



# BRoverbs - Measuring how much LLMs understand Portuguese proverbs

Thales Sales Almeida   [ Institute of Computing, University of Campinas. Maritaca AI | [t224732@dac.unicamp.br](mailto:t224732@dac.unicamp.br) ]

Giovana Kerche Bonás  [ Institute of Computing, University of Campinas. Maritaca AI | [g216832@dac.unicamp.br](mailto:g216832@dac.unicamp.br) ]

João Guilherme Alves Santos  [ Institute of Computing, University of Campinas | [j199624@dac.unicamp.br](mailto:j199624@dac.unicamp.br) ]

 Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, Campinas, SP, 13083-852, Brazil  
Maritaca AI, Campinas, SP, Brazil

Received: 31 March 2025 • Accepted: 17 July 2025 • Published: October 2025

**Abstract** Large Language Models (LLMs) exhibit significant performance variations depending on the linguistic and cultural context in which they are applied. This disparity signals the necessity of mature evaluation frameworks that can assess their capabilities in specific regional settings. In the case of Portuguese, existing evaluations remain limited, often relying on translated datasets that may not fully capture linguistic nuances or cultural references. Meanwhile, native Portuguese-language datasets predominantly focus on structured national exams or sentiment analysis of social media interactions, leaving gaps in evaluating broader linguistic understanding. To address this limitation, we introduce BRoverbs, a dataset specifically designed to assess LLM performance through Brazilian proverbs. Proverbs serve as a rich linguistic resource, encapsulating cultural wisdom, figurative expressions, and complex syntactic structures that challenge the model comprehension of regional expressions. BRoverbs aims to provide a new evaluation tool for Portuguese-language LLMs, contributing to advancing regionally informed benchmarking. The benchmark is available at <https://huggingface.co/datasets/Tropic-AI/BRoverbs>.

**Keywords:** Large Language Models, LLM Benchmark, Portuguese LLM Evaluation

## 1 Introduction

In recent years, significant investments have been made in the development and expansion of large language models (LLMs). As these models become more integrated into real-world applications [Li *et al.*, 2023; Liang *et al.*, 2024; Rane *et al.*, 2023], evaluating their abilities across various tasks has become increasingly important. Reliable and varied evaluation methods help measure how well LLMs understand language, handle different tasks, and perform across languages and cultures.

Although LLMs have gained increasing global attention, research has predominantly focused on English, leaving evaluation datasets in other languages, such as Portuguese, relatively scarce [Giagkou *et al.*, 2023]. As a result, many studies on Portuguese LLMs rely, in part, on translating existing tasks to measure performance [Almeida *et al.*, 2024; Corrêa *et al.*, 2024; Larcher *et al.*, 2023]. These translated tasks often fail to capture regional and cultural nuances. This issue was highlighted by Sabiá [Almeida *et al.*, 2024], which observed that after pretraining on Portuguese data, the model showed greater performance gains on tasks originally designed in Portuguese than on translated ones.

Given the increasing global expansion of LLM applications [Naveed *et al.*, 2023; Zhao *et al.*, 2023], it is crucial to assess whether they effectively fulfill regional capabilities such as linguistic comprehension, cultural adaptability, and performance in native language tasks.

Recent studies have shown that LLMs perform unevenly

across regions, particularly when dealing with underrepresented countries and contexts [Myung *et al.*, 2024; Moayeri *et al.*, 2024; Almeida *et al.*, 2025a]. These disparities underscore the importance of native benchmarks that reflect local knowledge and cultural specificity.

The performance gap observed in benchmarks likely reflects the low representation of certain regions in training data. Studies on data governance have revealed a significant imbalance in data sources, with Latin American representation remaining exceedingly low and showing no signs of increasing. [Longpre *et al.*, 2024] audited training datasets over recent years, showing that data representation is overwhelmingly concentrated in North America, Western Europe, and China, while Latin America contributes less than 0.5% of the total. Such audits motivate targeted evaluation—not because a single benchmark can resolve systemic bias, but because it can reveal where that bias manifests in model behavior.

This imbalance not only impacts model performance but also reinforces systemic biases, further marginalizing non-dominant languages and cultural expressions in AI-driven technologies. To better understand how such imbalances manifest, we propose BRoverbs, a dataset designed to evaluate the comprehension of Brazilian proverbs by LLMs. Proverbs are deeply rooted in cultural and linguistic contexts, making them a suitable signal of a model’s ability to grasp regional semantics and idiomatic expressions. Our benchmark represents a step forward in evaluating Brazilian regional knowledge.

Additionally, the creation of culturally grounded datasets

like BRoverbs serves a social purpose. By concentrating on commonly shared folk wisdom encoded in Brazilian proverbs, we underscore the need for inclusive large language models. Such models should be capable of understanding not only the direct meaning of words but also the deeply woven cultural, historical, and linguistic nuances that shape diverse communities. This approach encourages equitable access to advanced language technologies for Brazilian Portuguese speakers.

Our contributions are as follows.

- We introduce BRoverbs, a new benchmark designed to evaluate LLMs understanding of Brazilian proverbs, improving Brazil’s scenario of culturally aware LLM evaluation.
- We evaluate a diverse set of LLMs, including commercial and open-source models of varying sizes, highlighting disparities in performance and the impact of model scale and pretraining data composition.
- We discuss the implications of cultural underrepresentation in LLM training data and emphasize the importance of developing native benchmarks to ensure fair and accurate assessments of linguistic understanding in underrepresented languages.

## 2 Related work

The evaluation of large language models (LLMs) has been an active area of research, with numerous benchmarks designed to assess different aspects of model performance. These benchmarks focus on a wide variety of tasks, such as reading comprehension [Dua *et al.*, 2019; Lai *et al.*, 2017; Rajpurkar *et al.*, 2016; Kwiatkowski *et al.*, 2019; Yu *et al.*, 2020], sentiment analysis and subjectivity [Potts *et al.*, 2020; Alfina *et al.*, 2017], reasoning and factual knowledge retrieval [Yu *et al.*, 2020; Thorne *et al.*, 2018; Jiang *et al.*, 2020; Ho *et al.*, 2020], generation and creativity [Srivastava *et al.*, 2022; Chen *et al.*, 2021; Hasan *et al.*, 2021], and others [Rudinger *et al.*, 2018; Parrish *et al.*, 2021; Wang *et al.*, 2019]. While these benchmarks provide a broad foundation for evaluating LLMs, they predominantly focus on English, leaving gaps in assessing model performance in other languages. In Portuguese, evaluation resources remain scarce, often requiring the translation of existing English-language tasks, which can introduce inconsistencies and fail to capture the linguistic and cultural nuances necessary for robust assessments.

