as-a-judge framework, and human evaluation, show that GPT-4o series achieves the best generative performance in Portuguese, followed closely by the Sabiá-3 family. While slightly behind, the open-weight model Tucano stands out for its computational efficiency, making it a strong candidate for deployment in resource-constrained settings. The code used to conduct all experiments is publicly available at `https://github.com/MeLLL-UFF/brfauna-gen-eval`.

Keywords: LLMs, NLG-Evaluation, Question-Answering, Summarization, Simplification, Brazilian Portuguese

1 Introduction

Large Language Models (LLMs) have transformed a wide range of tasks and applications in Natural Language Processing (NLP), with their influence extending across domains such as law [Malaquias-Junior *et al.*, 2024; Pereira *et al.*, 2025], finance [Assis *et al.*, 2024b; Xie *et al.*, 2024], healthcare [Schneider *et al.*, 2021; Paiola *et al.*, 2024b], and creative fields like music [Yuan *et al.*, 2024] and image understanding and generation [Lee *et al.*, 2024]. Once a topic primarily confined to academic and industrial research, LLMs are now embedded in widely used technologies, increasingly serving as the backbone of everyday digital services and applications across the globe[1] — for example, through Google's search tools powered by Gemini [Gemini-Team *et al.*, 2023][2], assistants integrated into messaging applications such as WhatsApp[3], or popular chatbots like ChatGPT[4]. This widespread impact also resonates within the Brazilian context, where the development and adaptation of local LLMs have accelerated in recent years. In this work, we refer to this emerging ecosystem as the *LLM Fauna*, reflecting a trend in the international open-source community (*e.g.*, LLaMA [Touvron *et al.*, 2023a], Vicuna [Chiang *et al.*, 2023], Falcon [Almazrouei *et al.*, 2023]) of naming models after animals. Brazilian LLMs often follow this tradition, adopting names of native species such as Bode [Garcia *et al.*, 2024], Boto [Santa Brígida, 2024a], Tucano [Corrêa *et al.*, 2024b], and Sabiá [Pires *et al.*, 2023].

Despite this increasing deployment of LLMs in broad-range applications, their *evaluation* often remains mostly restricted to controlled, task-specific, and discriminative benchmarks typically designed for academic or research settings [Zellers *et al.*, 2019; Wang *et al.*, 2019; Chen *et al.*, 2021; Hendrycks *et al.*, 2021a; Rein *et al.*, 2024]. While this focus is understandable, considering the cost and complexity of evaluating large-scale models, it may obscure important aspects of model behavior. Benchmarks targeting multiple-choice knowledge, code generation, or mathematical reasoning naturally provide valuable insights into factual performance and logical capabilities. However, they do not capture more nuanced features, such as dialogue coherence, writing quality, and other pragmatic skills that are essential in real-world use cases [Chiang *et al.*, 2024].

Some evaluations attempt to measure these conversational capabilities of generative models [Zheng *et al.*, 2023; Chiang *et al.*, 2024], nonetheless, they remain overwhelmingly centered on English [Singh *et al.*, 2025]. This reinforces a significant gap with respect to other languages, for instance, Portuguese [Joshi *et al.*, 2020; Longpre *et al.*, 2025]. Moreover, delivering solutions for specific audiences requires alignment not only in linguistic terms but also in cultural and strategic dimensions [Pires *et al.*, 2023; Pawar *et al.*, 2025].

Given that users tend to interact with LLMs in their native language, they may naturally expect these systems to align

[1] `https://www.technollama.co.uk/a-gemini-report-how-many-people-are-using-generative-ai-on-a-daily-basis-a-gemini-report`
[2] `https://blog.google/products/search/generative-ai-google-search-may-2024/`
[3] `https://cohere.com/research/aya/whatsapp`
[4] `https://chatgpt.com/`

with locally situated communicative norms, pragmatic conventions, and culturally grounded expectations [Pawar *et al.*, 2025]. This alignment is especially relevant in tasks involving naturalistic interaction, which encompass a significant portion of generation-based applications. Prior studies in multilingual NLP have shown that models trained or adapted to specific languages and cultural settings tend to produce outputs that are more appropriate, intelligible, and aligned with user expectations [Hovy and Spruit, 2016; Joshi *et al.*, 2020; Pawar *et al.*, 2025]. Accordingly, assessing how effectively LLM-powered applications capture such contextual conditions is not only a matter of performance but of linguistic equity and inclusiveness. In this work, we take a step in this direction by focusing on Brazil and its official language, Portuguese (PT-BR).

Additionally, the training and deployment of LLMs require significant infrastructure [Corrêa *et al.*, 2024b], which developing countries like Brazil often struggle to establish [Lehdonvirta *et al.*, 2025]. To illustrate this disparity, Meta's planned AI infrastructure spending for 2025 alone amounts to $65 billion USD[5], while Brazil's national AI plan [Government of Brazil, 2025] allocates less than $5 billions USD for the entire 2024–2028 period. Likewise, the environmental costs are also concerning, involving substantial electricity usage, freshwater consumption, and large-scale CO_2 emissions [Li *et al.*, 2023]. Despite their importance, these impacts are often overlooked, even in dominant-language settings such as English [Luccioni *et al.*, 2025].

Within this scope, we raise the following central question: *What is the generative performance of LLM-powered solutions specifically designed for Brazilian Portuguese?* We complement this investigation with an analysis of environmental impact, measuring carbon emissions, computational time, and energy consumption. To empirically explore this question, we evaluate nine LLMs: (i) Sabiá-3 [Abonizio *et al.*, 2024], (ii) Sabiazinho-3 [Abonizio *et al.*, 2024], (iii) Tucano [Corrêa *et al.*, 2024b], (iv) Bode [Garcia *et al.*, 2024], (v) Cabra [BotBot-AI, 2024a], (vi) Periquito [Gibaut, 2023], and (vii) Boto [Santa Brígida, 2024b]. For comparative purposes, we also include two general-purpose state-of-the-art models: (viii) GPT-4o [OpenAI *et al.*, 2024b] and its compact variant, (ix) GPT-4o-mini. Our goal is to understand how these Brazil-driven models perform beyond traditional discriminative benchmarks, by also analyzing the nature and quality of their generative outputs.

We adopt three tasks in the Natural Language Generation (NLG) setting, namely text summarization [Souza *et al.*, 2024b], text simplification [Leal and Aluísio, 2024], and question answering (QA) [Cortes *et al.*, 2024]. These tasks serve as proxies to evaluate different key model attributes. For example, summarization can assess the ability to retain essential content; simplification evaluates the model's capacity to manipulate and transform language effectively; and QA examines its ability to answer context-sensitive questions. To provide a comprehensive evaluation across these tasks, we employ a broad set of metrics that cover lexical [Papineni *et al.*, 2002; Lin, 2004; Lavie and Agarwal, 2007; Leal *et al.*, 2024], syntactic [Leal *et al.*, 2024], and semantic [Zhang *et al.*, 2020b; Vasilyev *et al.*, 2020] aspects. Additionally, we incorporate LLM-as-a-judge [Zheng *et al.*, 2024; Liu *et al.*, 2023] evaluations as well as careful human inspection. This combined approach allows for a robust and multi-angle assessment of model generative performance.

Our findings suggest that:

- Unsurprisingly, models from the GPT-4o series top-rank in Portuguese generative tasks.
- The Sabiá-3 family closely follows, standing out as a nationally developed alternative that achieves state-of-the-art performance.
- Available open-weight models do not match the generative performance of proprietary solutions but offer viable options in terms of computational efficiency, with Tucano being a notable example.
- Open-weight models do not generalize their performance across all evaluated criteria, showing divergent strengths and weaknesses. This highlights the importance of evaluating them according to the specific requirements of each use case.
- Our human review of generated texts and evaluation outputs suggests that robust automatic evaluation benefits from combining LLM-powered and traditional NLG metrics, as each helps to offset the other's weaknesses. It also shows that certain nuances remain detectable only through human judgment.

Our contributions include:

- To the best of our knowledge, the first large-scale analysis of generative performance covering six Brazilian LLMs;
- A historical overview of the development of generative solutions for Brazilian Portuguese;
- An extensive evaluation plan for three NLG tasks, covering lexical, semantic, morphosyntactic, and generative metrics;
- A complementary evaluation of computational performance and environmental impact during inference with LLMs used at scale for generative tasks in Portuguese.

Beyond the introduction, this article contains seven sections: Section 2 defines the research goal; Section 3 reviews generative models in Brazil; Section 4 discusses related work; Section 5 explains the methodology; Section 6 covers experimental settings; Section 7 presents results; and Section 8 concludes with final remarks.

2 Problem Statement

The objective of this research can be formulated as follows:

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_i\}$ denote the set of language models under evaluation, $\mathcal{T} = \{T_1, T_2, \ldots, T_j\}$ the set of generative tasks, and $\mathcal{E} = \{E_1, E_2, \ldots, E_k\}$ the set of evaluation metrics. Our goal is to systematically analyze the generative performance of each model M_i when addressing each task T_j, as measured by the corresponding metrics E_k. Based on the results, we intend to assess the generative capacity of the currently available Brazilian LLMs.

[5]https://www.nytimes.com/2025/01/24/technology/meta-data-center.html

In particular, in this work, the task set is defined as

$$\mathcal{T} = \{\text{summarization}, \text{simplification}, \text{question answering}\}$$

as briefly delineated below. The details regarding the sets $\mathcal{M}$ and $\mathcal{D}$ are discussed in the next sections.

Text Summarization is the process of generating condensed versions of existing texts, resulting in summaries or abstracts [Souza *et al.*, 2024b]. Typically, summarization is based on a source text that conveys a central idea, structured information, a clear communicative purpose, and a coherent narrative. The process then involves identifying key content and reformulating it into a concise yet cohesive summary while preserving the original intent [Rino and Pardo, 2003]. Additionally, summarization can be classified into two major types: extractive — where elements from the original text are directly selected and used in the summary — and abstractive — where a new textual piece is generated. In line with the objectives of this research, we focus on the abstractive approach.

Text Simplification is the process of reducing a text's complexity while preserving its meaning and content [Al-Thanyyan and Azmi, 2021]. The literature identifies at least three dimensions for adjusting textual complexity: cognitive — related to a reader's ability to recognize the global and local structure of a text; conceptual — associated with the background knowledge required to understand certain topics; and linguistic — the most widely explored by automatic systems — focused on lexical and syntactic modifications [Leal and Aluísio, 2024]. Our study focuses on the use of LLMs for linguistic-level simplification.

Question Answering (QA) refers to automatically answering questions posed in natural language. QA systems are usually grouped by the nature of questions they handle (factual or open-ended), their subject domain (general or specialized, *e.g.*, legal [Hu *et al.*, 2025] and environmental [Paschoal *et al.*, 2021]), and their source of information (ranging from unstructured data accessed through retrieval techniques to structured sources or alternative methods) [Cortes *et al.*, 2024]. Considering the objective of evaluating the generative abilities of Brazilian LLMs, we adopt data associated with the country in an open-ended configuration. For the knowledge source, we restrict the evaluation to the information inherently encoded in the models.

3 Generative Large Language Models for Brazilian Portuguese

Since the consolidation of the Transformer architecture [Vaswani *et al.*, 2017] and the establishment of the Pre-Trained Language Model (PTLM) paradigm [Wang *et al.*, 2023] — where models are trained on large *corpora* and subsequently made available for direct use or fine-tuning in downstream tasks — a plethora of language models tailored to Brazilian Portuguese (PT-BR) have emerged. Corrêa *et al.* [2024b] undertakes the effort of chronologically enumerating key language models developed for the language, including the well-established encoder-only model BERTimbau [Souza *et al.*, 2020], the decoder-only models Sabiá [Pires *et al.*, 2023] and the pioneering GPorTuguese-2 [Guillou, 2020], as well as the encoder-decoder model PTT5 [Carmo *et al.*, 2020]. While that study provides a broad overview of language models for Brazilian Portuguese, the present article focuses specifically on text generative-oriented models for the language, thus emphasizing decoder-only and encoder-decoder solutions.

The approaches applied to build those PTLMs range from general-purpose models — such as Sabiá-3 [Abonizio *et al.*, 2024] and InternLM-ChatBode [Recogna-NLP, 2024g] — to domain-specific models, including Juru [Malaquias-Junior *et al.*, 2024] for legal contexts and GPT2-Bio-Pt [Schneider *et al.*, 2021] and DrBode [Recogna-NLP, 2024b; Paiola *et al.*, 2024b] for biomedical and medical domains. Most approaches rely on fine-tuning available open-weight models [Henrique, 2023b,a; BotBot-AI, 2023, 2024c,a,b], such as Falcon [Almazrouei *et al.*, 2023], Llama [Touvron *et al.*, 2023a,b; Meta-AI, 2024], Qwen [Bai *et al.*, 2023; Yang *et al.*, 2024], InternLm [Cai *et al.*, 2024], and earlier versions of GPT [Radford *et al.*, 2019]. However, some efforts involve pre-training models from scratch, as represented by architecturally Llama-based models TeenyTinyLlama [Corrêa *et al.*, 2024a] and Tucano [Corrêa *et al.*, 2024b].

Figure 1 presents a chronological record of initiatives in the development of generative PT-BR-models[6]. The timeline begins with GPortuguese-2 [Guillou, 2020], a fine-tuned version of GPT-2 small [Radford *et al.*, 2019] (≈120 million parameters) trained on the Portuguese portion of Wikipedia. Building on this as a base model, GPT2-Bio-Pt [Schneider *et al.*, 2021] is a specialized variant further fine-tuned on a 16-million-token biomedical *corpus*, enhancing its generative capabilities within this domain.

Beyond these models, PTT5 [Carmo *et al.*, 2020] adapts the encoder-decoder T5 [Raffel *et al.*, 2020] architecture for Portuguese, having been trained on the BrWaC *corpus* [Wagner Filho *et al.*, 2018], which comprises over 2 billion tokens. This model family includes six versions, ranging from 60 million to over 700 million parameters, with some implementations initialized from the multilingual vocabulary of the original model and others trained from scratch on Portuguese text. More recently, Piau *et al.* [2024] introduced PTT5-v2, built with the same Portuguese vocabulary as its predecessor while leveraging Continued Learning techniques on the Portuguese portion of mC4 [Xue *et al.*, 2021], a significantly larger *corpus* than the one previously used. This version also introduces a new parameter size option, featuring approximately 3 billion parameters. Notably, PTT5 represents a milestone as a "command-oriented" option for the language, given its ability to generate outputs based on the prefix accompanying the input. Previous models had purely focused on next-token prediction.

[6]Although we strive to cover as many models as possible, we acknowledge that numerous other model versions are available on platforms such as Hugging Face. On Open-PT-LLM-Leaderboard alone, there are over 100 entries specializing in Portuguese. Here, we focus on those that have been consolidated and have documented their development, including details such as the datasets used.

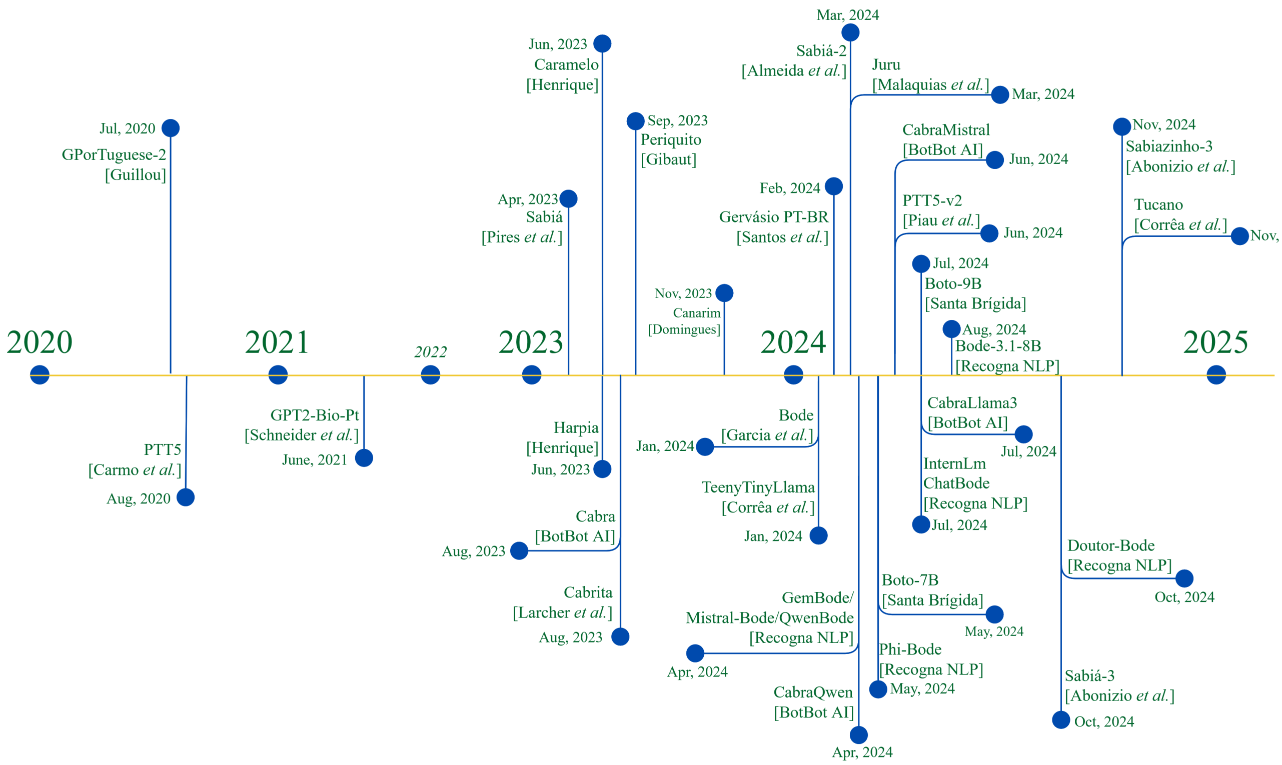

Figure 1. Timeline of initiatives for creating generative language models for Brazilian Portuguese, trained on the autoregressive task.

