Publication status: Not informed by the submitting author

Jabuticaba: The largest commercial corpus for LLMs in Portuguese

Marcellus Amadeus, William Alberto Cruz Castaneda, José Roberto Homeli da Silva, Rodrigo Scotti

https://doi.org/10.1590/SciELOPreprints.12696

Submitted on: 2025-07-18
Posted on: 2025-08-05 (version 1)
(YYYY-MM-DD)

JABUTICABA: THE LARGEST COMMERCIAL CORPUS FOR LLMS IN PORTUGUESE.

Marcellus Amadeus, William Alberto Cruz Castañeda, José Roberto Homeli da Silva, and Rodrigo Scotti

SoberanIA

{marcellus,william,homeli,scotti}@soberania.ai

https://orcid.org/0009-0002-7777-2562
https://orcid.org/0000-0002-9803-1387
https://orcid.org/0000-0002-8825-2362
https://orcid.org/0000-0002-9937-0129

ABSTRACT

Large Language Models provide a step towards intelligent communication systems by harnessing large repositories or datasets of written human knowledge to better predict and understand the world. However, Artificial Intelligence sovereignty is all about quality data because datasets serve as the foundational infrastructure that sustains the development of LLMs. Thus, this paper presents the Jabuticaba dataset, the most extensive Portuguese language corpus for LLMs with a total data size of 669 GB and over 139 billion tokens[1] consisting of clean, deduplicated words ready for use, including commercial use. Furthermore, Jabuticaba achieves a size comparable to and exceeding some state-of-the-art (SOTA) datasets in other languages. This paper outlines the methodological pipeline details used to build it to serve as a comprehensive reference for the research community in academia and industry in this field, as well as contributing to future studies. Resources are freely available at HuggingFace: `https://huggingface.co/datasets/soberania/jabuticaba`.

Keywords Datasets · Large Language Models · Artificial Intelligence · Natural Language Processing

[1]Token count calculated using OpenAI's tiktoken tokenizer.

1 Introduction

The emergence of Large Language Models (LLMs) has provoked a revolutionary transformation in all industries. The release of open-sourced LLMs, such as the Llama family, Phi, Qwen, Falcon, and others, has attracted increasing research attention and has become a hot research field. The benefits of these open-source models involve fine-tuning domain-specific information, cost-reduction, and collaborative-friendly business environments[1]. Nonetheless, the potential demonstrated by LLMs can be attributed, in part, to the datasets used for training and testing. Without high-quality datasets as the foundation, it is challenging to enable LLMs to perform exceptionally in text generation, natural language processing tasks, or other domains deployed.

For example, OpenAI trained GPT-3 [2] with WebText2's 19 billion byte-pair-encoded tokens (bpet), Books1's 12 billion bpet, Books2's 55 billion bpet, Wikipedia's 3 billion bpet and 45 TB of compressed plaintext from Common Crawl before filtering and 570 GB after filtering, equivalent to 400 billion bpet. Thus, curating and annotating a dataset brings a considerable challenge as (i) data scarcity in some domains results in imbalanced datasets that affect the model's ability to infer correctly; (ii) data annotation is required when fine-tuning LLMs for downstream tasks; (iii) concerns regarding the quality and biases of the training data and the potential for the unintentional dissemination of detrimental or inaccurate information; and (iv) tokenization sensitivity to input phrasing can lead to unintended outcomes when generating text, such as adversarial assaults and output variations based on minute input changes [3] [4].

Therefore, the construction and analysis of LLM datasets is an area worthy of attention, due to the composition and quality of these datasets influencing the performance of LLMs. If an LLM is trained with questionable datasets, they will be impacted by performance issues like bias and overfitting. Conversely, training an LLM with high-quality datasets enables a more accurate and coherent output[5]. Nowadays, companies around the world have realized that an LLM needs more than SOTA models and training methods. For example, Databricks-dolly-15k [6], an open-source dataset, contains 15,000 high-quality human-generated prompt/response pairs designed for instruction tuning Large Language Models. RedPajama-V2 [7], another open dataset for training Large Language Models, includes over 100B text documents coming from 84 CommonCrawl snapshots and processed using the CCNet [8] pipeline. Bloomberg [9] trained a transformer architecture from scratch with decades' worth of carefully curated financial data. The resulting BloombergGPT allows the financial company to empower its clients and perform existing financial-specific NLP tasks faster and with more accuracy. Likewise, Bigcode-Project[2] has developed a programmer-friendly model StarCode [10] by training it on code in different programming languages gathered from GitHub.