### 2.1 Evaluating LLMs in Portuguese and Iberian Languages

: Trends and Gaps

Recent efforts to evaluate LLMs in Portuguese have followed two main trends. The first trend focuses on structured exams used for human evaluation, such as benchmarks based on the Brazilian national university entrance exam (ENEM) [Silveira and Mauá, 2017] or other universities entrance exams [Almeida *et al.*, 2023] and based on the Order of Attorneys of Brazil national bar exam [Delfino *et al.*, 2017]. While the second trend, centers around social media classifications, such as sentiment analysis in tweets proposed by

TweetSentBR [Brum and Nunes, 2017], social media hate detection through datasets like HateBR [Vargas *et al.*, 2021] and PT Hate Speech [Fortuna *et al.*, 2019], and other studies targeting the analysis of informal, user-generated content, such as fake news detection or similar tasks.

While both approaches offer valuable insights into model performance, they fail to fully capture key aspects of factual recall, cultural adaptation, and regional linguistic variation. Notable exceptions include Faquad [Sayama *et al.*, 2019], an small extractive QA dataset, and TiEBe [Almeida *et al.*, 2025a], which focuses on open-domain QA about Brazilian events. However, these efforts remain limited in scope, highlighting the need for benchmarks that evaluate LLMs within richer cultural and linguistic contexts. A recent contribution in this direction is IberoBench [Baucells *et al.*, 2025] which targets the evaluation gap for Iberian languages—including Basque, Catalan, Galician, European Spanish, and European Portuguese. The benchmark comprises 62 tasks grouped into 179 subtasks across ten skill categories (e.g., commonsense reasoning, NLI, QA, summarization, translation) and assesses 33 LLMs in zero-shot and few-shot settings, revealing that performance in these languages still trails state-of-the-art results in English.

### 2.2 Regional Disparities in LLM Evaluation

A recent wave of research has introduced benchmarks designed to address biases and regional disparities in LLM factual recall. Notable among them are WorldBench [Moayeri *et al.*, 2024], TiEBe [Almeida *et al.*, 2025a] and BLEnD [Myung *et al.*, 2024]. These benchmarks reveal systemic biases in LLMs and highlight areas where model performance remains uneven.

WorldBench [Moayeri *et al.*, 2024] is designed to assess geographic disparities in factual recall using data per country from the World Bank, an international financial institution that provides funding, advice and research to help countries reduce poverty and promote sustainable development. The benchmark has revealed significant biases in LLM performance based on region and income level, showing that error rates are notably higher for lower-income countries, particularly in regions such as Sub-Saharan Africa. These disparities emphasize the need for more representative and globally inclusive evaluation methodologies.

Similarly, TiEBe [Almeida *et al.*, 2025a] expands LLM evaluation by addressing real-world knowledge and historical events with a dataset of over 11,000 question-answer pairs. On Wikipedia, there are annual pages that mention significant events both globally and in various countries. These pages list notable events, each accompanied by a set of external citation links, often pointing to journalistic sources that provide further context. TiEBe utilizes this information, spanning a 10-year interval, to generate question-answer pairs about the events. The purpose is to assess whether a language model can recall information from the original news sources, even though the model will not have direct access to these documents during testing. This approach ensures that the questions reflect major events and figures in a consistent and structured manner. It also reveals that LLMs often fail to represent events consistently across different regions, including

Portuguese-speaking countries such as Brazil and Portugal.

BLEnD [Myung *et al.*, 2024] takes a different approach by evaluating LLMs’ understanding of daily knowledge across multiple cultural and linguistic contexts. It includes 52.6k manually crafted question-answer pairs covering 16 countries and 13 languages, revealing that LLMs perform better in cultures and languages with stronger digital representation. While models generate more accurate responses in native languages for mid-to-high-resource languages, they often provide more reliable answers in English when dealing with low-resource languages. This reflects the ongoing challenge of training LLMs to effectively represent diverse linguistic and cultural knowledge.

Collectively, WorldBench, TiEBE, and BLEnD underscore the pressing need for culturally aware evaluation benchmarks that extend beyond mainstream datasets. These efforts expose significant gaps in LLMs’ regional knowledge and cultural comprehension, highlighting the broader challenge of ensuring AI systems equitably serve diverse global communities. At the same time, they demonstrate the potential for further research in this area, paving the way for new methodologies that assess how well models handle complex cultural expressions.

## 2.3 LLM Performance in Different Languages

While previous studies focused on regional disparities in factual recall, a growing body of work has turned its attention to a different but related challenge: the performance of LLMs across diverse languages, particularly those with limited digital representation. Linguistic diversity introduces unique obstacles, as models often struggle with languages that lack large-scale training data, affecting their ability to generate accurate and contextually appropriate responses.

Recent benchmarks have highlighted the limitations of LLMs in capturing region-specific knowledge, particularly in languages with limited digital presence. For example, Pariksha [Watts *et al.*, 2024] evaluates 30 LLM models in 10 Indic languages, conducting a large-scale comparison between human and LLM-based evaluations. The project reveals that discrepancies in evaluation arise in lower-resource languages such as Bengali and Odia, where human and machine evaluations diverge significantly, reinforcing concerns about how models interpret and assess culturally embedded concepts.

These studies reinforce the idea that cultural and linguistic biases in LLMs are a widespread issue, not limited to Portuguese. They also illustrate how regionally adapted benchmarks can help uncover deficiencies in the ability of LLMs to generalize knowledge between different cultures.