While 2022 marks a gap in the emergence of key PT-BR solutions, it also saw the global release of the proprietary GPT-3.5 [Brown *et al.*, 2020], which captivated the world with its remarkable ability to generalize and respond to instructions. During the early months of 2023, the open-source community initiated a response, with the launch of the first versions of models like LLaMA [Touvron *et al.*, 2023a] and Falcon [Almazrouei *et al.*, 2023].

Driven by these open initiatives, the development of Brazilian solutions experienced significant growth. One important development was the release of the first versions of the Sabiá model family [Pires *et al.*, 2023]. At the time, fine-tuned versions were introduced based on the 7B and 65B parameter models from the LLaMA family, along with an adaptation of GPT-J [Wang and Komatsuzaki, 2021]. These models were trained on the ClueWeb 2022 dataset [Overwijk *et al.*, 2022]. Over time, the Sabiá family has expanded and now includes proprietary models such as Sabiá-2 [Almeida *et al.*, 2024], Sabiá-3, and Sabiazinho-3 [Abonizio *et al.*, 2024], which currently stand as the state of the art in generative models for PT-BR. While only the initial Sabiá-7B model has been made publicly available, details regarding subsequent versions, including their base models and training *corpora*, remain undisclosed. Moreover, Juru [Malaquias-Junior *et al.*, 2024], a legally specialized model, is a fine-tuned adaptation of Sabiá-2, which is also unreleased to the public.

Similarly to the first version of Sabiá, the Caramelo [Henrique, 2023a] and Harpia [Henrique, 2023b] models represent fine-tuned alternatives based on the Falcon-7B model, utilizing, respectively, the machine-translated Portuguese version of the Alpaca *corpus* [Taori *et al.*, 2023][7] and a subset of the OASST1[8] dataset. Additionally, other models have been built on the OpenLLaMA 3B [Geng and Liu, 2023] architecture, namely Cabrita [Larcher *et al.*, 2023] — also fine-tuned using the PT-BR portion of mC4 [Xue *et al.*, 2021] — and Periquito [Gibaut, 2023] — which has also been adapted with the Portuguese-language data from Wikipedia.

The release of the Llama-2 architecture [Touvron *et al.*, 2023b] led to the development of several fine-tuned adaptations, including Canarim [Domingues, 2023], a 7B-parameter model fine-tuned on the CommonCrawl 2023-23 dataset[9]; Gervásio-PTBR [Santos *et al.*, 2024], another 7B-parameter model fine-tuned on the PT-BR ExtraGLUE-instruct dataset[10]; and the initial versions of Bode and Cabra. The first Bode models [Garcia *et al.*, 2024] (7B and 13B parameters) were trained on PT-BR Alpaca resources [Taori *et al.*, 2023][7], while Cabra [BotBot-AI, 2023] was developed as a 7B-parameter model fine-tuned on the Portuguese Dolly Instruct dataset[11].

The Bode model family has expanded through the replication of its methodology across various base models, including Phi-2 [Javaheripi *et al.*, 2023] and Phi-Bode [Recogna-NLP, 2024f], Mistral-7B [Jiang *et al.*, 2023] and Mistral-Bode [Recogna-NLP, 2024e], Gemma-7B [Gemma-Team *et al.*, 2024] and Gem-Bode [Recogna-NLP, 2024c], Qwen-1.8B [Bai *et al.*, 2023] and Qwen-Bode [Recogna-NLP, 2024g], Llama-3.1 [Meta-AI, 2024] and Bode-3.1-8B [Recogna-NLP, 2024a], as well as InternLm2 [Cai *et al.*, 2024] and InternLm-ChatBode [Recogna-NLP, 2024d]. The latter

[7] Hugging Face: dominguesm/alpaca-data-pt-br

[8] Hugging Face: timdettmers/openassistant-guanaco

[9] Hugging Face: dominguesm/CC-MAIN-2023-23

[10] Hugging Face: PORTULAN/extraglue-instruct

[11] PT-BR translation of Databricks Dolly [Conover *et al.*, 2023].

also features a further fine-tuned version on a *corpus* of medical texts, namely DrBode [Recogna-NLP, 2024b; Paiola *et al*., 2024b]. Similarly, a range of Cabra model variants has emerged, including CabraMistral [BotBot-AI, 2024b], CabraQwen [BotBot-AI, 2024c] (based on Qwen 1.5 7B [Qwen Team, 2024]) and CabraLlama3-8B [BotBot-AI, 2024a]. These recent versions of Cabra were trained on self-curated datasets, referred to as the Cabra datasets[12]. Likewise, the Boto series introduced other PTLMs based on Gemma. This includes a 7B version trained on the initial Gemma release, as well as 9B and 27B variants developed using Gemma-2 [Gemma-Team, 2024]. All Boto models have been trained on the Cetacean-PTBR dataset[13].

Although also based on Llama-2, Corrêa *et al*. [2024a,b] adopts a distinct experimental setup in their development, specifically performing pre-training from scratch rather than focusing on fine-tuning. In the first duo of released models, two versions named TeenyTinyLlama [Corrêa *et al*., 2024a] — one with 160 million and another with 460 million parameters — are introduced, both trained on the constructed Portuguese-Corpus Instruct dataset. Meanwhile, the Tucano [Corrêa *et al*., 2024b] model family comprises four versions, with approximately 160M, 630M, 1.1B, and 2.4B parameters each. Notably, its training is based on GigaVerbo [Corrêa *et al*., 2024b], the largest publicly available unified Portuguese *corpus* introduced to date. This dataset consolidates a range of *corpora*, including many previously mentioned, alongside other curated Portuguese-language resources. These model series also stand out for explicitly evaluating GPU usage and associated carbon emissions during pre-training.

Overall, it is evident that most efforts toward developing generative models for Brazilian Portuguese focus on adapting foreign solutions to the language. The importance and impact of open-source initiatives in fostering the local development ecosystem are particularly noteworthy. Additionally, most models developed so far do not exceed a few dozen billion parameters, a trend that may be linked to previous findings indicating that countries like Brazil — or languages such as Portuguese — often lack the resources necessary to build robust and high-performing solutions when compared to more resource-unrestricted environments, such as those associated with the English language [Joshi *et al*., 2020; Lehdonvirta *et al*., 2025; Longpre *et al*., 2025]. This brief historical contextualization highlights the significant progress made while underscoring that there is still a crucial path ahead to be explored.

4 Related Work

This section reviews traditional approaches to LLM benchmarking. It also outlines the research conducted on their application to the NLG-oriented tasks examined in this study, specifically summarization, text simplification, and question answering, with a special focus on PT-BR.

[12] Hugging Face: botbot-ai/Cabra3k

[13] Hugging Face: lucianosb/cetacean-ptbr

4.1 Large Language Model Evaluation

Evaluations on benchmarks accompany the release of most top-tier LLMs, to assess key capabilities and dimensions such as truthfulness [Lin *et al*., 2022], harmlessness [Zhang *et al*., 2024], reasoning and knowledge [Zellers *et al*., 2019; Hendrycks *et al*., 2021a; Wang *et al*., 2019; Srivastava *et al*., 2023; Mialon *et al*., 2024; Rein *et al*., 2024; Sprague *et al*., 2024]. These benchmarks vary in scope, with some covering general content, while others target specific areas such as mathematics [Cobbe *et al*., 2021; Hendrycks *et al*., 2021b], coding [Chen *et al*., 2021; Austin *et al*., 2021; Patil *et al*., 2024], and conversational abilities [Chiang *et al*., 2024; Zheng *et al*., 2023; Zhou *et al*., 2023]. Moreover, certain benchmarks are designed for specialized domains, including finance [Xie *et al*., 2024], law [Guha *et al*., 2023], and healthcare [Singhal *et al*., 2023].

Some benchmarks are based on traditional NLP tasks. For instance, HellaSwag [Zellers *et al*., 2019] is a multiple-choice question-answering dataset designed to measure models' commonsense natural language inference capabilities, requiring them to select the most coherent sentence completion. Among the most widely used benchmarks, MMLU (Massive Multitask Language Understanding) [Hendrycks *et al*., 2021a] provides a large multiple-choice question set covering 57 diverse subjects, including mathematics, history, law, and computer science. MMLU includes over 15,000 questions ranging from high school to expert level.

Furthermore, extending the GLUE benchmark [Wang *et al*., 2018], SuperGLUE [Wang *et al*., 2019] is designed to evaluate LLMs across multiple text-based tasks, such as natural language inference and question-answering, incorporating formats like multiple-choice and binary answers. Another notable benchmark, WinoGrande [Sakaguchi *et al*., 2021], builds upon the Winograd Schema Challenge, a natural language understanding task that assesses models' ability to resolve pronoun ambiguities in sentences. It is structured as a binary-choice problem. Meanwhile, BIG-bench [Srivastava *et al*., 2023] serves as a large-scale collaborative benchmark that evaluates LLMs' generalization across more than 200 diverse tasks, ranging from translation error detection and textual inference to arithmetic and logical reasoning. A particularly specialized benchmark, TruthfulQA [Lin *et al*., 2022], focuses on measuring LLMs' ability to generate factually accurate responses while minimizing the propagation of misinformation and misconceptions. It includes texts related to false beliefs, myths, pseudoscience, and misinformation in contexts such as health, law, finance, and politics. TruthfulQA features both multiple-choice and generative-oriented evaluation formats, initially requiring human assessment but also applicable with LLM-as-a-judge [Zheng *et al*., 2024] evaluation scenarios.

Focusing on mathematical reasoning, GSM8K [Cobbe *et al*., 2021] evaluates LLMs' ability to perform multi-step arithmetic operations expected of school students, evaluating accurate final results. In contrast, MATH [Hendrycks *et al*., 2021b] assesses advanced mathematical problem-solving skills in competition-style settings, requiring proficiency in algebra, calculus, geometry, and statistics. In parallel, for code generation, HumanEval [Chen *et al*., 2021] is one of the most established benchmarks, testing LLMs' ability to generate correct and functional code based on provided specifications. The correctness of the generated code is verified through unit tests. Another programming-related benchmark, MBPP (Mostly Basic Python Programming) [Austin *et al.*, 2021], evaluates LLMs' ability to produce basic Python scripts from natural language descriptions, assessing concepts such as list manipulation, conditional logic, fundamental algorithms, and string operations. The dataset includes test cases to validate output correctness.

Other benchmarks further explore the expanding capabilities of LLMs, particularly those related to multimodal functionalities, function calling, web access, and enhanced reasoning over extended contexts. GAIA [Mialon *et al.*, 2024] presents real-world challenging questions that typically require models to access online resources, integrate multimodal inputs (*e.g.*, visual reasoning), and engage in multi-step reasoning to arrive at a single answer. SWE-bench [Jimenez *et al.*, 2024] extends the evaluation of code-generation abilities by tasking LLMs with resolving full-fledged GitHub issues, necessitating the generation of patches that effectively fix reported problems. Meanwhile, the Gorilla benchmark [Patil *et al.*, 2024] assesses models' capacity to execute function calls and interact with APIs. Additionally, MuSR [Sprague *et al.*, 2024] introduces innovative tasks requiring multi-step reasoning, such as solving murder mysteries, determining object placements, and making team allocation decisions.

As safety became a key area of evaluation, benchmarks were also developed to assess potential risks. AgentHarm [Andriushchenko *et al.*, 2025] evaluates whether models can appropriately refuse to engage in harmful activities, such as fraud, cybercrime, and harassment. Similarly, SafetyBench [Zhang *et al.*, 2024] employs multiple-choice question-answering as a proxy to measure risks related to bias, illegal activities, offensive content, and mental health concerns associated with LLM usage.

Certain benchmarks are designed for highly specialized domains. GPQA [Rein *et al.*, 2024] consists of challenging multiple-choice questions authored by subject-matter experts in biology, physics, and chemistry, intended for individuals pursuing at least a *Ph.D.* in their respective fields. Additionally, traditional exams have also been repurposed for LLM evaluation, including the Law Bar Exam, which assesses legal knowledge and reasoning for prospective lawyers [Martínez, 2024; Katz *et al.*, 2024]; the SAT (Scholastic Assessment Test), a standardized test used for college admissions in the United States; the LSAT (Law School Admission Test), which evaluates logical reasoning, reading comprehension, and analytical thinking for law school applicants; and the Gaokao, China's highly competitive national college entrance exam [Zhong *et al.*, 2024]. Each of these assessments presents unique domain-specific challenges, measuring LLMs' proficiency in specialized fields and their ability to match human performance.

A comparative and publicly accessible option for evaluating the performance of LLMs was the Open LLM Leaderboard [Fourrier *et al.*, 2024][14], which aggregates the performance of open-source models across various benchmarks, including some previously mentioned or derived from those cited here, such as MuSR, MATH, GPQA, MMLU, and BIG-Bench. It also includes performance metrics for the IFEval benchmark [Zhou *et al.*, 2023], which assesses instruction-following capabilities, with scoring tied to the strict adherence to the requested format. Notably, it reports the carbon footprint of each model's evaluation.

Another well-known leaderboard is ChatBot Arena [Zheng *et al.*, 2023; Chiang *et al.*, 2024]. It focuses on evaluating the conversational capabilities of models through a crowdsourced framework, where human annotators compare responses from different LLMs in a blind pairwise setting. Users interact with two anonymized models side by side, selecting the response they find superior. These comparisons are then aggregated to generate rankings, providing a dynamic and human-in-the-loop assessment of model performance. The results are compiled based on user input prompts, with dedicated sections for code generation, creative text generation, and other task-specific or language-based evaluations. Built upon the results of ChatBot Arena and following a conversational approach, MT-Bench [Zheng *et al.*, 2023] focuses on evaluating models' capabilities in writing, extraction, reasoning, and other skills based on multi-turn dialogues between users and the models. The benchmark employs an LLM-as-a-Judge framework to perform evaluations, comparing its assessments with human preferences from ChatBot Arena to measure alignment and consistency, ultimately demonstrating its applicability.

Although some of the previously mentioned resources provide support for multiple languages — such as ChatBot Arena, which includes a section for results in languages like English, Spanish, Korean, Chinese, Russian, German, and French (but not Portuguese as of now), and SafetyBench, which offers a Chinese version alongside English — the majority of benchmarks remain English-oriented. As a result, various efforts have been made to translate these resources into multiple languages, including Portuguese, such as Global-MMLU [Singh *et al.*, 2025], MMMLU [OpenAI *et al.*, 2024a], and m-ArenaHard [Dang *et al.*, 2024]. However, coverage for Portuguese across a broader range of these resources remains limited.

In this context, the evaluation of Brazilian LLMs remains diverse, with each work utilizing the available resources at the time. In this regard, Pires *et al.* [2023], alongside the introduction of the Sabiá model, also presented Poeta (Portuguese Evaluation Tasks), a benchmark encompassing 14 downstream NLP datasets in Portuguese, covering tasks such as classification, multiple-choice question answering, sentiment analysis, and sentence entailment. The benchmark includes ASSIN 2 RTE and STS [Real *et al.*, 2020], ENEM Challenge [Cataneo Silveira and Deratani Mauá, 2018], ENEM 2022 [Nunes *et al.*, 2023], FaQuAD [Sayama *et al.*, 2019], TweetSentBr [Brum and Volpe Nunes, 2018], AG News [Zhang *et al.*, 2015], IMDB [Maas *et al.*, 2011], MASSIVE [FitzGerald *et al.*, 2023], MKQA [Longpre *et al.*,

[14]In March 2025, the leaderboard's discontinuation was announced, citing concerns that its evaluations might not capture all necessary dimensions for assessing advanced LLMs. More details at `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard/discussions/1135`

2021], BoolQ [Clark *et al.*, 2019], SST2 [Socher *et al.*, 2013], WSC [Melo *et al.*, 2019], and BLUEX [Almeida *et al.*, 2023], with some datasets being natively in Portuguese (ASSIN 2 RTE and STS, BLUEX, ENEM Challenge, ENEM 2022, FaQuAD, and TweetSentBr), while the others were translated. The first Bode [Garcia *et al.*, 2024] models were also evaluated using the same benchmark. However, Poeta was not explicitly released as a public benchmark, limiting reproducibility for evaluating other models, as reported by Corrêa *et al.* [2024a]. Nonetheless, Larcher *et al.* [2023] utilized available subsets of Poeta in isolation, while Corrêa *et al.* [2024b,a] incorporated translations of HellaSwag and TruthfulQA on it, along with additional datasets such as HateBR [Vargas *et al.*, 2022] and PT Hate Speech [Fortuna *et al.*, 2019], for evaluating their models.