Thus, the objective of this paper is to present the Jabuticaba dataset and also provide a description of the methodological pipeline steps to build it. The remainder of this paper is organized as follows. Section 2 presents the Jabuticaba methodology pipeline, outlining all the steps required to build it. Section 3 provides the corpus statistics and content on the collected data. Section 4 presents the conclusions and future research directions.

2 Related Works

The Instruct-PTBR-ENUS-11M[3] dataset is a mix of several instruct datasets found on HuggingFace (with 11,165,249 rows). Includes tasks such as RAG-focused question-answering, summarization, and keyword generation, among others. Most of the original dataset was in English (5,239,163 rows) and has been translated into Brazilian Portuguese (5,926,086 rows). The authors warn that the translation may contain errors. Other available datasets are ASSIN, ASSIN2, Mac-Morpho, and NURC-SPCorpusMinimo[4].

OLID-BR [11] is a high-quality offensive language identification dataset for Brazilian Portuguese. It contains 6,354 comments labeled and compatible with other languages datasets to enable training of multilingual models.The work of [12] proposes the creation of a large-scale linguistic corpus for Brazilian Portuguese, composed of publications collected from the social network Twitter. Carolina [13] is a large open corpus of texts in Brazilian Portuguese that is compiled from web-based and typologically diverse sources. The first public version has 653 million tokens, distributed in 7 types.

HateBR [14] is the first large-scale, expert-annotated corpus of Brazilian Instagram comments to detect hate speech and offensive language. The corpus was collected from the comments section of Brazilian politicians' accounts. Consists of 7,000 documents annotated according to three different layers: a binary classification, an offensiveness-level classification, and nine hate speech groups. UlyssesNER-Br [15] is a corpus of Brazilian legislative documents for NER. The corpus is composed of bills and legislative consultations from the Chamber of Deputies.

[2]https://www.bigcode-project.org/

[3]https://huggingface.co/datasets/cnmoro/Instruct-PTBR-ENUS-11M

[4]https://huggingface.co/nilc-nlp

The Brazilian Portuguese Web Corpus (brWaC) dataset [16] is one of the most popular in Portuguese due to its size and variety of content. This dataset was built by adopting an approach called WaCky (Web-As-Corpus KoolYinitiative), which can obtain data from multiple languages by extracting content from the Web. Contains over 3.5 million pages from 120,000 websites and over 2.7 billion tokens. Although this is one of the largest datasets in Portuguese and the content has been filtered, the quality of many instances is questionable due to poor writing and offensive/racist content.

ToLD-Br [17] is a toxic language dataset for Brazilian Portuguese composed of Twitter posts. A total of 21K tweets were manually annotated into seven categories: non-toxic, LGBTQ+phobia, obscene, insult, racism, misogyny, and xenophobia.

The work of [18] presents four textual datasets for language modeling in Brazilian Portuguese. MultiWOZ-PTBR is a direct translation dataset from the original Multi-WOZ. Cleaned BrWaC is a cleaned dataset version from BrWaC, which contains 96814 domains and 3166108 pages. MCCD Generated is a human conversational dataset generated using the methodology MCCD. Miner-XenForo2 derives from two different datasets: a) Adrenaline dataset based on a web Forum about technology, hardware, and games, and b) OuterSpace dataset based on a web Forum about games on different platforms and millions of generic messages.

GigaVerbo [19] is a dataset comprising 780 GB of text in Portuguese, being a concatenated version of several datasets available in HuggingFace. Contains over 200 billion unfiltered tokens. The token count without accounting for the repetition factor of the filtered portion of GigaVerbo is 129 billion tokens. It covers several sources, including crawled websites, articles, translated conversations, and legal documents.

3 Jabuticaba Methodological Pipeline

After collecting a large amount of text data, it is essential to preprocess the data for constructing a pre-training corpus for LLMs. Tasks such as removing noisy, redundant, irrelevant, and potentially toxic data affect the capacity and performance of LLMs. This section describes the methodological pipeline steps used to build the Jabuticaba dataset. The proposed methodological pipeline is based on recommendations raised in [3] and [20]. Figure 1 shows an integrated overview of all steps, which will be discussed in detail for the remainder of this section. Please note that certain tools may experience significant slowdowns due to their computational requirements, rendering them less suitable for large datasets. Additionally, not every step is required or feasible for every dataset, and therefore adaptability is crucial to tailor the approach according to the specific application domain.