## 2.4 Figurative-Language Benchmarks

Figurative language—idioms, metaphors, similes and, above all, proverbs—conveys meaning that is not compositionally recoverable from the literal sense of its words. Because large language models (LLMs) are typically trained to predict text token-by-token, their ability to recognize and interpret such nonliteral signals is a stringent test of genuine contextual understanding. Dedicated benchmarks have emerged to probe this ability.

ProverbEval [Azime *et al.*, 2024] contributes to the evaluation of the understanding of proverbs by LLMs, particularly in the context of low-resource languages. It introduces a multilingual benchmark to evaluate LLMs’ understanding of proverbs, with a particular emphasis on low-resource and typologically diverse languages, presenting evaluations in Amharic, Tigrinya, Ge’ez, Afaan Oromo, and English.

The benchmark is structured in three different types of tasks: (1) a meaning-based multiple-choice task, in which models select the correct interpretation of a proverb from four detailed options; (2) a fill-in-the-blank task, which tests the model’s ability to recognize conventional phrasings in known proverbs; and (3) a generation task, where models must retrieve the appropriate proverb given a detailed description of its meaning or situational use. This setup allows researchers to examine not only a model’s figurative-language understanding but also its ability to transfer proverb knowledge across languages and scripts.

While this project shares similarities with BRoverbs, particularly in its focus on assessing regional linguistic comprehension through culturally embedded expressions, it adopts a different methodological approach. ProverbEval’s approach sheds light on the key challenges in evaluating how well models grasp proverb meanings across different linguistic landscapes. In contrast, BRoverbs provides a focused analysis of Brazilian Portuguese, offering a detailed examination of the regional linguistic complexities unique to this language.

Complementary to these proverb-focused datasets, [Mi *et al.*, 2024] present Rolling the DICE on Idiomacity, a controlled contrastive benchmark that pairs 402 English idioms with twin sentences—one literal, one figurative—differing only in their surrounding context. Results on 13 popular LLMs (GPT-4o, Llama-3, etc.) show that even the strongest models struggle to resolve the literal-versus-figurative ambiguity, often defaulting to the figurative sense and achieving below-perfect “strict consistency” across sentence pairs. This finding highlights that figurative interpretation remains challenging even in high-resource languages and supports the need for culturally grounded evaluations like BRoverbs.

## 3 Methodology

In this section, we describe the pipeline for creating BRoverbs, which is illustrated in Figure 1. Our goal was to build a dataset that connects popular Brazilian proverbs to short narratives and propose tasks that assess how effectively language models can correctly associate each proverb with its corresponding story and vice versa.

### 3.1 Data Collection and Story Generation

Our process began by collecting Brazilian proverbs from online sources. We conducted three web searches<sup>1</sup> for Brazilian proverbs and gathered the top five results for each query, resulting in 15 different sources. Proverbs were extracted from all sites, then grouped using lexical clustering based on fuzzy

<sup>1</sup>As buscas foram: ‘lista de provérbios brasileiros’, ‘ditados populares brasileiros’, ‘provérbios populares brasileiros’



**Figure 1.** Methodological flow of *BRoverbs*, illustrating the steps of data collection, story generation, creation of questions and answers, and model evaluation.

string matching. Duplicates were then manually removed, yielding a final set of 196 unique proverbs.

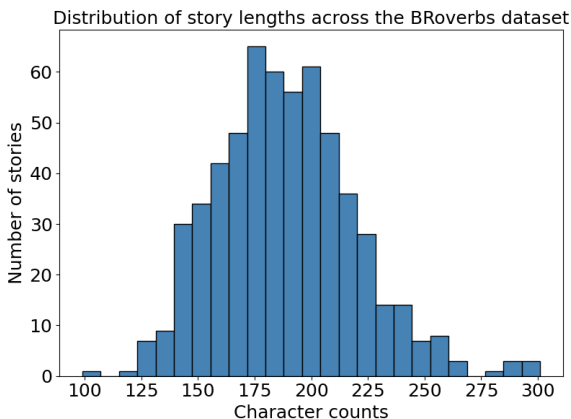
For each proverb, we used GPT-4 to generate three short narratives illustrating its meaning, without quoting or referencing the proverb directly. The prompt also asked GPT-4 to provide a brief explanation of the moral or practical meaning behind each proverb for verification purposes. The used prompt is available at appendix A.2.

Each of the three stories for each proverb was manually verified, and some adjustments were made when the meaning of the story did not correctly reflect the meaning of the expression, and in some rare cases, the stories were completely rewritten. In this process, we also dropped three proverbs completely, such as “Batatinha quando nasce, se esparrama pelo chão”, due to not reaching a consensus for their meaning. Thus, we obtained:

- 579 short stories (3 for each of the 193 proverbs left from manual verification).
- A concise explanatory summary for each proverb.

### 3.1.1 Distribution of Story Length in Characters

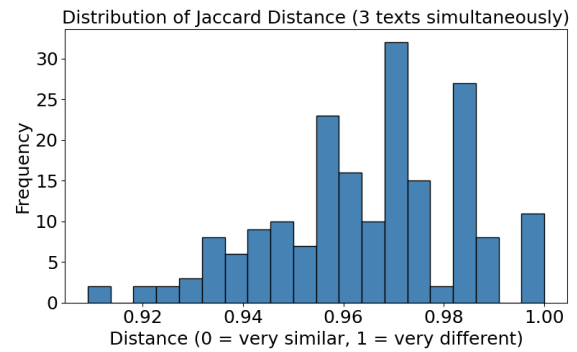
To gain deeper insights into the textual complexity of the 579 stories, we computed the length of each story by counting the total number of characters. Figure 2 shows the histogram of character counts (x-axis) against the number of stories (y-axis). We observe that most stories lie within a similar range of lengths, reflecting a relatively consistent output from the generative prompt.



**Figure 2.** Distribution of story lengths (in characters) across the *BRoverbs* dataset.

### 3.1.2 Jaccard Distance Across Triplet Stories

In addition to examining story length, we also measured how distinct the three short stories generated for each proverb were from one another, using the Jaccard distance. As shown in Figure 3, these distances are generally quite high—most values exceed 0.9. This indicates that the short stories typically share relatively few words in common, reflecting a high degree of lexical diversity. Such diversity is important to ensure each story presents a unique illustration of the same underlying proverb without simply repeating the same wording.