Moreover, national exams have also been employed in the Brazilian context for model assessment [Almeida *et al.*, 2023; Corrêa *et al.*, 2024b; Abonizio *et al.*, 2024; Almeida *et al.*, 2024], including the previous mentioned ENEM (*Exame Nacional do Ensino Médio*) — a multiple-choice exam used for university admissions in Brazil, designed to evaluate high school students' proficiency in subjects such as mathematics, languages, sciences, and humanities — and the OAB Exam (*Ordem dos Advogados do Brasil*) — a mandatory bar exam for law graduates in Brazil. Lastly, the Open Portuguese LLM Leaderboard [Garcia, 2024] serves as a centralized platform for evaluating LLMs on Portuguese-language tasks, encompassing the ENEM, BLUEX, OAB Exams, ASSIN 2 RTE and STS, FaQuAD NLI, HateBR, PT Hate Speech, and TweetSentBR. This leaderboard provides a platform for visualizing LLM performance on Portuguese tasks and allows model submissions to be evaluated.

Despite the diversity of domains and contexts, most existing evaluations focus on discriminative, structured and controlled tasks, such as classification, multiple-choice or binary question answering, sentiment analysis, and language entailment. Even in mathematical and programming assessments, correctness is typically determined through predefined answers or automated compilation checks. Furthermore, conversational evaluation approaches — more aligned with the objectives of this research — are more scarce in general and highly limited in Portuguese. This constraint is reasonable, as large-scale assessments often require structured environments to ensure systematic evaluation, and NLG metrics often lack direct interpretability [Sai *et al.*, 2022]. However, this paradigm overlooks key aspects of a model's generative capabilities, such as writing quality, tone consistency, and appropriateness in more open-ended tasks. Given that LLMs are not only expected to perform well on benchmarks but also to support real-world conversational systems — like chatbots — we evaluate the Brazilian LLMs in inherently generative tasks in this work, specifically generative question answering, summarization, and text simplification. Further details are provided in the following sections.

4.2 Natural Language Generation Tasks

NLG focuses on enabling machines to automatically produce coherent, contextually appropriate, and human-like text from structured or unstructured data. It encompasses a wide range of tasks, including automated report generation, dialogue systems, translation, transduction, text summarization, question answering, and text simplification, each with extensive applications and possibilities [Gatt and Krahmer, 2018]. This section examines explicitly the last three tasks, as they fall within the scope of this work. We concentrate on exploring existing approaches that employ LLMs and examining the methodologies used to evaluate them, emphasizing Portuguese-centered developments.

Previous research on PTLM [Rehman *et al.*, 2022] evaluation in abstractive summarization has explored Transformer-based models such as T5, Pegasus [Zhang *et al.*, 2020a], and BART [Lewis *et al.*, 2020] on English-based benchmarks. These approaches resonate with work done in Portuguese, such as [Paiola *et al.*, 2024a], which introduced RecognaSum, a new dataset for automatic news summarization, and evaluated it using PTT5 as a strong baseline. More recently, Sarmento and de Oliveira [2024] fine-tuned PTT5 and explored Flan-T5-based [Chung *et al.*, 2024] approaches for Portuguese abstractive summarization. Their study compares these models with open LLMs like Gemma-2 and Llama-3, as well as proprietary models such as GPT-3.5 and GPT-4o [OpenAI *et al.*, 2024b], concluding that they remain competitive with larger models. These findings align with our previous work [Assis *et al.*, 2024a], which investigated PTT5 and OPT-PTBR[15] under efficient fine-tuning conditions for summarization, also leveraging RecognaSum, and demonstrated their strong performance even in resource-constrained settings.

In general, works on summarization employ metrics such as BERTScore [Zhang *et al.*, 2020b] to assess the semantic similarity between generated pieces and reference texts. However, ROUGE [Lin, 2004] remains a predominant evaluation metric in summarization research, as observed in the works above. Jorge *et al.* [2025], however, highlight its limitations — both due to its reliance on reference summaries and its emphasis on lexical overlap as the primary indicator of summary quality — advocating for alternative approaches such as the reference-free BLANC metric [Vasilyev *et al.*, 2020]. Souza *et al.* [2024b] similarly argue for the need for qualitative assessments based on criteria such as grammaticality, non-redundancy, referential clarity, focus, structure, and coherence.

Kew *et al.* [2023] conducts a systematic evaluation of LLMs for sentence simplification, assessing 44 models on English-based datasets and demonstrating that these language models often outperform traditional simplification baselines such as MUSS [Martin *et al.*, 2022]. These findings align with those of Feng *et al.* [2023], who evaluate simplification in multiple idioms, including Portuguese, and highlight the strong performance of models like GPT. Focusing entirely on Portuguese, Scalercio *et al.* [2024] proposes a T5-based method incorporating adapters that capture stylistic markers. Their study also compares the proposed approach with the MUSS baseline and GPT-3.5, showing that both the new method and the use of an LLM like GPT surpass the baseline. In fact, the traditional MUSS system itself is built on a language model, BART [Lewis *et al.*, 2020], further re-

[15] Hugging Face: monilouise/opt125M_portuguese

inforcing the strong performance of LLM-based approaches in simplification tasks. Likewise, Pereira *et al.* [2025] evaluate T5- and Flan-T5-based approaches in comparison with GPT-3.5 and GPT-4o, but focused on the legal context in Portuguese, providing evidence of their applicability even in specialized domains. Scalercio *et al.* [2025] assessed a broad range of LLMs for Portuguese sentence simplification, highlighting that general models such as GPT-4o, Qwen-2.5, and Sabiá can even outperform task-specific approaches.

Regarding evaluation metrics for this context, again, BERTScore is commonly used to evaluate semantic preservation concerning reference texts [Kew *et al.*, 2023; Scalercio *et al.*, 2024; Pereira *et al.*, 2025]. However, the SARI metric [Xu *et al.*, 2016] — which accounts for addition, deletion, and retention operations during simplification — has become the primary indicator in recent studies due to its explicit design for assessing these transformations [Dong *et al.*, 2019; Kumar *et al.*, 2020; Kew *et al.*, 2023; Scalercio *et al.*, 2024; Pereira *et al.*, 2025]. Works such as [Kumar *et al.*, 2020] and [Pereira *et al.*, 2025] also report BLEU scores [Papineni *et al.*, 2002], as previous research has shown it may correlate with human judgments of fluency and meaning preservation [Xu *et al.*, 2016]. Nonetheless, other studies indicate that BLEU should not be used as a general indicator for text simplification [Sulem *et al.*, 2018]. In addition to these metrics, a range of linguistic indicators — based on syllable and character counts, the presence of specific grammatical classes, among other features — are also applied in this context [Leal and Aluísio, 2024]. Moreover, qualitative analyses performed by human evaluators, considering criteria such as fluency, adequacy, and simplicity, are also employed [Dong *et al.*, 2019; Kumar *et al.*, 2020; Feng *et al.*, 2023].

Within the scope of Question Answering (QA) research, while traditional approaches primarily rely on supporting textual contexts or retrieve relevant documents from external sources [Cortes *et al.*, 2024; Srivastava and Memon, 2024; Souza *et al.*, 2024a], Pirozelli *et al.* [2024] propose assessing models under the closed generative question answering (GCQA) setting. In this configuration, generative models are instructed to answer questions without auxiliary context, relying solely on the knowledge encoded in their weights (from pre-train or after fine-tuning). Their study conducts a bilingual evaluation in Portuguese and English, using T5 and PTT5 models for the respective languages. As expected, their findings confirm that this task is significantly more challenging for models, particularly in specialized domains. However, this setup is also valuable for assessing the knowledge embedded within these models. Additionally, it evaluates the language model's ability to answer questions, reducing dependence on the assessment of retrieval components or function calls, for example. This approach aligns with the one implemented in this work, as we intend to focus on the generative capacity of the models.

Srivastava and Memon [2024] conduct a comprehensive review of QA evaluation methods, considering the advancements brought by LLMs. For lexical assessment, they highlight variations of precision, recall, and F1-score, which measure the proportion of correctly predicted words relative to both predictions and references. They also discuss metrics such as ROUGE, BLEU, and METEOR [Lavie and Agarwal, 2007], which rely on n-gram overlap, each with distinct considerations. For semantic evaluation, they examine approaches like BERTScore, BARTScore, and BLEURT [Sellam *et al.*, 2020], which leverage language models to capture the semantic similarity between generated responses and references. Additionally, they explore the emergence of LLM-based evaluation metrics [Zheng *et al.*, 2024], such as G-Eval [Liu *et al.*, 2023], where state-of-the-art models act as judges, serving as proxies for human feedback alignment. Lastly, they also highlight the importance of qualitative evaluation, considering aspects such as factuality, relevance, completeness, clarity, and insightfulness in assessing system-generated responses.

The aforementioned studies demonstrate that summarization, simplification, and question answering are essential tasks for evaluating various attributes in generative contexts. Furthermore, prior research provides methodologies for assessing the performance of generative LLMs in these settings, including frameworks applicable to Portuguese. While some studies incorporate Brazilian LLMs in their evaluations, to the best of our knowledge, no work has yet considered such a broad range of PT-BR models within this context. Therefore, we aim to assess their generative capabilities comprehensively, extending beyond numerical accuracy on traditional and more "closed-system" benchmarks. The methodology applied to evaluate these models is detailed in the following section.

5 Evaluating the Generative Performance of BR-LLMs

This section describes the methodology applied in evaluating the generative abilities of PT-BR LLMs. It establishes the criteria for model selection, discusses their implementation in the examined generative tasks, and details the adopted evaluation approach.

5.1 Model Selection

The primary reference for selecting the models used in this research was the *Open Portuguese LLM Leaderboard*[16], which, as previously stated in the earlier section, provides a comprehensive overview of generative language models' performance on Portuguese tasks. While none of the tasks featured in the leaderboard are purely generative, as intended in this work, the leaderboard still serves as a centralized platform that compiles general evaluations of LLMs for Portuguese-language tasks. In fact, this apparent limitation may present an additional analytical opportunity to assess whether the ranking of models in generative tasks aligns with their positioning in the leaderboard.

To assess Brazilian models, a filtering criterion was first applied to the leaderboard, restricting it to LLMs for which Portuguese is the primary language, meaning they underwent a specific training (or tuning) stage for this language. Consequently, multilingual models, such as the original Llama and

[16]Hugging Face: eduagarcia/open_pt_llm_leaderboard

Qwen families, were excluded from the analysis. Additionally, most models in the leaderboard are open-weight, meaning their parameters are publicly available. While this facilitates accessibility, it demands local computational resources for deployment. Given the hardware constraints of this study (see Section 6), only models available via remote API or those with fewer than 10 billion parameters were retained. Moreover, as the leaderboard is an open platform where any user can submit a model, we established a minimum documentation criterion by selecting only those accompanied by a textual reference, such as a report, article, system description, or model card detailing the training dataset. Lastly, only one representative per model family was selected to ensure diversity in the evaluation. For instance, since multiple Sabiá models are present in the leaderboard, the selection was limited to the highest-ranked and yet accessible version.

Directly based on these restrictions, the top-ranking models selected were **(i) Sabiá-3** [Abonizio *et al.*, 2024], **(ii) Boto-9B** [Santa Brígida, 2024b], **(iii) Bode-3.1-8B** [Recogna-NLP, 2024a], and **(iv) CabraLlama8B** [BotBot-AI, 2024a]. Also, to investigate whether model size, in terms of parameters, correlates with the generative capacity of Brazilian models, we also included two additional models from the leaderboard that ranked lower: **(v) Tucano** [Corrêa *et al.*, 2024b] and **(vi) Periquito** [Gibaut, 2023] — both with less than 4 billion parameters, compared to the 8–9 billion in the previous selections. Additionally, for comparative purposes and as general state-of-the-art baselines, we incorporated **(vii) GPT-4o** [OpenAI *et al.*, 2024b] and **(viii) GPT-4o-mini** [OpenAI *et al.*, 2024b]. Lastly, we included the smaller version of Sabiá, **(ix) Sabiazinho-3** [Abonizio *et al.*, 2024], since it has been reported to achieve performance comparable to GPT-4o-mini, allowing us to analyze it as a more compact and efficient alternative to its larger counterpart.

5.2 Task Design and Model Application

The models are evaluated on three generative tasks, namely, abstractive text summarization, linguistic text simplification, and closed generative question answering. The selection of these tasks is motivated by the fact that we conjecture that they may be applicable as proxies for assessing key dimensions of LLMs related to their generative capabilities.

For instance, text summarization enables the evaluation of a model's ability to extract and condense essential information while preserving the core meaning of longer textual pieces, requiring both content retention and proper structuring [Souza *et al.*, 2024b]. The simplification task directly assesses how LLMs handle idiomatic and linguistic transformations in Portuguese, as it involves syntactic, lexical, and structural modifications [Cortes *et al.*, 2024]. Meanwhile, the question-answering task offers insight into both the models' encoded knowledge and their reasoning capabilities, as they must comprehend the question and generate a coherent and cohesive response [Leal and Aluísio, 2024]. Collectively, these tasks enable a robust and multidimensional analysis of the generative potential of Brazilian LLMs, encompassing various aspects of language comprehension, transformation, and structured text generation.

We use the instruction-tuned checkpoints available for the selected models to enable the models to perform these tasks, leveraging their pre-existing in-context capabilities [Brown *et al.*, 2020]. In addition, no fine-tuning of the model weights is performed for any of the tasks. Although fine-tuning could improve performance and optimize results for specific tasks, this is not the primary aim of our research. Instead, we argue that evaluating the models without additional adjustments provides a more precise and more direct understanding of their generative performance in their default state, while also revealing their natural strengths, which could later guide more targeted refinements. Similarly aligned with this premise, we instruct the models using zero-shot prompts, as outlined below. The design of each prompt, while kept simple, was crafted based on insights from previous research for each task [Assis *et al.*, 2024a; Sarmento and de Oliveira, 2024; Pirozelli *et al.*, 2024; Scalercio *et al.*, 2024].

Summarization Prompt

> Summarize the following *{text_type}* concisely and directly.
> *{text_type}*: *{original_text}*
> Summary:

Simplification Prompt

> Replace the complex sentence with a simple one.
> Maintain the same meaning but make it simpler.
> Complex sentence: *{text}*
> Simple sentence:

Question Answering Prompt

> Answer the following question based on your general knowledge about *{subject}*.
> Be objective.
> Question: *{question}*
> Answer:

Within each prompt, variables were replaced with their respective content based on the datasets used. Specifically, in the summarization task, the variable $text\_type$ corresponds to options such as "News" or "Report", while in question answering, $subject$ aligns with the central theme of the dataset, such as "Law" or "Climate Change". These elements serve as subtle cues to facilitate in-context learning, simulating how typical users would naturally provide context in generative conversational scenarios. Finally, the Portuguese versions of the prompts are available in Appendices A, B, and C, respectively.

5.3 Generative Evaluation

Our evaluation framework includes a diverse set of metrics designed to capture the generative capacity of the models, covering lexical [Papineni *et al.*, 2002; Lin, 2004; Lavie and Agarwal, 2007; Leal *et al.*, 2024], syntactic [Leal *et al.*, 2024], and semantic [Zhang *et al.*, 2020b; Vasilyev *et al.*, 2020; Leal *et al.*, 2024] dimensions while also addressing some communicative aspects like readability, insightfulness and completeness [Liu *et al.*, 2023; Zheng *et al.*, 2024; Leal *et al.*, 2024]. Alongside the generative metrics, we report inference time, estimated energy consumption, and equivalent CO_2 emissions, computed using the eco2AI [Budennyy *et al.*, 2023] library, to provide a broader perspective on model efficiency. We consider these consumption and impact indicators particularly important, especially in low-resource computational settings, where resource limitations necessitate careful consideration, and also given the scale of the adopted models.

Although some NLG metrics were designed initially for particular applications — such as BLEU [Papineni *et al.*, 2002] for text translation and ROUGE [Lin, 2004] for text summarization —, their use has expanded, making them standard evaluation tools across multiple tasks. Thus, we report their values for the three applied tasks, detailing any exceptions as necessary[17].

5.3.1 Lexical Metrics

Among the lexical metrics used, we selected **(i) BLEU**, which computes a precision-based score between the predicted text and the references. It incorporates a brevity penalty to discourage excessively short generations relative to the reference; **(ii) ROUGE**, another metric based on n-gram overlap between the reference and the generated text, applied in this work in its ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum variants, all of which are reported as percentages. Each of these variants differs in how the overlap is measured: ROUGE-1 is based on unigram (1-gram) matching, ROUGE-2 on bigram (2-gram) matching, ROUGE-L on the longest common subsequence, and ROUGE-Lsum, which splits the text using "$\backslash n$"[18]. Additionally, we compute **(iii) METEOR** [Lin, 2004], which, like the previous metrics, relies on term overlap but differs from BLEU by incorporating both precision and recall into its mechanism.

Furthermore, some of the mentioned metrics are already based on traditional scoring methods such as precision and recall. However, in the case of the question-answering task, the **(iv) F1-score** is often reported explicitly [Srivastava and Memon, 2024; Pirozelli *et al.*, 2024], which we also adopt in this work. The F1-score represents the harmonic mean between precision and recall. It is important to note that, in this context, these scores are computed by considering the number of correctly predicted words relative to the references. Precisely, precision is calculated as the ratio of correctly predicted words to the total number of generated words. At the same time, recall is the ratio of correctly predicted words to the total number of words in the reference [Srivastava and Memon, 2024].

To assess the simplification task, we adopt the **(v) SARI** [Xu *et al.*, 2016], which evaluates text transformations by comparing the generated output to both the original complex text and reference simplifications. SARI measures three key aspects: adding relevant words that enhance clarity, deleting unnecessary or complex words, and retaining essential words that preserve meaning.