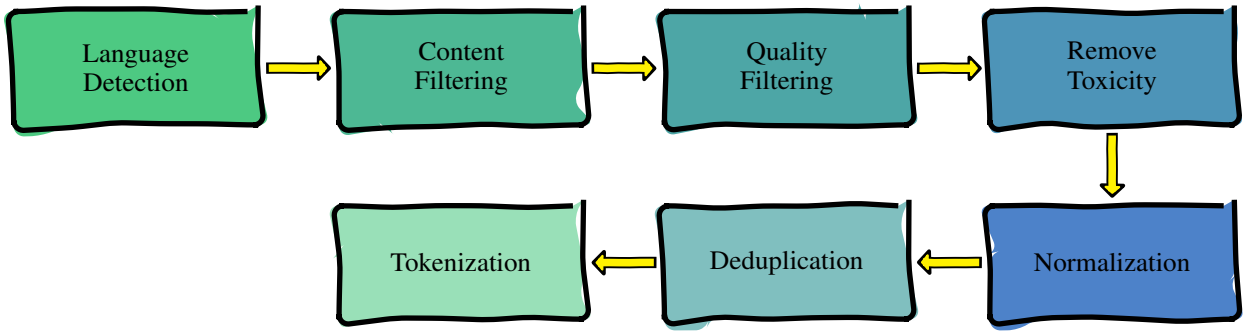

Figure 1: Jabuticaba methodological pipeline

3.1 Language Detection

Jabuticaba is mainly formed from Brazilian Portuguese, but it may contain European Portuguese. The search to structure the corpus was carried out on websites such as Linguateca, Kaggle, HuggingFace, and some university websites. A focus was given on natural written data[5], some oralized texts were included, e.g. TedTalks transcriptions[6]. Our search expanded to any variety of Portuguese (European or African). The corpora went through a selection based on some criteria, which are size (in MB), synchrony represented, and automatic translation. Each corpus received an ID annotation, domain, and, when available, author, date, and URL. The tools that enabled language detection were: (a)

[5]Texts that were written or planned somehow and someone has spoken them out.

[6]the dataset used already contained texts in Portuguese.

fasttext[7] model applied over each corpus and filtering out not Portuguese corpora; (b) the language-detection library *langdetect*[8] that can detect the language of texts with over 99% precision for 49 languages; and (c) *Lingua*[9] to produce accurate results on both long and short text, even on single words and phrases with more languages taking part in the decision process.

3.2 Content Filtering

This step begins by filtering out non-Portuguese documents. At this stage, pages that did not pass the established filters or that incorporated explicit content were removed. For specific tasks, fine-tuning and specific corpora are delimited genres and data sources when crawling the data. To do this, we delimited the domains we were interested in the crawler. Also, it removed URLs from the corpus based on a blacklist[10]. For filtering undesirable content, [21] suggests that documents with "bad words" be removed. One word-based filter tool example is the *List of Dirty, Naughty, Obscene, and Otherwise Bad Words*[11]. For domain list filter tools, examples are *pi-hole-blocklist*[12] (for blocking ads and trackers on the Internet that acts as a DNS service), *Bon-Appetit/porn-domains*[13] (collection of domains used for explicit adult content like porn websites), and *hosts*[14] (contain several reputable host files and merge them into a unified host file with duplicates removed.)

3.3 Quality Filtering

Most of the texts found on the web have insufficient quality to be useful for training an LLM. Many web pages contain primarily automatically generated content, text not intended for human consumption or from social media, which may lack context. Therefore, this step removes low-quality data by applying the necessary heuristic filters to remove any document that does not contain between 50 and 100,000 words, or whose average word length is outside the range of 3 to 10 characters. Removes any document with a symbol-to-word ratio greater than 0.1 (10%) for either the hash symbol ('#') or the ellipsis ('. . . '). Removes any document with more than 90% of lines starting with a bullet point, or more than 30% ending with an ellipsis. Requires that 80% of the words in a document contain at least one alphabetic character, and apply a "*stop word*" filter, to remove documents that do not contain at least two of the following words: the, be, too, of, and, that, have, with; this adequately deals with documents that contain no coherent text. Removes documents containing "*lorem ipsum*". This string is indicative that the document contains only placeholder text. Also, were used heuristic-based approaches to eliminate low-quality texts through a set of well-designed rules, which can be summarized as follows:

- **Metric based filtering.** Evaluation metrics about the generated. *Perplexity*[15] can be employed to detect and remove unnatural sentences.
- **Statistic based filtering.** Includes statistical features of a corpus, e.g., the punctuation distribution, symbol-to-word ratio, sentence length, minimum word count, minimum mean word size, maximum symbol-to-word ratio, symbols at the beginning and end of sentences, minimum ratio of alphanumeric characters, checking for the presence of required words, which can be utilized to measure the text quality and filter the low-quality data.
- **Keyword-based filtering.** Based on a specific keyword set, the noisy or unuseful elements in the text, such as HTML tags, hyperlinks, boilerplates, and offensive words, can be identified and removed.