**Figure 3.** Distribution of Jaccard distances among the three short stories for each proverb.

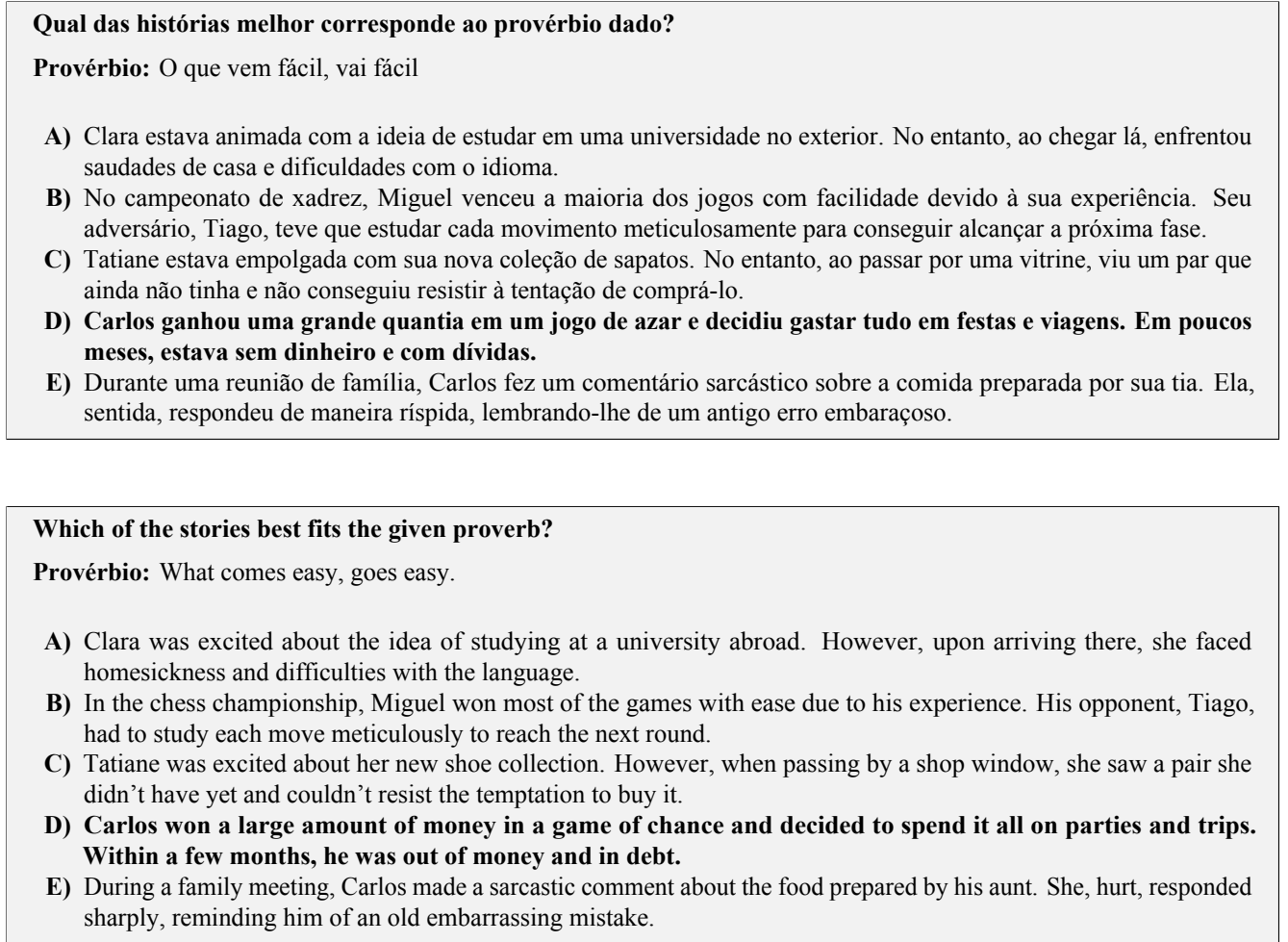
## 3.2 Generation of Question-Answer Pairs

After compiling the proverbs and their corresponding narratives, we designed two tasks to assess whether language models can accurately match proverbs to stories:

- **PtS (Proverb to Story):** Given a specific proverb, the model must identify which of the five provided short stories best represents that proverb. One of the short stories matches the proverb accurately, while the remaining four options were randomly selected from the full set of 579 stories. An example of this task can be seen in Figure 4.
- **StP (Story to Proverb):** Given a short story, the model must select which of the five presented proverbs is most closely related to that story. Here, too, one proverb was correct, and the remaining four were randomly chosen from the total pool of proverbs. An example of this task can be seen in Figure 5.

Because some proverbs have similar meanings—for example, “Quem com ferro fere, com ferro será ferido” and “Tudo que vai, volta”—randomly selecting stories and proverbs for the alternatives can result in more than one correct answer. To address this, we manually reviewed the distractor options to ensure that none of the three randomly chosen proverbs unintentionally matched the meaning of the correct answer, making corrections whenever such overlaps were found.

This process was performed in two stages. First, we reviewed each question in the dataset to ensure the alternatives were appropriate. Then, after running four commercial models (GPT-4o, GPT-4o-mini, Sabiazinho-3, and Sabiá-3), we re-examined all questions that were answered incorrectly by at least one of the models—this accounted for about 12% of



**Figure 4.** Example from the *BRoverbs* dataset: The first section shows a **PtS (Proverb to Story)** task in Brazilian Portuguese, and the second presents the **literal** English translation of the same task.

the total questions. From these questions, around 35% were modified.

### 3.3 Model Evaluation

We evaluated these tasks using a range of language models, including both open-source and commercial ones. Each model was tested on its ability to correctly align a proverb with its corresponding short story (PtS task) and vice versa (StP task). The performance outcomes served as a measure of how effectively each model could interpret and map the semantic content of the proverbs to the narrative situations generated for them. We used an RTX A6000 for all our experiments with open-source models. All evaluations were conducted in a 1-fewshot scenario. We randomly selected a different example from the dataset for each test instance to serve as the single-shot illustration, ensuring it was not the same as the input being evaluated, preventing overlap and avoiding potential data leakage.