[17] Appendix D provides implementation details for each metric to ensure reproducibility.

[18] `https://huggingface.co/spaces/evaluate-metric/rouge`

5.3.2 Semantic Metrics

Unlike lexical metrics, which primarily focus on term co-occurrence, semantic metrics aim to bridge the gap by evaluating generations that may not use the exact same words but remain semantically aligned. Among the semantic metrics, we adopt the widely used **(vi) BERTScore** [Zhang *et al.*, 2020b], considering its precision, recall, and F1 components. This metric is built upon BERT-based models and first obtains contextual word representations by independently processing both the generated and reference texts through a language model. It then computes pairwise cosine similarity scores between the embeddings, effectively capturing semantic similarity beyond surface-level lexical matches.

Specifically for summarization, we compute the **(vii) BLANC** [Vasilyev *et al.*, 2020] metric in its BLANC-help configuration. This metric is also BERT-based and measures how much a summary improves a masked language model's ability to fill in missing words in a given text. The idea is that a high-quality summary should enhance the model's contextual understanding, thereby increasing its accuracy in predicting masked tokens. BLANC differs from the widely used summarization metric ROUGE, as it does not require reference summaries and is not dependent on exact n-gram overlap, making it a more flexible alternative for evaluating summarization quality.

5.3.3 NILC-Metrix

The **(viii) NILC-Metrix** [Leal *et al.*, 2024] is an extensive set of 200 metrics designed for linguistic analysis in Portuguese, extracting a wide range of indicators to evaluate different dimensions of textual quality. These indicators include, for example, morphosyntactic features, cohesion measures, coherence metrics, and text complexity indices. By integrating these linguistic features, NILC-Metrix offers a comprehensive assessment of text quality beyond traditional generative metrics, thereby also being computed within our evaluation framework.

Although the full set of 200 indicators can be computed automatically, large-scale processing remains time-intensive. To address this, we applied NILC-Metrix to a sampled subset of model generations, constructed as follows. For each task, we selected a representative supporting metric based on its prominence and relevance in the literature — as discussed in Section 4. Specifically, we employed BLANC for summarization, SARI for text simplification, and F1 score for question answering. Model generations were then ranked according to the corresponding metric, and one example was randomly sampled from each percentile, up to a total of 10% of the generations per model had been collected, yielding a stratified sample. This sampling strategy allows for representative coverage across the full performance spectrum to be maintained for each task.

Furthermore, the high dimensionality of the indicators space complicates both interpretation and presentation. We address this by adopting a previously established approach [Assis *et al.*, 2024a,b], which groups indicators by evaluation criterion into normalized vectors. The Euclidean distance to the reference vector is then computed, offering a quantitative estimate of model alignment, where lower values denote stronger agreement.

5.3.4 LLM-based Metric

A recent trend in evaluation methods involves leveraging LLMs as reference-free evaluators, based on the rationale that their fine-tuning and reinforcement learning aligned with human preferences allow them to generate assessments that approximate human judgment [Bavaresco *et al*., 2024]. This approach, termed LLM-as-a-judge [Zheng *et al*., 2024], takes advantage of these models' built-in reasoning abilities to perform dynamic evaluations. We instantiate a framework grounded in this premise, **(ix) G-Eval** [Liu *et al*., 2023], which instructs the LLM to assess specific dimensions as indicated by the prompt, enabling more adaptable and context-sensitive evaluations. The choice of which model to use as a judge can directly impact the results; nevertheless, GPT4o has been widely adopted due to its state-of-the-art conversational alignment [Liu *et al*., 2023; Mitchell *et al*., 2023; Aakanksha *et al*., 2024a,b], which we also employ in this research. However, given the cost of relying on this top-tier judge and the substantial volume of generations, we applied the same sampling strategy described in the previous section to remain within our budget constraints.

The G-Eval framework requires the evaluation criteria to be specified in natural language and, usually, following a 1–5 Likert format hint to guide the LLM judge. The resulting scores are then normalized, considering the model's output probabilities when generating the ratings. To define these criteria, we grounded our approach in a literature review — covering summarization [Souza *et al*., 2024b][19], simplification [Dong *et al*., 2019; Kumar *et al*., 2020; Feng *et al*., 2023], and question answering [Srivastava and Memon, 2024] — to formulate instructions inspired by guidelines designed for human evaluation. The final criteria are listed below, with their Portuguese translations available in Appendices E, F, and G.

Summarization Evaluation Criteria

> A. Grammaticality
> Grammaticality (1-5) - The summary should not contain capitalization errors and obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
>
> B. Non-redundancy
> Non-redundancy (1-5) - The summary should not contain unnecessary repetition. Repetition might appear as whole sentences being repeated, repeated facts, or excessive use of a noun when a pronoun would suffice.
>
> C. Referential Clarity
> Referential Clarity (1-5) - It should be easy to identify who or what the pronouns and noun phrases in the summary refer to. If a person or entity is mentioned, their role in the story should be clear. A reference would be unclear if an entity is mentioned without making its identity or relation to the story evident.
>
> D. Focus
> Focus (1-5) - The summary should have a clear focus; sentences should only contain information relevant to the rest of the summary.
>
> E. Structure and Coherence
> Structure and Coherence (1-5) - The summary should be well-structured and well-organized. It should not just be a heap of related information but should build from sentence to sentence into a coherent body of information about a topic.

Simplification Evaluation Criteria

> A. Adequacy
> Adequacy (1-5) - To what extent is the meaning of the original sentence preserved in the simplified version? The simplified sentence should retain the essential content of the original without significant loss of information.
>
> B. Simplicity
> Simplicity (1-5) - Is the simplified sentence actually simpler than the original? The simplification should reduce the structural complexity of the sentence, making it easier to understand without introducing incorrect or redundant information.
>
> C. Fluency
> Fluency (1-5) - Is the simplified sentence grammatically correct and well-formed? The syntactic structure should be natural and readable, with no errors that hinder understanding.

Question Answering Evaluation Criteria

> A. Answer Quality
> Answer Quality (1-5) - Scoring Guidelines:
> - 1: The answer is completely incorrect. It is entirely different from or contradicts the reference.
> - 2: The answer shows some degree of semantic similarity and includes partially correct information. However, it still contains significant discrepancies or inaccuracies compared to the reference.
> - 3: The answer correctly addresses some aspects, aligning partially with the reference. However, there are still omissions or minor inaccuracies.
> - 4: The answer is mostly correct. It provides accurate information but may contain one or more minor omissions or inaccuracies.
> - 5: The answer is correct. It demonstrates a high degree of accuracy and semantic similarity to the reference.

Lastly, also according to the related instructions found in the literature, the input provided to the model varies by task:

[19] `https://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt`

for summarization, only the summary is given; for simplification, the original sentence is provided along with the generated text; and for QA, the model receives the question, the generated answer, and the reference.

5.3.5 Human Evaluation

To complement the automatic evaluation with a human-centered perspective, we conducted a manual review of a subset of model generations. Following a sampling procedure similar to the one described in Subsection 5.3.3, we ranked the instances based on a supporting metric — this time using the G-Eval score. Each human annotator was then presented with the same information available to the LLM-as-a-judge framework. Additionally, they were provided with the rationale generated by the automatic judge and the final score assigned to each instance.

The annotators were tasked with verifying whether the rationale was appropriate in light of the generation and the evaluation criterion, and whether it plausibly justified the assigned score. In addition, they were required to independently rate the generation on a Likert scale for the given criterion and to provide a brief explanatory comment. This setup allowed for both an assessment of generation quality and a validation of the reliability of the LLM-grading approach.

All three authors of the article contributed to the process, each with at least a bachelor's degree (two holding PhDs) and with academic or professional backgrounds in computational linguistics or natural language processing.

6 Experimental Setup

This section outlines the experimental configurations used to instantiate the generative evaluation of BR-LLMs, detailing the datasets, the hardware setup, hyperparameter settings, and model implementation specifics.

6.1 Data

We selected an established dataset well-suited to the respective evaluation objectives for each adopted task. Despite only one dataset being applied per task, we emphasize the thorough evaluation process. The specifics of dataset usage are detailed below.

RecognaSumm [Paiola *et al.*, 2024a] (Summarization) This dataset originates from diverse sources, comprising news articles from various information providers. Such diversity results in a collection of documents covering various topics and journalistic styles. Moreover, RecognaSumm contains approximately 135,000 instances in Portuguese, from which only the columns corresponding to the news text and its summary were selected for this study, with the latter serving as the reference for evaluation metrics. In the dataset's standard split, 27,100 instances are allocated for testing. In this study, we sampled 600 instances to adhere to our cost limitations associated with running the adopted language models. To ensure diversity and broader coverage, a new sampling round was conducted for each execution.

PorSimplesSent [Leal *et al.*, 2018] (Text Simplification) A Portuguese dataset composed of aligned sentence pairs and triplets, originally created to investigate sentence readability assessment in the language. Although its initial purpose was sentence pair classification — where the model must determine whether one sentence is simpler than the other or if both share the same level of simplicity — the *corpus* can be adapted for the text simplification task. For this purpose, only the more complex sentence was considered as the input, while the simpler one served as the target, disregarding cases where both sentences had equivalent simplification levels. After this preprocessing step, as in the previous dataset, the Portuguese instances from the test set were used, ensuring a consistent basis for evaluating the models' ability to perform meaningful text simplification.

Pirá 2.0 [Paschoal *et al.*, 2021; Pirozelli *et al.*, 2024] (QA) This dataset focuses on topics related to the ocean, the Brazilian coast, and climate change. Constructed from scientific abstracts and reports on these subjects, Pirá represents a versatile and highly specialized linguistic resource, ideal for testing the ability of language models to acquire specialized scientific knowledge. It is a bilingual dataset, with instances available in both Portuguese and English. In this study, only the Portuguese test instances were used to evaluate the generative performance of the language models, following the standard dataset split.

6.2 Computational Environment

Each model was executed under the same inference conditions, running independently on an NVIDIA RTX 4090 GPU. Notably, the computational resources were physically located in *Rio de Janeiro*, which influences the reported energy consumption and equivalent CO_2 emissions.

6.3 Model Implementation

The Sabiá and GPT models were accessed through their respective APIs[20]/[21], whereas the openly available models were implemented using the Hugging Face Transformers library [Wolf *et al.*, 2020]. Specific hyperparameters were selected for each task based on insights from the literature, data analysis, and empirical experiments. For summarization, we set `max_new_tokens` to 85 while maintaining `temperature` at 0.3, `top_p` at 0.95, and `repetition_penalty` at 2.5. In the task prompt, the `text_type` parameter was configured as `'News'`. Similarly, for text simplification, we used `max_new_tokens` of 100 but increased `temperature` to 0.8 while keeping `top_p` at 0.95 and `repetition_penalty` at 2.5. Meanwhile, for question answering, we set `max_new_tokens` to 100, `temperature` to 0.3, `top_p` to 0.95, and `repetition_penalty` to 2.5. Additionally, the LLM-as-a-judge evaluation using GPT-4o leveraged the G-Eval framework implemented by DeepEval [Ip and Vongthongsri, 2025]. Finally, to ensure consistency, all generative runs were repeated three times.

[20]Maritaca AI (Sabiá) API: `https://www.maritaca.ai/`
[21]Open AI API: `https://openai.com/index/openai-api/`

Table 1. Computational impact indicators and generative results for summarization on the RecognaSumm dataset.

Model	Rank (Gen.)	Inference Consumption (↓) Time (h)	Energy (kWh)	CO_2eq (kg)	Generative Metrics (↑) BLANC	METEOR	BLEU	ROUGE R1	R2	RL	RLs	BERTScore prec.	recall	f1	G-Eval
Tucano (2.44B params)	4	0.699	0.142	**0.016**	0.059	0.109	0.000	12.750	1.530	8.300	8.560	0.648	0.681	0.664	0.399
Periquito (3.55B params)	8	**0.575**	**0.134**	**0.015**	0.018	0.043	0.000	6.870	0.500	5.160	5.160	0.616	0.636	0.625	0.322
Bode (8.03B params)	6	1.325	0.349	0.040	0.037	0.068	0.000	9.590	0.790	6.870	7.370	0.620	0.656	0.637	0.228
Cabra (8.03B params)	7	1.334	0.352	0.040	0.025	0.057	0.000	8.590	0.670	6.190	6.870	0.609	0.648	0.628	0.225
Boto (9.24B params)	5	1.997	0.515	0.058	0.043	0.074	0.000	10.710	1.160	7.430	7.460	0.649	0.675	0.661	0.306
Sabiazinho-3 (#params unknown)	1	–	–	–	0.243	0.294	**0.055**	**37.010**	**14.910**	**24.650**	**24.660**	**0.721**	0.742	**0.731**	**0.495**
GPT4o-mini (#params unknown)	2	–	–	–	**0.268**	**0.314**	0.051	36.540	14.740	24.080	24.080	0.713	**0.750**	**0.731**	**0.503**
Sabiá-3 (#params unknown)	3	–	–	–	0.242	0.291	0.052	**36.950**	14.690	**24.490**	**24.500**	0.718	**0.741**	0.729	**0.500**
GPT4o (#params unknown)	2	–	–	–	**0.271**	**0.313**	0.053	36.420	14.750	24.020	24.020	0.713	0.749	**0.731**	**0.518**

Table 2. NILC-metrix distances between the LLM generations and the reference summaries on RecognaSumm. Lower is better.

Model	Simplicity	Readability	Morphosyntax	Referential Cohesion	Semantic Cohesion	Syntactic Complexity	Semantic Information
Tucano (2.44B params)	0.929	0.759	1.739	0.653	0.644	**0.386**	0.708
Periquito (3.55B params)	1.026	0.592	1.639	0.641	1.165	0.501	0.748
Bode (8.03B params)	1.046	0.511	1.841	0.582	1.083	0.454	0.706
Cabra (8.03B params)	1.024	**0.472**	1.635	0.750	1.194	0.468	0.740
Boto (9.24B params)	0.991	0.637	1.731	0.557	1.101	0.486	0.734
Sabiazinho-3 (#params unknown)	0.868	0.522	1.536	**0.476**	0.556	0.577	0.720
GPT4o-mini (#params unknown)	**0.792**	0.524	**1.394**	**0.476**	**0.470**	0.461	0.667
Sabiá-3 (#params unknown)	0.833	0.556	1.540	0.626	0.673	0.467	**0.653**
GPT4o (#params unknown)	0.863	0.513	1.570	0.518	0.551	0.439	**0.655**

7 Results

This section presents the experimental results for generative tasks using LLMs, combining automatic evaluation metrics with, when applicable[22], estimates of energy consumption and carbon impact. It also includes findings from human inspection of the generated outputs. **Boldface** in the result tables indicates the best-performing scores, accounting for statistical ties based on significance testing.

7.1 Quantitative Results

Table 1 presents the comparative results for the summarization task on the RecognaSumm dataset. Among the Portuguese-focused models, *Sabiazinho-3* and *Sabiá-3* stand out, with *Sabiazinho-3* achieving the best results across most metrics, while its larger counterpart, *Sabiá-3*, follows closely. This is particularly noteworthy given that *Sabiazinho-3* is a smaller model than *Sabiá-3*. These results suggest that both models are effective for PT-BR summarization, successfully balancing lexical and semantic adequacy.

Notably, the *Sabiá-3* series even outperforms or matches the general-purpose *GPT4o* models, except for the BLANC and METEOR metrics, where the OpenAI solutions achieve slightly better — though not substantially superior — scores. However, a substantial gap persists between proprietary and open-weight models. This is particularly evident in metrics such as BLANC and ROUGE, and although the differences are smaller in BERTScore and G-Eval, they remain distinguishable, with the proprietary models — *GPT* and *Sabiá* — consistently occupying the top ranks.

Among the open-access models, *Tucano* stands out due to its low CO_2 equivalent emissions while also leading in performance across several metrics within its tier. In contrast, *Bode* and *Boto* achieve comparable results in some generative metrics, such as ROUGE and BERTScore, but at the cost of significant energy and time consumption. It is worth noting that both *Bode* and *Boto* have considerably more parameters than *Tucano*, which may partially explain their performance advantage. However, this trend of efficient performance in smaller models does not extend to *Periquito*, which records the lowest scores overall. Given the similarity in size and architecture between *Tucano* and *Periquito*, differences in pretraining *corpora* or strategies may be key factors underlying their divergent outcomes.

Table 2 presents the NILC-Metrix linguistic indicators for the summarization task. Overall, with a few exceptions, the results are relatively close across most models and metrics in

[22] API-based models are excluded from energy and carbon assessments due to lack of hardware and location transparency.

Table 3. Computational impact indicators and generative results for simplification on the PorSimplesSent dataset.