Another indicator of poor-quality data is an excessive repetition of certain words or phrases within a document. Excessive repetition is often linked with uninformative content. Furthermore, a well-studied failure mode of current language models is to repeat themselves during sampling, which may be partially attributed to repetitious training data. Some tasks performed are:

[7] https://fasttext.cc/docs/en/language-identification.html

[8] https://github.com/Mimino666/langdetect

[9] https://github.com/pemistahl/lingua-py

[10] https://github.com/mhhakim/pihole-blocklist/raw/master/list.txt, https://blocklistproject.github.io/Lists/alt-version/abuse-nl.txt, https://blocklistproject.github.io/Lists/alt-version/gambling-nl.txt, https://blocklistproject.github.io/Lists/alt-version/malware-nl.txt, https://blocklistproject.github.io/Lists/alt-version/scam-nl.txt, https://blocklistproject.github.io/Lists/alt-version/porn-nl.txt

[11] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

[12] https://github.com/zangadoprojets/pi-hole-blocklist

[13] https://github.com/Bon-Appetit/porn-domains

[14] https://github.com/StevenBlack/hosts

[15] Perplexity needs to run a model, so it can be slow. Furthermore, it is advisable to ensure that the model used is well-suited for the specific application at hand, such as being trained in the same language. It may be more appropriate to use perplexity for evaluating the model rather than the data itself.

- To ensure the quality and uniqueness of our documents, we use multiple approaches to identify and remove those with high proportions of repeated lines, paragraphs, or *n-grams*. As well as those containing short or long duplicate passages, making sure to identify both types by using multiple approaches to calculate the proportion of duplicate content.
- For lines and paragraphs separately, we calculate over the document both the fraction that are duplicates, and the fraction of characters contained within those duplicates.
- For each $n \in \{2, \ldots, 4\}$, we calculated the fraction of characters contained within the most frequent *n-gram*; and for each $n \in \{5, \ldots, 10\}$, we calculated the fraction of characters contained within all duplicate *n-grams*, not counting characters that occur in overlapping *n-grams* more than once. Thus, it filters out documents whose duplicate content surpasses any of the thresholds detailed in Table 1. Those thresholds are a criterion to avoid injecting data that will cause overfitting of LLMs during training [22].

Measurement	Threshold
Duplicate line fraction	0.30
Duplicate paragraph fraction	0.30
Duplicate line character fraction	0.20
Duplicate paragraph character fraction	0.20
Top *2-gram* character fraction	0.20
Top *3-gram* character fraction	0.18
Top *4-gram* character fraction	0.16
Duplicate *5-gram* character fraction	0.15
Duplicate *6-gram* character fraction	0.14
Duplicate *7-gram* character fraction	0.13
Duplicate *8-gram* character fraction	0.12
Duplicate *9-gram* character fraction	0.11
Duplicate *10-gram* character fraction	0.10

Table 1: Thresholds for repetitious text for each measurement of text repetition.

3.4 Remove Toxicity

This step removes toxicity using the *detoxify*[16] model that returns a filtered dataset with entry texts whose toxicity is below an established threshold of 0.2.

3.5 Normalization

This step converts into human-readable texts and excludes documents made up of less than 200 unique tokens. The *ftfy*[17] library helps to fix Unicode that's broken in various ways. Takes in bad Unicode and outputs good Unicode. Normalization techniques in *Byte Pair Encoding*, such as NFKC (Unicode normalization algorithm), may degrade the tokenization performance.

3.6 Deduplication

This step removes all exact duplicates to obtain a set of unique documents. In addition to exact duplicates, there are many documents with significant *n-gram* overlap, and exact substring matches can be efficiently identified by the *ExactSubstr* algorithm described in [23]. There is an implementation in this remote repository[18]. This procedure is applied to identify repeated substrings within documents where, for example, two documents may be distinct and still contain repeated substrings. Additionally, approximate matches can be found through the *MinHash* algorithm to compute *13-gram Jaccard* similarities to determine which documents are near-duplicates of each other. The algorithm defines two documents to be similar when their *Jaccard* similarity exceeds 0.8 and randomly removes one of them. [23] suggests further comparing the pairs through *edit distance* similarity before deleting one of the possibly repeated occurrences. That's why it is important to normalize whitespaces and ignore punctuation when constructing the *n-grams*. Another important characteristic is privacy redaction, which removes the personally identifiable information (PII) from

[16]https://github.com/unitaryai/detoxify

[17]https://pypi.org/project/ftfy

[18]https://github.com/google-research/deduplicate-text-datasets

the pre-training corpus. One direct and effective approach is to employ rule-based methods, such as keyword spotting, to detect and remove PII, such as names, addresses, and phone numbers.