## 4 Results

We now present the performance of various LLMs on the *BRoverbs* benchmark, analyzing how model size, training

data, and language exposure affect task success.

### 4.1 Model Results

We evaluated our dataset using a diverse selection of LLMs, spanning both open-source and commercial models accessible via API. Among the commercial models, we included OpenAI's GPT-4o and GPT-4o-mini, Anthropic's Claude 3.5 Sonnet and Haiku, as well as Maritaca AI's Sabiá-3 and Sabiazinho-3, which represent Brazilian commercial alternatives. On the open-source side, we tested the Qwen 2.5B family [Yang *et al.*, 2024], ranging from 1.5B to 14B parameters, a state-of-the-art option, along with three generations of LLaMA models [Touvron *et al.*, 2023a,b; Dubey *et al.*, 2024] at approximately 7B parameters each. We also evaluated Sabiá 7B [Pires *et al.*, 2023], a further pretrained Portuguese variant of LLaMA-1 7B. To assess smaller models, we included Tucano 1.1B and 2.4B [Corrêa *et al.*, 2024], a recent family trained from scratch in Portuguese. Additionally, we considered TinyLlama 1T and 3T, 1.1B-parameter models trained on a large corpus of English data. Finally, we tested Curió-1.1B, which builds on TinyLlama 1T with an additional 150B tokens of Portuguese pretraining.

We separate the commercial models into two categories, "Large" and "Small", based on their price range. As for the

**Qual dos provérbios melhor corresponde à história dada?**

**História:** Mariana viu um homem com roupas rasgadas sentado na praça e logo pensou que ele era desleixado. Mais tarde, descobriu que ele era um renomado artista de rua, famoso por sua generosidade.

- A) Em time que está ganhando, não se mexe
- B) Um homem prevenido vale por dois
- C) **Não julgue um livro pela capa**
- D) Todos os caminhos levam à Roma
- E) Cada um sabe onde o sapato aperta

**Which proverb best matches the given story?**

**Story:** Mariana saw a man in torn clothes sitting in the square and immediately thought he was careless. Later, she discovered that he was a renowned street artist, famous for his generosity.

- A) In a team that is winning, don't make changes.
- B) A forewarned man is worth two.
- C) **Don't judge a book by its cover.**
- D) All roads lead to Rome.
- E) Each one knows where the shoe pinches.

**Figure 5.** Example from the *BRoverbs* dataset: The first section shows a **StP (story to proverb)** task in Brazilian Portuguese, and the second presents the **literal** English translation of the same task.

open source models, we call "Medium" models above 7B parameters and "small" models with less than 7B parameters. The chosen models were selected to study how models with different sizes and training sizes perform in our tasks; the full results on both *BRoverbs* tasks, models sizes and pretrain sizes of the models are shown in Table 1, pretrain sizes in the format "X+Y" indicate a model under continued pretraining, where X is the number of tokens used in the base model, and Y the number of tokens used in the second pretrain.

First, we can see that both the 'Small' and 'Large' commercial models perform well in the task, all achieving an average between the two tasks superior to 95%. Here is worth noting that while GPT-4o was used to help in story elaboration, it does not show a significant performance gap when compared to the other commercial models of similar cost.

One notable trend across almost all models is that the StP variant of the task is generally easier than the PtS task. This can be observed in the consistently higher scores for StP, even in models that perform poorly overall. The discrepancy between the two tasks suggests that identifying a proverb given a story is a simpler task than recognizing a story that matches the meaning of a given proverb. This is likely because many proverbs are distinct and memorable phrases, making them easier for models to recognize when presented as answer choices. In contrast, the PtS task requires the model to recognize a story that matches the proverb's meaning, which demands a deeper interpretation of the story's nuances and the proverb's meaning.

Looking at the medium-level open-sourced models, we can see a big variety of performance. First, we can see the evolution of performance of the Llama Models, LLama1 and LLama2 show performances very close to the random expected score, and the performance improvement from LLama1 to

LLama2 is negligible, a possible reason for this is that, while LLama2 trains in the double of tokens, the proportion of Portuguese tokens in it's train is still very low at 0.09% of the training data [Touvron *et al.*, 2023b], the low contact with Portuguese documents may have hindered the models understanding of local expressions. LLama3-8B however, does show an expressive improvement over LLama2-7B, LLama3-8B is trained on 7 times more tokens, but the language distribution of the training data is not known, making it hard to discuss the impact of Portuguese documents in this performance improvement.

Still on the topic of the Llama models, Sabiá-7B is a model further pretrained in 10B Portuguese starting from Llama-7B; we can see that this Portuguese specialization did bring some gains in the StP task, but not so much in the PtS task. It is interesting to see that Sabiá-7B slightly outperforms LLama2-7B by training in Portuguese documents in a much smaller scale, as LLama2 trains in 2T tokens, and Sabiá-7B total train is 1T tokens, herded from using Llama 7B as base, plus 10B Portuguese tokens, this results points to the importance of regional documents in model pretraining.

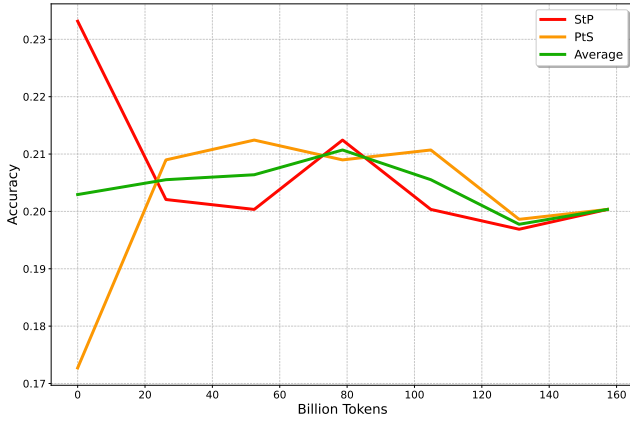
Looking at the Qwen models, they clearly outperform their counterparts with the same number of parameters and show a progressive improvement as the parameter count goes up, with Qwen 2.5 14B reaching performance close to the commercial models. Qwen models are pretrained in 18T tokens, but little is known about the training data, thus limiting our discussion here. Nevertheless, we can observe a positive correlation between model size and model performance in the *BRoverbs* tasks.