Model	Rank (Gen.)	Inference Consumption (↓) Time (h)	Energy (kWh)	CO_2eq (kg)	Generative Metrics (↑) SARI	METEOR	BLEU	ROUGE R1	R2	RL	RLs	BERTScore prec.	recall	f1	G-Eval
Tucano (2.44B params)	5	0.084	**0.015**	**0.002**	22.376	0.175	0.000	17.760	3.840	14.390	15.170	0.689	0.741	0.713	0.424
Periquito (3.55B params)	7	**0.071**	**0.016**	**0.002**	20.107	0.098	0.000	11.210	1.400	8.860	8.860	0.679	0.714	0.696	0.289
Bode (8.03B params)	8	0.255	0.067	0.008	20.996	0.121	0.000	10.860	1.000	7.840	9.240	0.617	0.698	0.655	0.378
Cabra (8.03B params)	9	0.257	0.068	0.008	20.261	0.096	0.000	8.540	0.670	6.390	6.950	0.614	0.694	0.651	0.222
Boto (9.24B params)	6	0.168	0.036	0.004	20.362	0.114	0.000	13.510	1.890	11.340	11.370	0.713	0.738	0.725	0.448
Sabiazinho-3 (#params unknown)	2	–	–	–	**40.091**	0.547	0.237	59.980	40.300	53.710	53.710	0.870	**0.856**	0.863	**0.841**
GPT4o-mini (#params unknown)	1	–	–	–	**41.104**	**0.563**	**0.261**	**61.990**	**43.330**	**56.630**	**56.630**	**0.882**	**0.858**	**0.869**	**0.814**
Sabiá-3 (#params unknown)	4	–	–	–	37.849	0.487	0.180	54.660	33.580	49.290	49.290	0.861	0.844	0.852	**0.838**
GPT4o (#params unknown)	3	–	–	–	38.710	0.498	0.204	56.650	37.470	51.290	51.290	0.873	0.843	0.857	**0.829**

Table 4. NILC-metrix distances between the LLM generations and the reference simplifications on PorSimplesSent. Lower is better.

Model	Simplicity	Readability	Morphosyntax	Referential Cohesion	Semantic Cohesion	Syntactic Complexity	Semantic Information
Tucano (2.44B params)	0.877	0.583	1.605	0.315	1.162	0.391	0.727
Periquito (3.55B params)	0.898	0.545	1.714	0.233	0.670	0.465	0.740
Bode (8.03B params)	1.109	0.831	2.294	0.658	2.035	0.522	0.750
Cabra (8.03B params)	1.078	0.861	1.934	0.433	1.704	0.521	0.721
Boto (9.24B params)	1.048	0.487	1.526	0.167	0.332	0.414	0.760
Sabiazinho-3 (#params unknown)	**0.701**	**0.316**	**0.920**	**0.000**	0.122	**0.328**	0.555
GPT4o-mini (#params unknown)	0.782	0.343	1.124	**0.000**	**0.079**	0.418	0.566
Sabiá-3 (#params unknown)	0.802	0.441	0.950	0.000	0.112	0.384	**0.552**
GPT4o (#params unknown)	0.910	0.408	0.999	**0.000**	0.092	0.404	0.545

dimensions such as simplicity, readability, and morphosyntax. However, more pronounced differences emerge in the semantic cohesion dimension, where *GPT4o-mini* leads, followed by its larger counterpart, the *Sabiá* models, and *Tucano*. The remaining open-weight models trail further behind. Additionally, there is a recurring trend toward proprietary models maintaining a slight dominance in dimensions such as semantic information, simplicity, and morphosyntax. Still, this tendency is not absolute: *Cabra* achieves the top score in readability, and *Tucano* outperforms all others in the syntactic complexity indicator.

Table 3 shows that *GPT4o-mini*, closely followed by *Sabiazinho-3*, achieves the highest overall scores across most generative metrics for sentence simplification on the PorSimplesSent dataset. In contrast, smaller open-weight models such as *Tucano*, *Periquito*, and *Boto* register noticeably lower scores across several generative metrics. Particularly striking are the differences in the simplification-specific indicator SARI and in the LLM-powered G-Eval evaluation and the traditional BLEU metric, where the gap between proprietary and open models can reach up to double the score. Nevertheless, certain open models show distinct results within their class. *Tucano*, in particular, leads among open-weight models in generative quality while maintaining consistent efficiency. Once again, *Boto* follows in terms of generative performance, though with higher resource consumption.

Similarly, the sentence simplification task on the PorSimplesSent dataset reveals a clear stratification in model performance, as detailed in Table 3. *GPT4o-mini* achieves the highest scores across most generative metrics, with *Sabiazinho-3* following closely. Interestingly, these smaller models consistently outperform their larger versions, indicating that scale is not always the decisive factor in quality. On the other hand, smaller open-weight models — including *Tucano*, *Periquito*, and *Boto* — record significantly lower scores. The contrast is especially evident in SARI, BLEU, and G-Eval, where the gap can reach up to 100%. Nonetheless, performance within the open-weight category is not homogeneous. *Tucano* demonstrates better generative capabilities with high efficiency, while *Boto* offers a similar output quality at the expense of greater computational demand. In contrast to the summarization task, however, *Periquito* ranks as the third-best simplifier among open-weight models, whereas *Cabra* consistently underperforms, emerging as the weakest model overall in this scenario.

Further observations are presented in Table 4, which reports NILC-metrix distances between generated and reference simplifications. Once again, *Sabiazinho-3* and *GPT4o-mini* stand out, achieving the lowest distances in dimensions such as simplicity, readability, and morphosyntax. Although

Table 5. Computational impact indicators and generative results for question answering on the Pirá dataset.

Model	Rank	Inference Consumption (↓)			Generative Metrics (↑)										
		Time	Energy	CO_2eq	F1	METEOR	BLEU	ROUGE				BERTScore			G-Eval
	(Gen.)	(h)	(kWh)	(kg)				R1	R2	RL	RLs	prec.	recall	f1	
Tucano (2.44B params)	8	**0.091**	**0.017**	**0.002**	0.057	0.122	0.001	9.390	1.280	6.700	7.260	0.596	0.676	0.633	0.167
Periquito (3.55B params)	9	**0.090**	0.021	**0.002**	0.047	0.092	0.000	8.060	0.720	5.570	5.570	0.606	0.682	0.641	0.126
Bode (8.03B params)	6	0.161	0.043	0.005	0.063	0.114	0.002	10.250	1.720	7.720	8.270	0.607	0.682	0.642	0.221
Cabra (8.03B params)	5	0.150	0.039	0.004	0.075	0.123	0.003	11.430	2.000	8.660	8.900	0.624	0.690	0.655	0.170
Boto (9.24B params)	7	0.279	0.061	0.007	0.064	0.120	0.002	10.140	1.640	7.570	7.980	0.606	0.683	0.641	0.159
Sabiazinho-3 (#params unknown)	3	–	–	–	**0.123**	0.189	0.010	16.040	4.760	12.640	12.830	0.636	0.711	0.671	**0.300**
GPT4o-mini (#params unknown)	2	–	–	–	**0.125**	0.197	0.010	16.440	5.450	13.140	13.510	0.641	**0.720**	0.678	**0.334**
Sabiá-3 (#params unknown)	4	–	–	–	0.120	0.189	0.008	15.830	4.530	12.310	12.480	0.630	0.710	0.667	**0.297**
GPT4o (#params unknown)	1	–	–	–	**0.128**	**0.208**	**0.013**	**17.300**	**5.880**	**13.980**	**14.150**	**0.644**	**0.723**	**0.681**	**0.360**

Table 6. NILC-metrix distances between the LLM generations and the reference answers on Pirá. Lower is better.

Model	Simplicity	Readability	Morphosyntax	Referential Cohesion	Semantic Cohesion	Syntactic Complexity	Semantic Information
Tucano (2.44B params)	1.090	0.782	2.068	0.686	2.217	0.565	0.983
Periquito (3.55B params)	**0.995**	0.803	1.968	0.776	**1.471**	**0.551**	0.995
Bode (8.03B params)	1.084	0.857	2.182	**0.532**	1.979	0.560	1.013
Cabra (8.03B params)	1.076	0.790	2.040	0.824	2.028	0.626	1.075
Boto (9.24B params)	1.221	0.894	1.970	0.572	1.724	0.724	1.077
Sabiazinho-3 (#params unknown)	1.150	0.761	**1.745**	1.112	1.607	0.720	0.940
GPT4o-mini (#params unknown)	1.073	0.795	1.856	0.828	1.812	0.752	**0.934**
Sabiá-3 (#params unknown)	1.170	0.837	2.131	0.740	1.717	0.857	2.131
GPT4o (#params unknown)	1.039	**0.747**	1.921	0.751	1.569	0.752	0.966

smaller open-weight models occasionally approach the performance of larger ones in certain aspects of cohesion or syntactic complexity, they generally yield slightly higher distances across metrics. A noteworthy observation is that, unlike the summarization task (Table 2), all models tend to perform better in referential and semantic cohesion indices. Remarkably, the *Sabiá* and *GPT* models often reach near-optimal or even perfect alignment in these dimensions. However, such differences are likely a consequence of the task design — summarization entails constructing broader textual structures that integrate multiple propositions, whereas, on PorSimplesSent dataset, simplification typically operates at the sentence level.

Table 5 presents the QA outcomes on the Pirá dataset. Once again, the top ranks are dominated by proprietary models from the *GPT4o* and *Sabiá-3* families. *GPT4o* leads the question answering task, with *GPT4o-mini* following closely in second place. *Sabiá-3* ranks just below its smaller counterpart, *Sabiazinho-3*, completing the top four. In this task, the open-access model that stands out as the best generative performer is *Cabra*, followed by *Bode* and *Boto*, respectively. Notably, despite their efficiency, *Tucano* and *Periquito* did not appear among the top results this time — even within the open-weight category. Taken together with previous results (Tables 1 and 3), this outcome suggests that the performance of open-access solutions for Portuguese generation remains variable across tasks.

Although the proprietary *Sabiá* and *GPT* models perform slightly better than open solutions in this task, the overall lower scores in metrics such as F1 and G-Eval are consistent with the findings of [Pirozelli *et al.*, 2024], which highlight that question answering in settings relying solely on internally encoded information within LLMs — without access to external context or tools — is particularly challenging, especially in specialized domains.

Turning to the NILC-metrix results for QA answers in Table 6, we observe a notably different pattern in this case. Although, once again, the overall results across models are relatively close, within this task, open-weight models occasionally lead in specific criteria — *Bode* in referential cohesion, and *Periquito* in simplicity and semantic cohesion. That said, despite more reference-aligned scores in dimensions such as syntactic complexity, readability, and referential cohesion, higher distances — often approaching a value of 2 — are more frequently observed in semantic indicators.

Figure 2 offers a visual overview of both generative performance and computational impact across the evaluated tasks. The chart echoes earlier findings: the *Sabiá* and *GPT* model families consistently occupy the top positions, while a gap remains between them and open-weight alternatives.

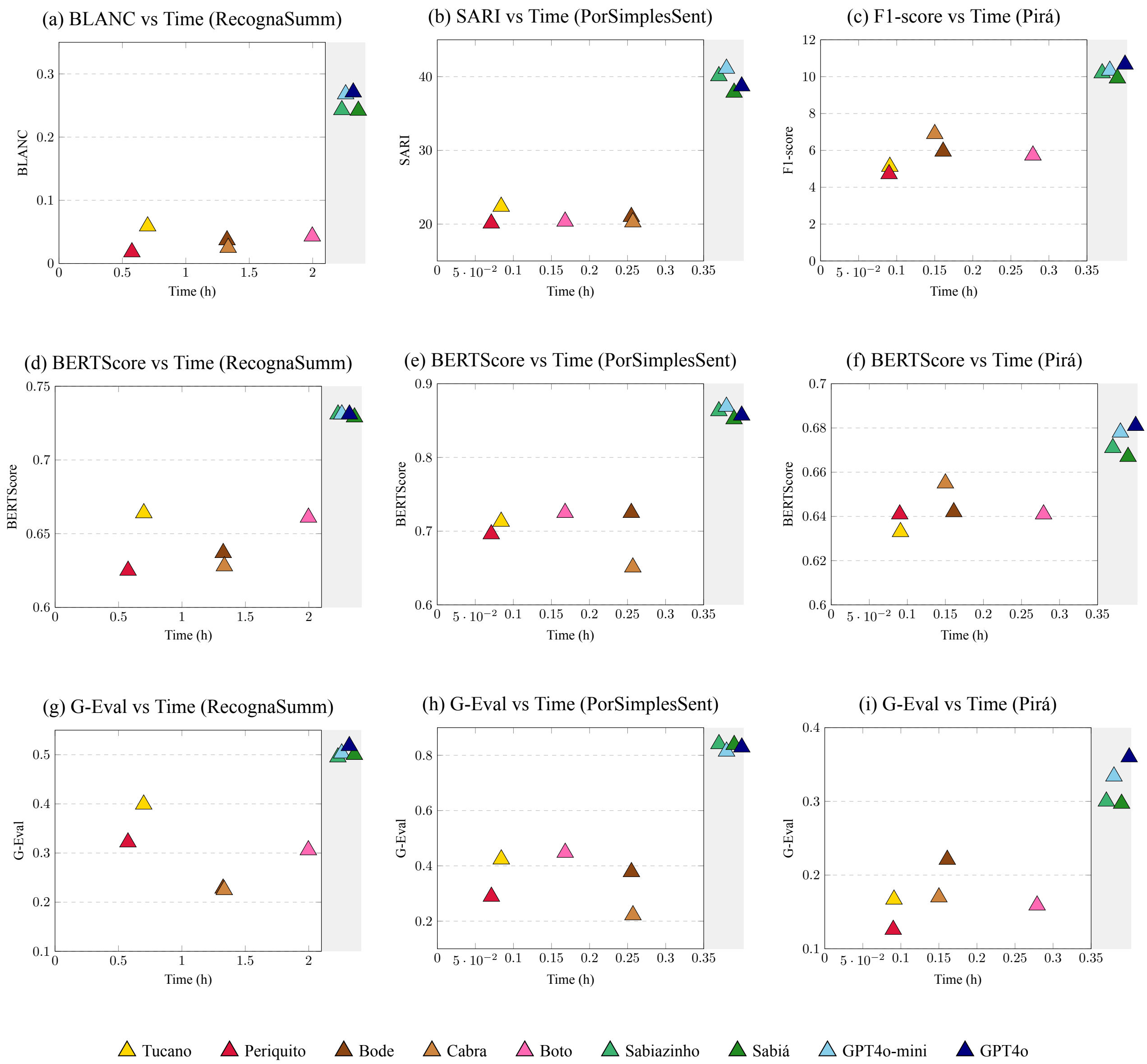

Figure 2. Comparative performance in relation to Time and Generative Metrics for the RecognaSumm, Pirá, and PorSimples datasets. Commercial models accessed via API and run on external hardware are shown in the shaded area, as their runtime is not directly comparable.

Among the open models, performance varies considerably across tasks, with *Tucano*, *Bode*, *Cabra*, and *Boto* alternating among themselves as the top performers within their category. Despite these fluctuations, *Tucano* consistently balances moderate performance with lower computational footprint, positioning it as an attractive open-weight solution.

By comparing performance on classification and more discriminative tasks — as reflected in the Open PT LLM Leaderboard (OPLL) [Garcia, 2024] — with results obtained in this research, Table 7 provides notable observations. Although superior performance on discriminative tasks may occasionally indicate strong generative capabilities, as demonstrated by *Sabiá-3* and *GPT4o-mini*, this correlation is not absolute. Rankings can vary significantly across tasks. A clear example is *Tucano*, which ranks second-to-last among seven models on discriminative tasks but excels in generative evaluations, trailing only *Sabiá-3* and *GPT4o-mini*, and surpassing *Boto*, *Bode*, and *Cabra*. This also reinforces the observation that a model's parameter count does not consistently correlate with generative task performance, considering that *Tucano* is smaller than many of its counterparts.

Even among generative tasks, performance rankings vary considerably. For instance, models positioned just below the top tier exhibit fluctuating performance depending on the specific generative task. *Tucano*, despite performing strongly in summarization and simplification, ranks second-to-last in QA. Conversely, *Cabra* leads among open-weight models in QA but falls to the bottom ranks in other generative tasks.

Overall generative rankings place *GPT4o-mini* at the forefront, followed closely by a tie between *GPT4o* and *Sabiazinho-3*, with *Sabiá-3* in third position. Subsequently, *Tucano* stands out within the open-weight group, being followed by *Boto*. These findings underscore the importance of directly evaluating generative capabilities, as model suitability may differ substantially across different evaluation scenarios.

Extending beyond the results provided by automatic NLG metrics, Table 8 includes the average number of words in both the generated texts and their references, offering a complementary perspective on model behavior across tasks. These indicators generally show that LLMs produce longer texts compared to their references, except for the proprietary *Sabiá-3* and *GPT4o* model families in the simplification task.

Such verbosity indicates a misalignment between the generated outputs and the expectations established by the reference texts. However, this discrepancy may stem from several factors. For instance, in the context of question answering — where lower performance was previously observed — this misalignment may indicate that models are deviating from the reference answers by including extraneous or inappropriate information. Conversely, it might also reflect that models produce paraphrased responses that are, in certain cases, more detailed or comprehensive than the references — a scenario that was corroborated by some manual inspections. These aspects are further analyzed through human assessments in the following section.

Table 7. Comparison of rankings between the Open PT LLM Leaderboard (OPLL) and those from the evaluated tasks, including the sum of ranks for generative tasks (G-Sum) and their corresponding order (G-Rank).

Model	OPLL	Summ.	Simp.	QA	G-Sum	G-Rank
Sabiá-3 (#params unknown)	1	3	4	4	11	3
GPT4o-mini (#params unknown)	2	2	1	2	5	1
Boto (9.24B params)	3	5	6	7	18	5
Bode (8.03B params)	4	6	8	6	20	6
Cabra (8.03B params)	5	7	9	5	21	7
Tucano (2.44B params)	6	4	5	8	17	4
Periquito (3.55B params)	7	8	7	9	24	8
Sabiazinho-3 (#params unknown)	–	1	2	3	6	2
GPT4o (#params unknown)	–	2	3	1	6	2

Table 8. Average Word Count per Generative Task.