Finally, looking at the smaller models, Tucano, Tinyllama, and Curió all show performance close to the random expected score of 20%. This is interesting since we know that a pos-



**Table 1.** Results for various LLMs in both BRoverbs tasks. The table also shows the model size and size of pretrain for open source models.

Model	Model size	Pretrain size	PtS	StP	Average
Large Commercial Models					
GPT-4o [Hurst <i>et al.</i> , 2024]	-	-	0.97	0.98	0.97
Claude-3.5-sonnet [Anthropic, 2024b]	-	-	0.96	0.98	0.97
Sabiá-3 [Abonizio <i>et al.</i> , 2024]	-	-	0.96	0.96	0.96
Small Commercial Models					
GPT-4o-mini [Hurst <i>et al.</i> , 2024]	-	-	0.94	0.97	0.95
Claude-3.5-haiku [Anthropic, 2024a]	-	-	0.95	0.94	0.95
Sabiazinho-3 [Abonizio <i>et al.</i> , 2024]	-	-	0.94	0.96	0.95
Medium open source models					
Qwen 2.5 14b [Yang <i>et al.</i> , 2024]	14B	18T	0.92	0.94	0.93
Qwen 2.5 7B [Yang <i>et al.</i> , 2024]	7B	18T	0.83	0.83	0.83
Llama3-8B [Dubey <i>et al.</i> , 2024]	8B	15T	0.48	0.75	0.61
Llama2-7B [Touvron <i>et al.</i> , 2023b]	7B	2T	0.22	0.25	0.23
Llama1-7B [Touvron <i>et al.</i> , 2023a]	7B	1T	0.21	0.24	0.22
Sabiá-7B [Pires <i>et al.</i> , 2023]	7B	1T + 10B	0.23	0.33	0.28
Small Open source models					
Qwen 2.5 3B [Yang <i>et al.</i> , 2024]	3B	18T	0.51	0.68	0.59
Qwen 2.5 1.5B [Yang <i>et al.</i> , 2024]	1.5B	18T	0.41	0.57	0.49
Tucano 2.4B [Corrêa <i>et al.</i> , 2024]	2.4B	500B	0.19	0.20	0.19
Tucano 1.1B [Corrêa <i>et al.</i> , 2024]	1.1B	250B	0.19	0.20	0.19
TinyLlama 3T [Zhang <i>et al.</i> , 2024]	1.1B	3T	0.19	0.20	0.19
TinyLlama 1T [Zhang <i>et al.</i> , 2024]	1.1B	1T	0.18	0.22	0.20
Curió 1.1B [Almeida <i>et al.</i> , 2025b]	1.1B	1T + 150B	0.20	0.20	0.20

**Figure 6.** Accuracy of Curió in Broverbs tasks through training

itive score is possible for this model size, since Qwen 1.5B performs well above the random score. However, even with TinyLlama 3T—trained on three times more English tokens than its 1T counterpart—we see no sign of improvement. Moreover, incorporating Portuguese data does not appear to help in this scenario. Curió 1.1B, which builds on TinyLlama 1T with an additional 150B Portuguese tokens, still performs at random chance level. Likewise, Tucano 2.4B, trained from scratch on 500B Portuguese tokens, also fails to show meaningful gains.

## 4.2 Does More Portuguese Data Improve Cultural Understanding?

Since intermediate checkpoints for Curió-1.1B are available, we can track their performance throughout training. This setup is particularly relevant because Curió starts from an English-only model and transitions the training to Portuguese data, allowing us to observe the impact of the continued pre-training in our task. Given that our evaluation requires familiarity with Brazilian proverbs, we expected to see performance improvements. Figure 6 shows Curió-1.1 B’s results over time for both BRoverbs tasks.

However, the results are mixed. While PtS shows performance slightly improves, StP declines, leading to no overall gain compared to TinyLlama 1T, as well as the average score never distancing itself from the random performance of 20% accuracy. This suggests that despite training on Portuguese text, the data may not have included enough relevant information about Brazilian proverbs. Another possibility is that Curió-1.1B requires more training for this ability to emerge, as seen in some large-scale studies of emergent capabilities [Wei *et al.*, 2022].

The impact of training data quality and language exposure on task performance remains unclear. For instance, TinyLlama 3T, despite being trained on 3 trillion English tokens, performs at chance level, while Qwen 2.5 1.5B achieves notably better results after 18 trillion tokens. However, since details about Qwen’s training data are not publicly available, it’s difficult to pinpoint the exact cause of this improvement. It

could arise from the benefits of large-scale training on general reasoning capabilities, the presence of substantial Portuguese content in Qwen’s training corpus, or a combination of both factors.

The results from Sabiá-7B and Curió-1.1B suggest that models may need to reach a certain capacity threshold to handle our task effectively. This threshold doesn’t necessarily require pretraining on Portuguese data; however, once the model attains it, exposure to Portuguese—especially documents that are better aligned with the task—can significantly enhance performance.

## 5 Future Work

In our analysis, we use the amount of Portuguese data included in a model’s pretraining as a proxy for its potential exposure to Brazilian proverbs, since such proverbs are more likely to appear in Portuguese documents. A more direct approach for future work would be to measure the actual presence of Brazilian proverbs in large Portuguese corpora, such as Gigaverbo [Corrêa *et al.*, 2024] (used for training Tucano) and the Portuguese subset of Clueweb22 [Overwijk *et al.*, 2022] (used for continual pretraining of Sábia 7B).

Our results also indicate that current state-of-the-art commercial models already achieve strong performance on the current version of the dataset. To increase the difficulty of the BRoverbs benchmark, future iterations could include more than five alternatives per question. We limited the present version to five alternatives to keep manual validation feasible, as verifying a larger set of options for each question becomes increasingly challenging.

While our work introduces a benchmark and resources for evaluating Brazilian proverbs in Portuguese. Future work can use our benchmark, in addition to others such as ProverbEval [Azime *et al.*, 2024], to make a more nuanced analysis of LLMs’ abilities to understand proverbs across multiple regions and languages.

Finally, the selection of models used in this work is still limited; Future work could expand the types and sizes of models evaluated for both open and closed source models, possibly achieving further insights.

## 6 Conclusion

In this paper, we introduced BRoverbs, a new benchmark designed to evaluate LLMs’ ability to understand and associate Brazilian proverbs with corresponding stories. The dataset consists of two core tasks: Proverb to Story, where a model must select the correct story that embodies a given proverb, and Story to Proverb, where a model must identify the proverb that best represents a given story. Through these tasks, BRoverbs provides insights into how well LLMs capture linguistic and cultural nuances in Brazilian Portuguese.

Our evaluation of various LLMs revealed a significant disparity in performance across model sizes and architectures. Large commercial models such as GPT-4o and Sabiá-3 achieved near-perfect scores, while smaller open-source models struggled, with some performing close to random selection.

Notably, Qwen models demonstrated superior performance among open-source models.

Our analysis of Curió-1.1B, a model that underwent continued pretraining in Portuguese, revealed no significant gains in BRoverbs performance despite its exposure to 150B additional tokens. Meanwhile, Sabiá-7B, which underwent continued pretraining in 10B, showed slightly improved performance in the StP task. This raises questions about the role of pretraining data composition versus the overall training and model scale. Future work should explore whether specific exposure to idiomatic expressions and cultural narratives could enhance proverb comprehension.

Overall, BRoverbs advances the scenario of culturally grounded evaluation datasets for Brazil. As LLMs continue to evolve, benchmarks like BRoverbs could play a crucial role in ensuring fair and accurate assessments of model capabilities in underrepresented languages.

## Declarations

### Acknowledgements

We would like to thank Maritaca AI for providing the necessary computational resources for this research.

### Funding

The computational resources used in this Research were provided by Maritaca AI.

### Authors’ Contributions

We enumerate below the areas of contribution of the three authors: Thales Sales Almeida (TSA), Giovana Kerche Bonas (GKB), and João Guilherme Alves Santos (JGAS). We follow the CRediT taxonomy<sup>2</sup>

- **Conceptualization** — TSA
- **Data curation** — JGAS
- **Formal analysis** — GKB
- **Funding acquisition** — TSA
- **Investigation** — TSA
- **Methodology** — TSA, GKB, JGAS
- **Project administration** — TSA
- **Resources** — TSA
- **Software** — TSA, GKB, JGAS
- **Supervision** — TSA
- **Validation** — TSA
- **Visualization** — TSA, GKB, JGAS
- **Writing — original draft** — TSA, GKB, JGAS
- **Writing — review & editing** — TSA, GKB, JGAS

### Competing interests

The authors declare that they have no competing interests regarding this work.

### Availability of data and materials

The benchmark product of this study is available at HuggingFace at <https://huggingface.co/datasets/Tropic-AI/BRoverbs>.

<sup>2</sup><https://credit.niso.org/>



## References

- Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*. DOI: 10.48550/arXiv.2410.12049.
- Alfina, I., Mulia, R., Fanany, M. I., and Ekanata, Y. (2017). Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 international conference on advanced computer science and information systems (ICACSIS)*, pages 233–238. IEEE. DOI: 10.1109/icacsis.2017.8355039.
- Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*. DOI: 10.48550/arXiv.2403.09887.
- Almeida, T. S., Bonás, G. K., Santos, J. G. A., Abonizio, H., and Nogueira, R. (2025a). Tiebe: A benchmark for assessing the current knowledge of large language models. *arXiv preprint arXiv:2501.07482*. DOI: 10.48550/arXiv.2501.07482.
- Almeida, T. S., Laitz, T., Bonás, G. K., and Nogueira, R. (2023). Bluex: A benchmark based on brazilian leading universities entrance exams. In *Brazilian Conference on Intelligent Systems*, pages 337–347. Springer. DOI: 10.1007/978-3-031-45368-7\_2.
- Almeida, T. S., Nogueira, R., and Pedrini, H. (2025b). Building high-quality datasets for portuguese llms: From common crawl snapshots to industrial-grade corpora. *To Appear*. DOI: 10.48550/arXiv.2509.08824.
- Anthropic (2024a). Introducing claude 3.5 haiku. Available at: <https://www.anthropic.com/claude/haiku>.
- Anthropic (2024b). Introducing claude 3.5 sonnet. Available at: <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Azime, I. A., Tonja, A. L., Belay, T. D., Chanie, Y., Balcha, B. F., Abadi, N. H., Ademtew, H. B., Nerea, M. A., Yadeta, D. D., Geremew, D. D., *et al.* (2024). Proverbval: Exploring llm evaluation challenges for low-resource language understanding. *arXiv preprint arXiv:2411.05049*. DOI: 10.18653/v1/2025.findings-naacl.350.
- Baucells, I., Aula-Blasco, J., de Dios-Flores, I., Suárez, S. P., Pérez, N., Salles, A., Docio, S. S., Falcão, J., Saiz, J. J., Sepúlveda-Torres, R., *et al.* (2025). Iberobench: A benchmark for llm evaluation in iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519. Available at: <https://aclanthology.org/2025.coling-main.699/>.
- Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*. DOI: 10.48550/arxiv.1712.08917.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., *et al.* (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. DOI: 10.48550/arxiv.2107.03374.
- Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024). Tucano: Advancing neural text generation for portuguese. *arXiv preprint arXiv:2411.07854*. DOI: 10.1016/j.pat-ter.2025.101325.
- Delfino, P., Cuconato, B., Haeusler, E. H., and Rademaker, A. (2017). Passing the brazilian oab exam: data preparation and some experiments. In *Legal knowledge and information systems*, pages 89–94. IOS Press. DOI: 10.3233/978-1-61499-838-9-89.
- Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. (2019). Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*. DOI: 10.48550/arxiv.1903.00161.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. DOI: 10.48550/arXiv.2407.21783.
- Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., *et al.* (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104. DOI: 10.18653/v1/w19-3510.
- Giagkou, M., Lynn, T., Dunne, J., Piperidis, S., and Rehm, G. (2023). European language technology in 2022/2023. In *European Language Equality: A Strategic Agenda for Digital Language Equality*, pages 75–94. Springer. DOI: 10.1007/978-3-031-28819-7\_4.
- Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*. DOI: 10.18653/v1/2021.findings-acl.413.
- Ho, X., Nguyen, A.-K. D., Sugawara, S., and Aizawa, A. (2020). Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*. DOI: 10.18653/v1/2020.coling-main.580.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., *et al.* (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. DOI: 10.48550/arxiv.2410.21276.
- Jiang, Z., Anastasopoulos, A., Araki, J., Ding, H., and Neubig, G. (2020). X-factr: Multilingual factual knowledge retrieval from pretrained language models. *arXiv preprint arXiv:2010.06189*. DOI: 10.18653/v1/2020.emnlp-main.479.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., *et al.* (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. DOI: 10.1162/tacl.00276.
- Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*. DOI: 10.18653/v1/d17-1082.
- Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., and Caridá, V. (2023). Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*. DOI: <https://doi.org/10.48550/arxiv.2308.11878>.