Souce	Summarization	Simplification	Question Answering
Reference	36.192	18.095	14.401
Tucano	62.880	39.537	77.866
Periquito	38.252	33.214	69.665
Bode	40.042	56.255	53.655
Cabra	42.993	57.109	49.446
Boto	46.428	26.312	66.740
Sabiazinho-3	47.613	16.158	53.126
GPT4o-mini	57.135	14.953	56.332
Sabiá-3	47.171	15.516	54.617
GPT4o	57.449	13.674	54.659

7.2 Human Inspection

This section discusses human perceptions regarding the outputs generated by the LLMs. It also includes a human review of the automatic LLM-based metric used to evaluate generations in Portuguese. Lastly, it discusses how these human insights relate to the quantitative results obtained from traditional NLG metrics, as reported in Section 7.1.

7.2.1 Generated Texts Inspection

Figure 3 presents the human ratings for all criteria covered by the LLM-based G-Eval metric, aiming to complement the automatic evaluation conducted using traditional and recent NLG indicators. First, it shows that the trend observed in previous quantitative results — where proprietary models *Sabiá-3* and *GPT4o* stand out — also aligns with human perception. These models receive the majority of their ratings as 5 or 4 on the Likert scale (the highest and second-highest possible scores, respectively), while the remaining models more frequently receive scores around 2 or even as low as 1.

Furthermore, a human interpretation that slightly diverges from the quantitative results concerns the *Sabiá-3* model and its smaller variant, *Sabiazinho-3*. The larger version more frequently received higher ratings than the smaller one — and never received scores below 3, unlike the latter — even though, in the quantitative results, the smaller version ranked higher. This difference is particularly notable in the QA task, where the larger *Sabiá-3* stands out overall, even outperforming *GPT-4o*. In this specific context, the result may suggest that the specialized information required to answer questions in the Pirá dataset — rooted in Brazilian-specific content — is more effectively encoded in the larger Brazilian model, a nuance that may not have been fully captured by the automatic metrics.

As a general negative highlight, the *Cabra* model concentrates a large number of ratings of 1 assigned to its generations across tasks. A common observation made by the annotators in their comments is the presence of spelling and grammatical agreement errors in the outputs produced by this model. The following examples illustrate some cases.

Cabra Error Examples

> **Summarization**
> *Um grupo dos cidadãoSobrevivendo à Guerreiros nde Demidov (Ucraína) criarà-lhes artificialmenteeles trancando portonas [...]*
>
> **Simplification**
> *Portando isso para videogames (jogar em um computador ou console), androidezinhão/robota chatona que falam apenas com você sozinhas [...]*
>
> **Question Answering**
> *A formação de jandaíras esta localizada na regiã oceântica aberta da plataforma continental sul-atlantICA. [...]*

Still referring to Figure 3, *Tucano* is the open model that most consistently receives human ratings on the favorable end of the spectrum across tasks — though its scores are not as high as those of the proprietary *Sabiá* and *GPT* models, and it is not always the very top performer among open models. Exceptions include the criteria of simplicity and fluency, where it is outperformed by *Boto*, as well as the QA task, which proved challenging for all models. Importantly, *Tucano* is also among the most computationally efficient models. However, even this model is not exempt from incon-

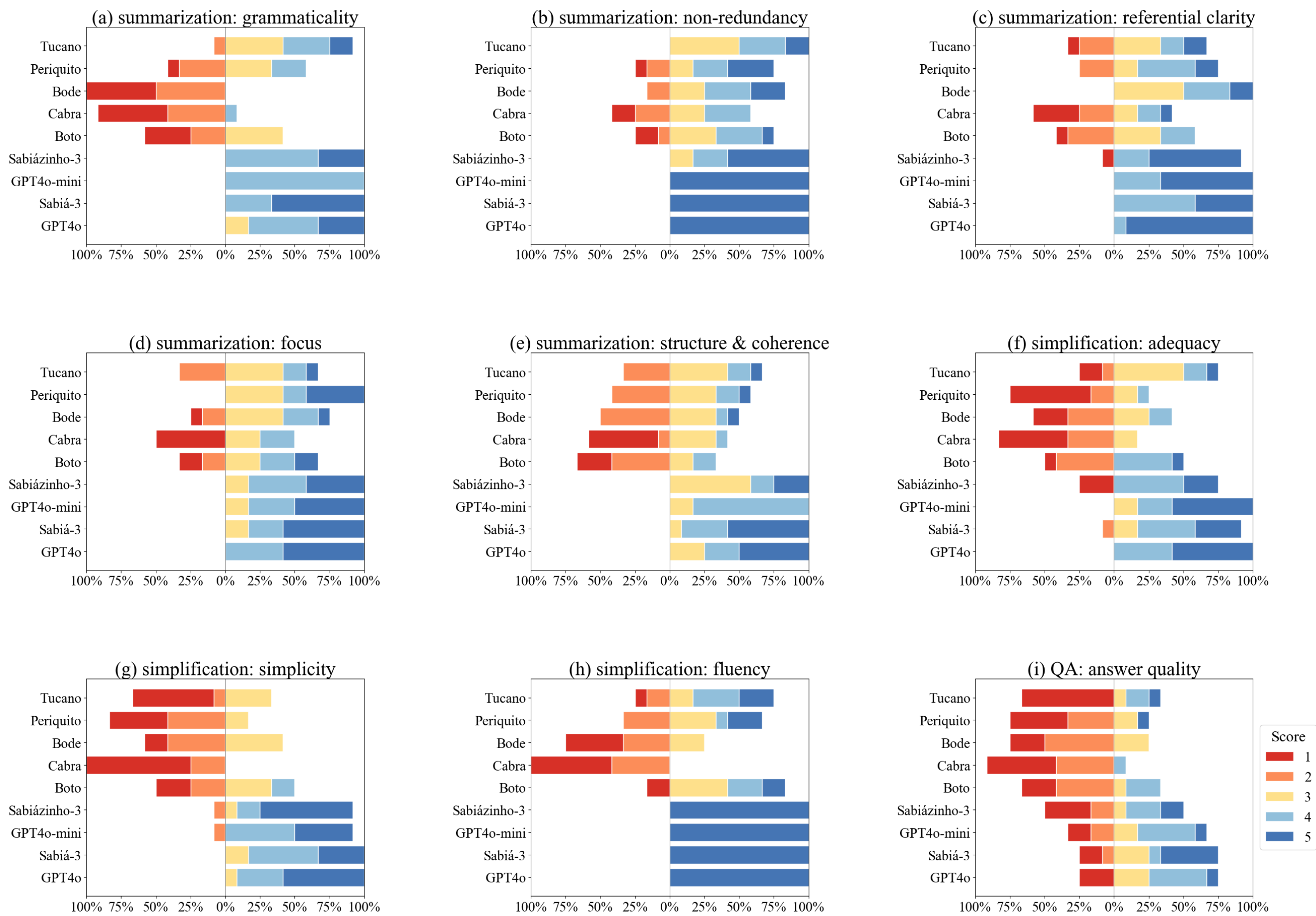

Figure 3. Combined Likert-scale ratings (1–5) for each model across summarization (a–e), simplification (f–h), and QA (i) criteria.

sistencies in its generations, with frequent reports of out-of-context information, distortions, or hallucinations, as illustrated in the simplification example below.

Tucano Error Example

Reference
Há 75 vagas para um universo estimado em pelo menos mil moradores de rua.

Tucano
Existem <u>750</u> lugares num Universo Estimado que tem cerca dos 1000 residentes sem-abrigo. [...]

For all models, most of the human ratings on the negative scale (scores 1 and 2) were accompanied by comments pointing to grammatical, spelling, agreement, or factual inconsistencies. Another frequent observation concerned the tendency of LLMs to produce long outputs.This pattern often caused the generations to exceed the configured token limit, resulting in incomplete sentences. Such outputs frequently led to inconsistencies that negatively impacted the evaluated criteria, mainly in summarization and QA.

This increased length was even more evident in the QA task (as shown in Table 8), where annotators frequently left comments such as "*The answer strays from the question*" or "*The response lacks substance and is overly wordy.*" However, in other cases, a different pattern emerged: the model-generated responses often paraphrased the references or included more specific and detailed information (and were therefore longer). While still correct and appropriate, these responses were penalized by the automatic metrics due to their misalignment with the reference texts.

Example of a Longer Generation that is Appropriate

Question (English-translated)
Who are the people who benefit the most from a fish-based diet?

Reference Answer (English-translated)
Poor rural populations.

GPT-4o Answer (English-translated)
People living in coastal communities, fishers, and populations that rely on fishing as their main source of protein and livelihood are those who benefit the most from a fish-based diet. Additionally, regions with limited access to other sources of protein may also depend significantly on fish.

The example above illustrates a case where the generated response correctly answers the given question. Specifically, we consulted the dataset context and verified its correctness. However, the reference answer originally present in the dataset is extremely concise — arguably excessively so

— and lacks relevant details. It could also be argued that the reference answer is inaccurate, as the context makes no explicit mention of rural areas, despite referring to regions with limited resource access. The automatic evaluator assigned this response a score of 0.280, whereas one human annotator gave it a 4, and another even assigned the maximum score.

This further underscores the challenges inherent in the evaluation of open-ended text generation. Our inspection of the model outputs also reveals instances in which the generated texts complement the existing benchmark references. The example below, drawn from the simplification task, illustrates such a case. While the reference transforms the term "`zoning`" into "`division of the territory into zones`", the automatic generation preserves the original term but proposes a valid simplification ("`represents the first step in shaping`" → "`marks the beginning of`"). Thus, it becomes evident that, while reference texts in evaluation datasets may serve as useful guides, they should not always be treated as absolute "*ground truths*", as is often the case in more discriminative benchmarks.

Example of a Complementary Simplification[23]

> **Original Sentence (English-translated)**
> *Zoning represents the first step in shaping a forest policy in Rio Grande do Sul.*
>
> **Original Reference (English-translated)**
> *The division of the territory into zones represents the first step in preparing a forest policy in Rio Grande do Sul.*
>
> **GPT-4o Generation (English-translated)**
> *Zoning marks the beginning of a forest policy in Rio Grande do Sul.*

As a final observation, overall, human evaluators showed a preference for the generations produced by *Sabiá-3* and *GPT-4o*, with their smaller variants ranking just behind, completing the group of top performers. In contrast, open-weight models often displayed perceptible flaws, as previously discussed, indicating that further refinement is needed for these LLMs to reach the level of the proprietary group.

7.2.2 Human Agreement with the LLM-based Metric

We leveraged the human inspection not only to evaluate the model generations for the tasks in Portuguese, but also to assess the performance of the LLM-as-a-judge approach implemented using the G-Eval framework. Table 9 presents the average agreement between human annotators and the LLM judge's reasoning for each criterion, along with the percentage of inter-annotator agreement specifically regarding the model's reasoning and its corresponding score. Annotators were asked to evaluate both aspects, as the automatic judge may, for instance, provide a reasonable justification but assign a final score that is entirely inconsistent, even with its own reasoning.

[23]Translation in Portuguese in Appendix H

Table 9. Average agreement with G-Eval reasoning, along with Inter-Annotator Agreement (IAA) percentages for reasoning and final score alignment.

Criteria	**AVG Agreement w/G-Eval**	**Reason *(IAA)***	**Score *(IAA)***
summarization:grammaticality	.833	.694	.833
summarization:non-redundancy	.657	.638	.750
summarization:referential clarity	.639	.638	.917
summarization:focus	.648	.500	.667
summarization:structure & coherence	.472	.556	.750
simplification:adequacy	.852	.722	.778
simplification:simplicity	.787	.639	.667
simplification:fluency	.722	.472	.639
question answering: answer quality	.843	.639	.694

Moreover, percentage agreement values are reported as indicators of consistency, given that the number of fully examined samples may occasionally lead to unstable agreement coefficients, with compromised reliability. This metric still provides a signal of consistency for future scenarios that may include annotators drawn from a population with backgrounds similar to those considered in this work, specifically a more specialized group of annotators [Artstein and Poesio, 2008]. Lastly, the sampling strategy employed, as described in Section 5.3.5, also ensures broad coverage of the metric's evaluations.

A general alignment is observed when analyzing the average human agreement with the reasoning provided by G-Eval. Notably, we highlight that agreement values in less objective tasks are typically, and naturally, lower than those observed in more objective ones [Reidsma and op den Akker, 2008]. A very low agreement value appears, however, in the summarization criterion of structure & coherence.

In this case, the LLM judge occasionally penalizes summaries for lacking an explicitly segmented structure, *e.g.*, an introduction, body, and conclusion. While such an organization could be valuable in certain contexts, it is not a stated requirement of the evaluation criterion and appears to reflect an assumption made exclusively by the evaluator model. Notably, this expectation is also not reflected in the reference summaries.

Furthermore, annotator comments frequently indicate that the judge model tends to over-penalize incomplete generations (likely due to token limits), which in turn affects its assessments across multiple criteria. An illustrative example is provided below, where the model evaluates the output as lacking focus in the overall text. The negative aspect raised by the model is the abrupt ending. However, human annotators noted that this could be acceptably resolved with the insertion of a final period, preserving a coherent idea overall. If penalized at all, the issue would be more appropriately categorized under other criteria, such as grammaticality. Nonetheless, the metric assigned a final score of 0.3 to the focus criterion.

G-Eval Focus Judgment

GPT-4o Summary (Portuguese)
O atacante brasileiro Vinícius Júnior foi novamente vítima de racismo durante uma partida contra o Valencia no estádio Mestalla, somando onze casos de racismo sofridos na Espanha em menos de dois anos. O episódio mais recente levou à paralisação do jogo e à expulsão de Vinícius. Ao longo desse período, o jogador enfrentou insultos racistas em diversas ocasiões, incluindo partidas contra Barcelona,

GPT-4o Summary (English-translated)
The Brazilian forward Vinícius Júnior was once again the target of racism during a match against Valencia at the Mestalla stadium, bringing the number of racist incidents he has suffered in Spain to eleven in less than two years. The most recent episode led to the suspension of the match and to Vinícius being sent off. Over this period, the player has faced racist abuse on several occasions, including matches against Barcelona,

G-Eval Reason
The Actual Output highlights recurring incidents of racism suffered by Brazilian player Vinícius Júnior during matches in Spain, with the most recent taking place against Valencia at the Mestalla stadium. However, the summary ends abruptly and lacks a proper conclusion.

The fluency criterion in the simplification task exhibited a higher degree of disagreement among annotators regarding the reasoning provided by G-Eval. According to annotator comments, one evaluator did not penalize the inclusion of additional observations on certain aspects, while the other two considered these remarks more relevant to different criteria and therefore evaluated them more critically.

Similarly, a comparable pattern was observed for the focus criterion in the summarization task. Two annotators frequently noted in their comments that the main focus of the generation could still be identified in many cases, even when some lexical or grammatical inconsistencies were present. In contrast, the third annotator more often agreed with the judge model's rationale that such errors could hinder the evaluation of focus. These cases highlight how subjective interpretations can influence not only machine-based evaluations but also human judgment in the assessment of text generation.

Conversely, human evaluation also showcases the positive capabilities of the automatic evaluation framework. Grammatical errors and adequacy problems are frequently identified correctly by the LLM-based judge, as evidenced by the high alignment in these criteria and the corresponding inter-annotator agreement. On the other hand, human annotators noted that similar errors are not always scored consistently, as illustrated in the following example. Both texts associated with these reasons received a score of 5 from all human annotators.

G-Eval Grammaticality Judgment

G-Eval Reason on Summary A
The Actual Output mostly follows the rules of capitalization and syntax but has a sentence fragment causing incomplete information at the end.
→ G-Eval Score = 0.400

G-Eval Reason on Summary B
Most sentences in the Actual Output start with a capital letter and are grammatically correct. However, the last sentence is incomplete, possibly due to an abrupt ending, affecting sentence structure and overall completeness.
→ G-Eval Score = 0.827

These overall challenges align with findings in the literature showing that evaluations performed by LLMs are not free from limitations and biases [Chen *et al.*, 2024]. In this regard, it becomes essential not to rely exclusively on such metrics, but rather to complement them with additional indicators that can address these shortcomings. These results also suggest that evaluation becomes more robust when applied to larger volumes of text, thereby grounding the findings in statistical relevance. Both of these aspects were addressed in this research.

7.2.3 Traditional NLG Metrics & Human Feedback

While human annotators were not explicitly instructed to assess scores provided by traditional NLG metrics beyond the LLM-judge-based method, it remains valuable to explore how well these metrics align or not with human judgments. Generally, automatic metrics underscored a clear distinction between open-weight and proprietary solutions, a trend confirmed by human perceptions. However, within each group, these metrics did not reveal substantial internal variations. Such a pattern might initially suggest similar model strengths; nevertheless, when considered alongside human judgments, they likely indicate the limited sensitivity of automatic metrics in some contexts, as previously noted in previous research [Freitag *et al.*, 2023, 2024].

In particular, the NILC-Metrix indicators could be interpreted as suggesting that all evaluated models demonstrate comparable lexical-syntactic strengths, even though semantic discrepancies remain evident. Human evaluations, on the other hand, clarified that models scoring similarly to top-ranked approaches according to these metrics still exhibited noticeable shortcomings in their generated texts. Therefore, while NILC-Metrix may be effective at detecting the presence of lexical and syntactic structures, caution is warranted in interpreting its scores, as the metric does not capture how these structural components are articulated.

In general, this comparison between human evaluations and automatic metrics highlights the inherent limitations of relying solely on automated measures. Consequently, these metrics should be interpreted with caution and used in complementary combinations, as explored in this work, to better account for and mitigate their individual shortcomings. Furthermore, human evaluation remains an important component in this process.

8 Conclusion

Addressing the central question posed — *What is the generative performance of LLM-powered solutions specifically designed for Brazilian Portuguese?* — this work concludes that performance can vary depending on the task and the specific Brazilian LLM adopted. Specifically, we evaluated six families of Brazilian models, including *Sabiá-3* and its smaller counterpart *Sabiazinho-3*, *Bode-3.1-8B*, *CabraLlama-8B*, *Boto-9B*, *Tucano*, and *Periquito*. In addition, we included the *GPT* family representatives, *GPT4o* and *GPT4o-mini*, as state-of-the-art generalist baselines. We designed an evaluation framework involving NLG tasks, namely summarization, simplification, and generative question answering, as proxies to assess key criteria related to the generative capabilities of these models.

Drawing on a broad set of lexical, morphosyntactic, and semantic metrics, as well as incorporating the recent LLM-as-a-judge paradigm and human evaluation, this work contributes valuable insights into the current generative capacity of Brazilian LLM solutions and into evaluation practices, particularly in a context that has thus far been centered around large-scale, however, discriminative benchmarks. Our findings demonstrate that previously reported results from discriminative tasks in the literature do not directly map to generative performance. This reinforces the importance of evaluating specifically generative aspects, as done in this research, for developing robust conversational applications.

Moreover and perhaps unsurprisingly, our results indicate that the *GPT4o* models — with outstanding performance by the mini version — extend their well-established general-purpose strengths to Portuguese generative tasks, despite not being explicitly designed for this language. Furthermore, they are closely followed by *Sabiá-3* and *Sabiazinho-3*, which stand out as the best-performing Brazilian models in generative terms. Nevertheless, all of these are proprietary solutions. Among the openly available models, *Boto* and *Tucano* emerge as the most competitive, with *Tucano* further distinguished as a viable option in terms of environmental impact and computational efficiency. With 2.4 billion parameters, it consistently ranks among the most efficient models while at times matching the performance of the 9B-parameter *Boto*. Despite these promising outcomes, we observed a tendency among smaller and open-weight models to struggle with generalization across the various evaluated dimensions. While proprietary solutions show little fluctuation in top-tier rankings, open models such as *Tucano* and *Cabra* vary more substantially.

Curating high-quality Portuguese datasets across diverse tasks and domains is a crucial step toward empowering open national solutions to reach commercial-level performance. Given that most national advancements currently rely on adjusting pre-trained foreign models, an open question arises: to what extent could fully developing models tailored to national contexts better address the particularities and demands of local applications? This remains as an interesting research and technological avenue, yet one that is highly costly to pursue and verify. Moreover, it is important to acknowledge that while general-purpose solutions can be effective, they are not the sole viable option. It is our view that smaller models exhibiting high performance in specialized domains and tasks also constitute a valuable alternative. For instance, *Tucano* shows strong results in summarization and simplification tasks, yet ranks second-to-last in QA. Conversely, *Cabra* leads among open-weight models in QA but remains less competitive in other generative tasks.

In addition to evaluating model generations, we also conducted a human inspection of the judgment produced by the LLM-powered metric *G-Eval*, using *GPT-4o* as its backbone. We highlight that despite showing promising and useful results, particularly for enabling large-scale evaluation, LLM-as-a-judge frameworks still present limitations regarding adherence to the intended criteria and, ideally, should not be used in isolation. We further explored how traditional NLG metrics relate to human judgments, noting their limitations in detecting fine-grained distinctions among model outputs. Our results underscore the importance of a robust, multi-indicator evaluation strategy.

In future work, we intend to expand the scope of human evaluation, both by increasing the number of annotators and by examining a broader set of instances. These aspects are acknowledged as current limitations and we emphasize that a more comprehensive and expressive evaluation is essential to better align model assessments with end-user preferences. We further emphasize the importance of revising current benchmarks to reflect the capabilities of modern LLMs, especially given the complexities involved in open-ended generation. Complementarily, a future path involves releasing the structured evaluation process as a framework, enabling its replication in the assessment of generative capabilities across models. Additionally, in light of the limited generalization observed in open-weight models across multiple criteria, we suggest that efficient fine-tuning represents a promising strategy to improve performance in applications that rely on *current* models, as previously demonstrated in other Portuguese-language contexts [Assis *et al.*, 2024a]. Hybrid approaches that integrate and leverage the individual strengths of these models, such as agent-based systems [Masterman *et al.*, 2024], also offer a promising avenue to narrow the gap between open-weight and proprietary solutions.

Declarations

Authors' Contributions

Gabriel Assis was the primary contributor to writing the original draft and led the efforts in conceptualization, data curation, formal analysis, investigation, methodology, validation, and visualization. **Cláudia Freitas** contributed to conceptualization, methodology, and validation. **Aline Paes** contributed to review and editing, and also in conceptualization, methodology, supervision, validation, and funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors declare the following competing interests: One of the authors of this manuscript also served as a guest editor for the special issue. In accordance with the agreement established with the editors-in-chief, this submission was handled exclusively by the other three guest editors. The author who served as a guest editor had no involvement in or access to the review process, ensuring an impartial and transparent evaluation.

Funding

This research was financed by CNPq (National Council for Scientific and Technological Development), FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also acknowledge the support of the Maritaca AI credits program, which enabled access to Sabiá models.

Availability of data and materials

The code used to conduct all experiments is publicly available at `https://github.com/MeLLL-UFF/brfauna-gen-eval`. The authors will provide any additional resources or clarifications upon request.

References

Aakanksha, Ahmadian, A., Ermis, B., Goldfarb-Tarrant, S., Kreutzer, J., Fadaee, M., and Hooker, S. (2024a). The multilingual alignment prism: Aligning global and local preferences to reduce harm. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.671.

Aakanksha, Ahmadian, A., Goldfarb-Tarrant, S., Ermis, B., Fadaee, M., and Hooker, S. (2024b). Mix data or merge models? optimizing for performance and safety in multilingual contexts. In *Neurips Safe Generative AI Workshop 2024*. Available at: `https://openreview.net/forum?id=L1Hxp8ktiT`.

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 Technical Report. Available at: `https://arxiv.org/abs/2410.12049`.

Al-Thanyyan, S. S. and Azmi, A. M. (2021). Automated Text Simplification: A Survey. *ACM Comput. Surv.*, 54(2). DOI: 10.1145/3442695.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models. DOI: 10.48550/arxiv.2311.16867.

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A New Generation of Portuguese Large Language Models. Available at: `https://arxiv.org/abs/2403.09887`.

Almeida, T. S., Laitz, T., Bonás, G. K., and Nogueira, R. (2023). Bluex: A benchmark based on brazilian leading universities entrance exams. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 337–347, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45368-7$_2$2.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. (2019). EASSE: Easier Automatic Sentence Simplification Evaluation. In Padó, S. and Huang, R., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-3009.

Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, J. Z., Fredrikson, M., Gal, Y., and Davies, X. (2025). AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. In *The Thirteenth International Conference on Learning Representations*. Available at: `https://openreview.net/forum?id=AC5n7xHuR1`.

Artstein, R. and Poesio, M. (2008). Survey article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. DOI: 10.1162/coli.07-034-R2.

Assis, G., Vasconcelos, A., de Azevedo, L., Ferro, M., and Paes, A. (2024a). Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 97–107, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/stil.2024.245362.

Assis, G., Vianna, D., Pappa, G. L., Plastino, A., Meira Jr, W., da Silva, A. S., and Paes, A. (2024b). Analysis of material facts on financial assets: A generative AI approach. In Chen, C.-C., Liu, X., Hahn, U., Nourbakhsh, A., Ma, Z., Smiley, C., Hoste, V., Das, S. R., Li, M., Ghassemi, M., Huang, H.-H., Takamura, H., and Chen, H.-H., editors, *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 103–118, Torino, Italia. Association for Computational Linguistics. Available at: `https://aclanthology.org/2024.finnlp-1.11`.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., and Sutton, C. (2021). Program synthesis with large language models. Available at: `https://arxiv.org/abs/2108.07732`.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. (2023). Qwen technical report. DOI: 10.48550/arxiv.2309.16609.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. (2024). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403. DOI: 10.48550/arXiv.2406.18403.

BotBot-AI (2023). Cabra. Available at: `https://huggingface.co/botbot-ai/Cabra`.

BotBot-AI (2024a). Cabra Llama-3 8B. Available at: `https://huggingface.co/botbot-ai/CabraLlama3-8b`.

BotBot-AI (2024b). CabraMistral. Available at/:`https://huggingface.co/botbot-ai/CabraMistral-v3-7b-32k`. Accessed: 5 March 2025.

BotBot-AI (2024c). CabraQwen. Available at: `https://huggingface.co/botbot-ai/CabraQwen7b`. Accessed: 5 March 2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc. Available at/:`https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Brum, H. and Volpe Nunes, M. d. G. (2018). Building a sentiment corpus of tweets in Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Available at/:`https://aclanthology.org/L18-1658/`.

Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V., Akhripkin, V. S., Pavlov, I. V., Oseledets, I. V., Barsola, I. S., Egorov, I. V., Kosterina, A. A., and Zhukov, L. E. (2023). eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*. DOI: 10.1134/S1064562422060230.

Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Xu, C., Xu, R., Yan, H., Yan, Y., Yang, X., Ye, H., Ying, H., Yu, J., Yu, J., Zang, Y., Zhang, C., Zhang, L., Zhang, P., Zhang, P., Zhang, R., Zhang, S., Zhang, S., Zhang, W., Zhang, W., Zhang, X., Zhang, X., Zhao, H., Zhao, Q., Zhao, X., Zhou, F., Zhou, Z., Zhuo, J., Zou, Y., Qiu, X., Qiao, Y., and Lin, D. (2024). Internlm2 technical report. DOI: 10.48550/arxiv.2403.17297.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. Available at: `https://arxiv.org/abs/2008.09144`.

Cataneo Silveira, I. and Deratani Mauá, D. (2018). Advances in automatically solving the ENEM. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48. DOI: 10.1109/BRACIS.2018.00016.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? a study on judgement bias. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.474.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. (2021). Evaluating large language models trained on code. DOI: 10.48550/arxiv.2107.03374.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Available at: `https://lmsys.org/blog/2023-03-30-vicuna/`.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org. DOI: 10.48550/arxiv.2403.04132.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53. Available at: `http://jmlr.org/papers/v25/23-0870.html`.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1300.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. DOI: 10.48550/arxiv.2110.14168.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. (2023). Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. Available at: `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`.

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and De Oliveira, N. (2024a). TeenyTinyLlama: Open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications*, 16:100558. DOI: 10.1016/j.mlwa.2024.100558.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024b). Tucano: Advancing neural text generation for portuguese. DOI: 10.1016/j.patter.2025.101325.

Cortes, E. G., Vieira, R., and Barone, D. A. C. (2024). Perguntas e respostas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 16. BPLN, 2 edition. Available at: `https://brasileiraspln.com/livro-pln/2a-edicao/parte-interacao/cap-qa/cap-qa.html`.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., Locatelli, A., Lin, H., Talupuru, D., Venkitesh, B., Cairuz, D., Yang, B., Chung, T., Ko, W.-Y., Shi, S. S., Shukayev, A., Bae, S., Piktus, A., Castagné, R., Cruz-Salinas, F., Kim, E., Crawhall-Stein, L., Morisot, A., Roy, S., Blunsom, P., Zhang, I., Gomez, A., Frosst, N., Fadaee, M., Ermis, B., Üstün, A., and Hooker, S. (2024). Aya expanse: Combining research breakthroughs for a new multilingual frontier. DOI: 10.48550/arxiv.2412.04261.

Domingues, M. (2023). Canarim-7B-Instruct. Available at: `https://huggingface.co/dominguesm/Canarim-7B-Instruct`. Accessed: 5 March 2025.

Dong, Y., Li, Z., Rezagholizadeh, M., and Cheung, J. C. K. (2019). EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1331.

Feng, Y., Qiang, J., Li, Y., Yuan, Y., and Zhu, Y. (2023). Sentence simplification via large language models. DOI: 10.48550/arxiv.2302.11957.

FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., and Natarajan, P. (2023). MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-long.235.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/W19-3510.

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. (2024). Open LLM Leaderboard v2. Available at: `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Freitag, M., Mathur, N., Deutsch, D., Lo, C.-K., Avramidis, E., Rei, R., Thompson, B., Blain, F., Kocmi, T., Wang, J., Adelani, D. I., Buchicchio, M., Zerva, C., and Lavie, A. (2024). Are LLMs Breaking MT Metrics? results of the WMT24 metrics shared task. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.wmt-1.2.

Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., and Foster, G. (2023). Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.wmt-1.51.

Garcia, E. A. S. (2024). Open portuguese llm leaderboard. Available at: `https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard`.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L. C. S., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based task. DOI: 10.48550/arxiv.2401.02909.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170. DOI: 10.1613/jair.5477.

Gemini-Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., and et al. (2023). Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805. DOI: 10.48550/ARXIV.2312.11805.

Gemma-Team (2024). Gemma 2: Improving Open Language Models at a Practical Size. Available at: `https://arxiv.org/abs/2408.00118`.

Gemma-Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. DOI: 10.48550/arxiv.2403.08295.

Geng, X. and Liu, H. (2023). OpenLLaMA: An Open Reproduction of LLaMA. Available at: `https://github.com/openlm-research/open_llama`.

Gibaut, W. (2023). Periquito-3B. Available at `https://huggingface.co/wandgibaut/periquito-3B`.

Government of Brazil (2025). Brazilian Artificial Intelligence Plan (PBIA) 2024–2028: AI for the Good of All. Available at: `https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2025/06/publicada-versao-final-do-plano-brasileiro-de-inteligencia-artificial-sob-coordenacao-do-mcti`. Accessed on June 29, 2025.

Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J. J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. DOI: 10.2139/ssrn.4583531.

Guillou, P. (2020). GPorTuguese-2: a Language Model for Portuguese Text Generation. Available at: `https://huggingface.co/pierreguillou/gpt2-small-portuguese`.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021a). Measuring massive multitask language understanding. In *International Conference on Learning Representations*. Available at: `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021b). Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. Available at: `https://openreview.net/forum?id=7Bywt2mQsCe`.

Henrique, B. (2023a). Caramelo-7B. Available at: `https://huggingface.co/Bruno/Caramelo_7B`. Accessed: 5 March 2025.

Henrique, B. (2023b). Harpia-7B. Available at: `https://huggingface.co/Bruno/Harpia-7b-guanacoLora`. Accessed: 5 March 2025.

Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics. DOI: 10.18653/v1/P16-2096.

Hu, Y., Gan, L., Xiao, W., Kuang, K., and Wu, F. (2025). Fine-tuning large language models for improving factuality in legal question answering. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4410–4427, Abu Dhabi, UAE. Association for Computational Linguistics. Available at: `https://aclanthology.org/2025.coling-main.298/`.

Ip, J. and Vongthongsri, K. (2025). DeepEval. Available at: `https://github.com/confident-ai/deepeval`.

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., *et al.* (2023). Phi-2: The surprising power of small language models. *Microsoft Research Blog*. Available at: `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-`

of-small-language-models/.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. *CoRR*, abs/2310.06825. DOI: 10.48550/ARXIV.2310.06825.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. (2024). SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=VTF8yNQM66.

Jorge, G. A. Z., Bezerra, D. A., Xavier, C. C., and Pardo, T. A. S. (2025). Multilingual extractive summarization: Investigating state-of-the-art methods for english and brazilian portuguese. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 212–223, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-79032-4$_1$5.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.560.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. (2024). GPT-4 passes the bar exam. *Philos Trans A Math Phys Eng Sci*, 382(2270):20230254. DOI: 10.1098/rsta.2023.0254.

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., and Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.821.

Kumar, D., Mou, L., Golab, L., and Vechtomova, O. (2020). Iterative edit-based unsupervised sentence simplification. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.707.

Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., and Caridá, V. (2023). Cabrita: closing the gap for foreign languages. DOI: 10.48550/arxiv.2308.11878.

Lavie, A. and Agarwal, A. (2007). Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics. Available at: https://aclanthology.org/W07-0734/.

Leal, S. E. and Aluísio, S. M. (2024). Complexidade textual e suas tarefas relacionadas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 25. BPLN, 3 edition. Available at: https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-complexidade-textual/cap-complexidade-textual.html.

Leal, S. E., Duran, M. S., and Aluíso, S. M. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 401–413, Santa Fe, New Mexico, USA. Available at: https://aclanthology.org/C18-1034/.

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. DOI: 10.1007/s10579-023-09693-w.

Lee, S., Yu, S., Park, J., Yi, J., and Yoon, S. (2024). Interactive text-to-image retrieval with large language models: A plug-and-play approach. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–809, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.46.

Lehdonvirta, V., Wú, B., and Hawkins, Z. (2025). *Compute North vs. Compute South: The Uneven Possibilities of Compute-Based AI Governance Around the Globe*, page 828–838. AAAI Press. DOI: 10.1609/aies.v7i1.31683.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.703.