- Li, Y., Wang, S., Ding, H., and Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382. DOI: 10.1145/3604237.3626869.
- Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., *et al.* (2024). Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*. DOI: h10.48550/arxiv.2404.01268.
- Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., Mahari, R., Dey, M., Hamdy, M., Saxena, N., *et al.* (2024). Bridging the data provenance gap across text, speech and video. *arXiv preprint arXiv:2412.17847*. DOI: 10.48550/arXiv.2412.17847.
- Mi, M., Villavicencio, A., and Moosavi, N. S. (2024). Rolling the dice on idiomaticity: How llms fail to grasp context. *arXiv preprint arXiv:2410.16069*. DOI: 10.18653/v1/2025.acl-long.362.
- Moayeri, M., Tabassi, E., and Feizi, S. (2024). Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228. DOI: 10.1145/3630106.3658967.
- Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., *et al.* (2024). Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146. DOI: 10.48550/arxiv.2406.09948.
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*. DOI: 10.1145/3744746.
- 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*, pages 3360–3362. DOI: 10.1145/3477495.3536321.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. (2021). Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*. DOI: 10.18653/v1/2022.findings-acl.165.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. pages 226–240. DOI: 10.1007/978-3-031-45392-2\_15.
- Potts, C., Wu, Z., Geiger, A., and Kiela, D. (2020). Dynasent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*. DOI: 10.18653/v1/2021.acl-long.186.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. DOI: 10.18653/v1/d16-1264.
- Rane, N. L., Tawde, A., Choudhary, S. P., and Rane, J. (2023). Contribution and performance of chatgpt and other large language models (llm) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, 5(10):875–899. DOI: 10.56726/irjmet.45213.
- Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*. DOI: 10.18653/v1/n18-2002.
- 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. IEEE. DOI: 10.1109/bracis.2019.00084.
- Silveira, I. C. and Mauá, D. D. (2017). University entrance exam as a guiding test for artificial intelligence. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 426–431. IEEE. DOI: 10.1109/bracis.2017.44.
- Srivastava, A., Rastogi, A., Rao, A., Shoen, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., *et al.* (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*. DOI: h10.48550/arxiv.2206.04615.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*. DOI: 10.18653/v1/n18-1074.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., *et al.* (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv: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., *et al.* (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. DOI: 10.48550/arxiv.2307.09288.
- Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., and Pardo, T. A. S. (2021). Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*. DOI: 10.48550/arxiv.2103.14972.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Super-glue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32. DOI: 10.48550/arxiv.1905.00537.
- Watts, I., Gumma, V., Yadavalli, A., Seshadri, V., Swaminathan, M., and Sitaram, S. (2024). Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv:2406.15053*. DOI: 10.18653/v1/2024.emnlp-main.451.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., *et al.* (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*. DOI: 10.48550/arxiv.2206.07682.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., *et al.* (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*. DOI: 10.48550/arXiv.2412.15115.
- Yu, W., Jiang, Z., Dong, Y., and Feng, J. (2020). Reclor: A reading comprehension dataset requiring logi-

cal reasoning. *arXiv preprint arXiv:2002.04326*. DOI: 10.48550/arxiv.2002.04326.

Zhang, P., Zeng, G., Wang, T., and Lu, W. (2024). Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*. DOI: 10.48550/arxiv.2401.02385.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., *et al.* (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2). DOI: 10.48550/arxiv.2303.18223.

## A Appendix

### A.1 Carbon emission estimates

As mentioned before, we used an RTX A6000 to perform our evaluations; all experiments ran in under 5 hours. By using the ML CO<sub>2</sub> Impact calculator<sup>3</sup>, we arrive at an estimated carbon emission of 0.85 Kg of CO<sub>2</sub>.

### A.2 Prompts for story generation

This Appendix presents the prompt that was employed to generate the short stories in the dataset.

Você receberá um provérbio brasileiro e deverá criar três situações narrativas curtas que ilustrem a ideia desse provérbio.

Regras:

1. A situação deve ser específica e realista, como um pequeno trecho de uma história.
2. Não mencione o provérbio ou palavras-chave dele diretamente. A ideia deve ser transmitida indiretamente.
3. A história deve ter um personagem e uma ação, mostrando uma consequência ou aprendizado.
4. O formato de saída deve ser assim:

```
{
  "proverbio": "",
  "explicacao": "",
  "historia_curta_1": "",
  "historia_curta_2": "",
  "historia_curta_3": ""
}
```

"proverbio": repita o mesmo provérbio que recebeu.

"explicacao": explique o sentido ou a moral do provérbio de forma curta.

"historia\_curta": crie uma história curta em português com um ou mais personagens que vivenciem uma situação que ilustre a ideia principal do provérbio. Não cite o provérbio na história.

Não inclua nada além do JSON de resposta. Por exemplo, não inclua texto fora do objeto JSON.

Exemplos

...

Agora, gere a resposta para este provérbio:

```
{proverbio}
```

**Figure 7.** Prompt completo utilizado para geração das histórias (redimensionado).

<sup>3</sup><https://calculator.linkeddata.es/>