Li, P., Yang, J., Islam, M. A., and Ren, S. (2023). Making AI less "Thirsty': Uncovering and Addressing the Secret Water Footprint of AI models. DOI: 10.48550/arXiv.2304.03271.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Available at: https://aclanthology.org/W04-1013.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.229.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–

2522, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.153.

Longpre, S., Lu, Y., and Daiber, J. (2021). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406. DOI: 10.1162/tacl_a_00433.

Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., Mahari, R., Obeng-Marnu, N., Dey, M., Hamdy, M., Saxena, N., Anis, A. M., Alghamdi, E. A., Chien, V. M., Yin, D., Qian, K., LI, Y., Liang, M., Dinh, A., Mohanty, S., Mataciunas, D., South, T., Zhang, J., Lee, A. N., Lund, C. S., Klamm, C., Sileo, D., Misra, D., Shippole, E., Klyman, K., Miranda, L. J. V., Muennighoff, N., Ye, S., Kim, S., Gupta, V., Sharma, V., Zhou, X., Xiong, C., Villa, L., Biderman, S., Pentland, A., Hooker, S., and Kabbara, J. (2025). Bridging the Data Provenance Gap Across Text, Speech, and Video. In *The Thirteenth International Conference on Learning Representations*. Available at: `https://openreview.net/forum?id=G5DziesYxL`.

Luccioni, S., Gamazaychikov, B., Strubell, E., Hooker, S., Jernite, Y., Wu, C.-J., and Mitchell, M. (2025). Ai energy score. Available at: `https://huggingface.github.io/AIEnergyScore/`.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics. Available at: `https://aclanthology.org/P11-1015/`.

Malaquias-Junior, R., Pires, R., Romero, R., and Nogueira, R. (2024). Juru: Legal Brazilian Large Language Model from Reputable Sources. Available at: `https://arxiv.org/abs/2403.18140`.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2022). MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association. Available at: `https://aclanthology.org/2022.lrec-1.176/`.

Martínez, E. (2024). Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*. DOI: 10.1007/s10506-024-09396-9.

Masterman, T., Besen, S., Sawtell, M., and Chao, A. (2024). The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. Available at: `https://arxiv.org/abs/2404.11584`.

Melo, G., Imaizumi, V., and Cozman, F. (2019). Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/eniac.2019.9334.

Meta-AI (2024). The Llama 3 Herd of Models. Available at: `https://arxiv.org/abs/2407.21783`.

Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. (2024). GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*. Available at: `https://openreview.net/forum?id=fibxvahvs3`.

Mitchell, E., Rafailov, R., Sharma, A., Finn, C., and Manning, C. (2023). An emulator for fine-tuning large language models using small language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. DOI: 10.48550/arxiv.2310.12962.

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. (2023). Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. DOI: 10.48550/arxiv.2303.17003.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., and et al. (2024a). Openai o1 system card. DOI: 10.48550/arxiv.2412.16720.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., and et al. (2024b). Gpt-4o system card. DOI: 10.48550/arxiv.2410.21276.

Overwijk, A., Xiong, C., and Callan, J. (2022). ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3477495.3536321.

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024a). RecognaSumm: A novel Brazilian summarization dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceed-*

ings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 575–579, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. Available at: https://aclanthology.org/2024.propor-1.63.

Paiola, P. H., Garcia, G. L., Manesco, J. R. R., Roder, M., Rodrigues, D., and Papa, J. P. (2024b). Adapting LLMs for the Medical Domain in Portuguese: A Study on Fine-Tuning and Model Evaluation. Available at: https://arxiv.org/abs/2410.00163.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Paschoal, A. F. A., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2021). Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, page 4544–4553. ACM. DOI: 10.1145/3459637.3482012.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive APIs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Available at: https://openreview.net/forum?id=tBRNC6YemY.

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., Haznitrama, F. G., Song, I., Oh, A., and Augenstein, I. (2025). Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, pages 1–96. DOI: 10.1162/COLI.a.14.

Pereira, F. V., Frazão, A., and Moreira, V. P. (2025). Automatic text simplification for the legal domain in brazilian portuguese. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 31–45, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-79038-6_3.

Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 324–338, Porto Alegre, RS, Brasil. SBC. Available at: https://sol.sbc.org.br/index.php/bracis/article/view/33603.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45392-2-15.

Pirozelli, P., José, M. M., Silveira, I., Nakasato, F., Peres, S. M., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2024). Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63. DOI: 10.1162/dint_a_00245.

Qwen Team (2024). Introducing Qwen1.5. Available at: https://qwenlm.github.io/blog/qwen1.5/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *Open AI*. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pd

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1). DOI: 10.48550/arxiv.1910.10683.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). The assin 2 shared task: A quick overview. In Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., and Gonçalves, T., editors, *Computational Processing of the Portuguese Language*, pages 406–412, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-41505-$1_3$9.

Recogna-NLP (2024a). Bode-3.1-8B-Instruct. Available at: https://huggingface.co/recogna-nlp/Bode-3.1-8B-Instruct-full. Accessed: 5 March 2025.

Recogna-NLP (2024b). Doutor-bode. Available at: https://huggingface.co/recogna-nlp/doutor-bode-7b-360k. Accessed: 5 March 2025.

Recogna-NLP (2024c). GemBode-7B. Available at: https://huggingface.co/recogna-nlp/gembode-7b. Accessed: 5 March 2025.

Recogna-NLP (2024d). InternLm-ChatBode. Available at: https://huggingface.co/recogna-nlp/internlm-chatbode-20b. Accessed: 5 March 2025.

Recogna-NLP (2024e). Mistral-Bode. Available at: https://huggingface.co/recogna-nlp/mistral-bode. Accessed: 5 March 2025.

Recogna-NLP (2024f). Phi-Bode. Available at: https://huggingface.co/recogna-nlp/phi-bode-2-ultraalpaca. Accessed: 5 March 2025.

Recogna-NLP (2024g). QwenBode. Available at: https://huggingface.co/recogna-nlp/qwenbode_1_8b_chat_ultraalpaca. Accessed: 5 March 2025.

Rehman, T., Das, S., Sanyal, D. K., and Chattopadhyay, S. (2022). An Analysis of Abstractive Text Summarization Using Pre-trained Models. In Mandal, L., Tavares, J. M. R. S., and Balas, V. E., editors, *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing*, pages 253–264, Singapore. Springer Nature Singapore. DOI: 10.48550/arXiv.2303.12796.

Reidsma, D. and op den Akker, R. (2008). Exploiting 'Subjective' Annotations. In Artstein, R., Boleda, G., Keller, F., and Schulte im Walde, S., editors, *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee. Avaialble at: https://aclanthology.org/W08-1203/.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*. Avaialble at: https://openreview.net/forum?id=Ti67584b98.

Rino, L. H. M. and Pardo, T. A. S. (2003). A sumarização automática de textos: Principais características e metodologias. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA)*, pages 203–245, Campinas-SP. Avaialble at: `https://www.di.ubi.pt/~jpaulo/competence/papers/(2003)RinoPardo.pdf`.

Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2022). A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.*, 55(2). DOI: 10.1145/3485766.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2021). WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106. DOI: 10.1145/3474381.

Santa Brígida, L. (2024a). Boto-7B. Available at:`https://huggingface.co/lucianosb/boto-7B`.

Santa Brígida, L. (2024b). Boto-9B. Available at:`https://huggingface.co/lucianosb/boto-9B`.

Santos, R., Silva, J. R., Gomes, L., Rodrigues, J., and Branco, A. (2024). Advancing Generative AI for Portuguese with Open Decoder gervásio PT*. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 16–26, Torino, Italia. ELRA and ICCL. Available at:`https://aclanthology.org/2024.sigul-1.3/`.

Sarmento, M. and de Oliveira, H. (2024). Sumarização automática de artigos de notícias em português: Da extração à abstração com abordagens clássicas e modelos de neurais. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 139–148, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/stil.2024.245395.

Sayama, H. F., Araujo, A. V., and Fernandes, E. R. (2019). Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. DOI: 10.1109/BRACIS.2019.00084.

Scalercio, A., de Souza, E. A., Finatto, M. J. B., and Paes, A. (2025). Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria. Association for Computational Linguistics. To appear. DOI: 10.18653/v1/2025.acl-long.1193.

Scalercio, A., Finatto, M., and Paes, A. (2024). Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15076–15091, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.895.

Schneider, E. T. R., de Souza, J. V. A., Gumiel, Y. B., Moro, C., and Paraiso, E. C. (2021). A GPT-2 Language Model for Biomedical Texts in Portuguese. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 474–479. DOI: 10.1109/CBMS52027.2021.00056.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.704.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ko, W.-Y., Ruder, S., Smith, M., Bosselut, A., Oh, A., Martins, A. F. T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. (2025). Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. DOI: 10.48550/arxiv.2412.03304.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. DOI: 10.1038/s41586-023-06291-2.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics. DOI: 10.18653/v1/d13-1170.

Souza, E., Silva, P., Gomes, D., Batista, V., Batista, E., and Pacheco, M. (2024a). TableRAG: A Novel Approach for Augmenting LLMs with Information from Retrieved Tables. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 182–191, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/stil.2024.245371.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer. DOI: 10.1007/978-3-030-61377-8_28.

Souza, J. W. d. C., Cardoso, P. C. F., and Paixão, C. A. (2024b). Sumarização automática. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 24. BPLN, 3 edition. Available at:`https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-as/cap-as.html`.

Sprague, Z. R., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. (2024). MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*. Available at:`https://openreview.net/forum?id=jenyYQzue1`.

Srivastava, A. and Memon, A. (2024). Toward Robust Evaluation: A Comprehensive Taxonomy of Datasets and Metrics for Open Domain Question Answering in the Era of Large Language Models. *IEEE Access*, 12:117483–117503. DOI: 10.1109/ACCESS.2024.3446854.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., and et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification. DOI: 10.48550/arxiv.2206.04615.

Sulem, E., Abend, O., and Rappoport, A. (2018). BLEU is not suitable for the evaluation of text simplification. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1081.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-following LLaMA model. Available at:`https://github.com/tatsu-lab/stanford_alpaca`.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971. DOI: 10.48550/ARXIV.2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288. DOI: 10.48550/ARXIV.2307.09288.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association. Available at:`https://aclanthology.org/2022.lrec-1.777/`.

Vasilyev, O., Dharnidharka, V., and Bohannon, J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. In Eger, S., Gao, Y., Peyrard, M., Zhao, W., and Hovy, E., editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.eval4nlp-1.2.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. DOI: 10.48550/arxiv.1706.03762.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Available at:`https://aclanthology.org/L18-1686/`.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA. DOI: 10.48550/arXiv.1905.00537.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/W18-5446.

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. Available at:`https://github.com/kingoflolz/mesh-transformer-jax`.

Wang, H., Li, J., Wu, H., Hovy, E., and Sun, Y. (2023). Pre-Trained Language Models and Their Applications. *Engineering*, 25:51–65. DOI: 10.1016/j.eng.2022.04.024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Lan-*

guage Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-demos.6.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., XIONG, G., Deng, Z., Jiang, Y., Yao, Z., Li, H., Yu, Y., Hu, G., Jiajia, H., Liu, X.-Y., Lopez-Lira, A., Wang, B., Lai, Y., Wang, H., Peng, M., Ananiadou, S., and Huang, J. (2024). Finben: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Available at:`https://openreview.net/forum?id=loDHZstVP6`.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415. Available at:`https://www.aclweb.org/anthology/Q16-1029`.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.naacl-main.41.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. (2024). Qwen2 technical report. DOI: 10.48550/arxiv.2407.10671.

Yuan, R., Lin, H., Wang, Y., Tian, Z., Wu, S., Shen, T., Zhang, G., Wu, Y., Liu, C., Zhou, Z., Xue, L., Ma, Z., Liu, Q., Zheng, T., Li, Y., Ma, Y., Liang, Y., Chi, X., Liu, R., Wang, Z., Lin, C., Liu, Q., Jiang, T., Huang, W., Chen, W., Fu, J., Benetos, E., Xia, G., Dannenberg, R., Xue, W., Kang, S., and Guo, Y. (2024). ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6252–6271, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.373.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1472.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org. DOI: 10.48550/arxiv.1912.08777.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. Available at:`https://openreview.net/forum?id=SkeHuCVFDr`.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press. DOI: 10.48550/arXiv.1509.01626.

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., and Huang, M. (2024). SafetyBench: Evaluating the safety of large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.830.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Available at:`https://openreview.net/forum?id=uccHPGDlao`.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.2306.05685.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. (2024). AGIEval: A human-centric benchmark for evaluating foundation models. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-naacl.149.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. (2023). Instruction-following evaluation for large language models. DOI: 10.48550/arxiv.2311.07911.

A PT-BR Summarization Prompt

Sumarize de forma breve e direta *{text_type}* a seguir.

{text_type}: *{original_text}*
Sumário:

B PT-BR Simplification Prompt

Substitua a frase complexa por uma frase simples. Mantenha o mesmo significado, mas torne-a mais simples.
Frase complexa: *{text}*
Frase Simples:

C PT-BR QA Prompt

Responda à seguinte pergunta com base em seu conhecimento geral sobre *{subject}*.
Seja objetivo
Pergunta: *{question}*
Resposta:

D Metrics Implementation Details

To ensure reproducibility, we describe the computation of each evaluation metric used in this work. The Hugging Face evaluate[24] library (version 0.4.1) was used to compute BLEU, ROUGE, BLANC (`phucdev/blanc_score`), METEOR, and BERTScore, with only the language parameter adjusted to Portuguese when applicable. For SARI, we employed the public implementation available in the EASSE [Alva-Manchego *et al.*, 2019] GitHub repository[25]. G-Eval was applied using the implementation provided by DeepEval[26]. The NILC indicators were computed based on their official repository[27].

For the QA F1 score, following Srivastava and Memon [2024], we compute it as the harmonic mean of precision and recall, defined in this context as:

$$\text{Precision} = \frac{\text{Number of Correctly Predicted Words}}{\text{Number of Words in the Prediction}}$$

$$\text{Recall} = \frac{\text{Number of Correctly Predicted Words}}{\text{Number of Words in the Reference}}$$

Finally, the metrics related to carbon emissions, runtime, and energy consumption were calculated using the official implementation of eco2AI[28].

E PT-BR Summarization Evaluation Criteria

A. Gramaticalidade
Gramaticalidade (1-5) - O sumário não deve conter erros de capitalização ou frases claramente agramaticais (como fragmentos ou componentes ausentes) que dificultem a leitura.

B. Não-redundância
Não-redundância (1-5) - O sumário não deve conter repetições desnecessárias. A repetição pode ocorrer na forma de frases inteiras repetidas, fatos repetidos ou o uso excessivo de um substantivo quando um pronome poderia ser utilizado.

C. Clareza Referencial
Clareza Referencial (1-5) - Deve ser fácil identificar a quem ou a quê os pronomes e expressões nominais do sumário se referem. Se uma pessoa ou entidade for mencionada, seu papel na história deve estar claro. Uma referência será considerada pouco clara se uma entidade for citada sem que sua identidade ou relação com a história fique evidente.

D. Foco
Foco (1-5) - O sumário deve ter um foco claro; as frases devem conter apenas informações relevantes para o restante do sumário.

E. Estrutura e Coerência
Estrutura e Coerência (1-5) - O sumário deve ser bem estruturado e organizado. Não deve ser apenas um amontoado de informações relacionadas, mas sim construir um corpo coerente de informações sobre um tópico.

F PT-BR Simplification Evaluation Prompt

A. Adequação
Adequação (1-5) - Quantos significados da frase original são preservados na versão simplificada? A frase simplificada deve manter o conteúdo essencial da original sem perda significativa de informações.

B. Simplicidade
Simplicidade (1-5) - A frase simplificada é de fato mais simples do que a original? A simplificação deve reduzir a complexidade da estrutura da frase, tornando-a mais fácil de entender sem introduzir informações erradas ou redundantes.

C. Fluência
Fluência (1-5) - A frase simplificada é gramaticalmente correta e bem formada? A estrutura sintática deve ser natural e legível, sem erros que dificultem a compreensão.

[24] https://huggingface.co/docs/evaluate/
[25] https://github.com/feralvam/easse
[26] https://github.com/confident-ai/deepeval
[27] https://github.com/nilc-nlp/nilcmetrix
[28] https://github.com/sb-ai-lab/Eco2AI

G PT-BR QA Evaluation Prompt

A. Qualidade da Resposta
Qualidade da Resposta (1-5) - Diretrizes de pontuação:
- 1: A resposta está completamente incorreta. É totalmente diferente da referência ou a contradiz.
- 2: A resposta apresenta algum grau de similaridade semântica e inclui informações parcialmente corretas. No entanto, ainda contém discrepâncias significativas em relação à referência ou imprecisões.
- 3: A resposta aborda corretamente alguns aspectos, alinhando-se à referência. No entanto, ainda há omissões ou pequenas imprecisões.
- 4: A resposta está majoritariamente correta. Fornece informações precisas, mas pode conter uma ou mais omissões ou imprecisões menores.
- 5: A resposta está correta. Demonstra um alto grau de precisão e similaridade semântica com a referência.

H Example of a Complementary Simplification in Portuguese

Original Sentence
O zoneamento representa o primeiro passo na formatação de uma política florestal no Rio Grande do Sul.

Original Simplification Reference
A divisão do território em zonas representa o primeiro passo na preparação de uma política florestal no Rio Grande do Sul.

Simplification Generated by GPT-4o
O zoneamento é o começo de uma política florestal no Rio Grande do Sul.