3.7 Tokenization

This step aims to segment raw text into sequences of individual tokens, which are subsequently used as the inputs of LLMs. The use of a tokenizer specially designed for the pre-training corpus can be highly beneficial, especially for the corpus that consists of diverse domains, languages, and formats. Therefore, several recent LLMs train their customized tokenizers. Here, OpenAI's tokenizer *tiktoken*[19] was used just for token counting purposes.

4 Dataset Analysis

According to [24], Jabuticaba is classified as **general pre-training corpora**, which comprises large-scale text data mixtures from different domains and topics to provide universal language knowledge and data resources for NLP tasks. Thus, Jabuticaba as a pre-training corpus influences the direction of pre-training and the potential of models in the future, playing several pivotal roles in providing generality, enhancing generalization ability, elevating performance levels, and supporting multilingual processing. Jabuticaba is a large-scale dataset composed of extensive text from diverse domains and sources, as listed in Table 2. Also, Table 5 shows the distribution of corpora alongside their domains.

Major Class	Domain	Content
Arts	Literature	books, poetry, novels.
Documents	Academic	papers, journals, dissertations, conference proceedings.
Documents	Legal	contracts, legislation, court rulings.
Internet	Informational	blog pages, wiki pages, how-to guides.
Internet	User-Generated Content	forum posts, personal blogs, reviews, opinion pieces.
Internet	Web Scraping	varied content scraped from the internet.
Research	NER	manual data created for the NER task.

Table 2: Taxonomy of major classes, domains, and their prototypical contents featured in the Jabuticaba dataset.

Their primary characteristic is that the text content is not confined to a single domain, making it suitable for training general foundational models. As presented in Table 2, the data types are categorized into five major classes: Arts, Documents, Internet, Media, and Research. The collected and organized information about Jabuticaba corpus is an aggregation of multiple public datasets, and Table 5 in appendix A presents the complete list of corpora contained in Jabuticaba alongside other information.

When comparing Jabuticaba regarding its size with the public, and open-source multilingual datasets from the list of 49 general pre-training corpora [24], it is observed in Figure 2 that Jabuticaba ranks fifth. CulturaX [25] is a multilingual dataset with 6.3 trillion tokens in 167 languages, which merges the latest version of mC4 with all available OSCAR corpora, encompassing distributions 20.19, 21.09, 22.01, and 23.01.

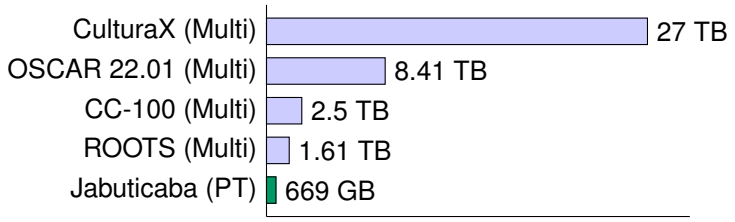

Figure 2: Jabuticaba size comparison with the state-of-the-art general pre-training public and open-source multilingual datasets. Multi indicates Multilingual, and PT indicates Portuguese.

[19] https://github.com/openai/tiktoken

OSCAR 2201[20] is the OSCAR version from January 2022, the November/December 2021 dump of Common Crawl. It is document-oriented, which means that rather than extracting lines and sorting them in language subcorpora, we identify documents as a whole. The main differences are that sentences in a document are contiguous and should make sense one after another, but sentences are not guaranteed to be of the subcorpus' language. CC-100 [26] attempts to recreate the dataset used for training XLM-R comprising monolingual data for 100+ languages and also includes data for Romanized languages. It was constructed using the URLs and paragraph indices provided by the CC-Net repository by processing January-December 2018 Commoncrawl snapshots. ROOTS [27] spans 59 languages that were used to train the 176-billion-parameter BigScience Large Open-science Open-access Multilingual (BLOOM) language model.

On the other hand, Figure 3 shows that when comparing Jabuticaba only with the Portuguese portion of the CulturaX dataset, it is found that Jabuticaba is superior by 2B more tokens, from 139B to 136.94B respectively. In this way, Jabuticaba ranks first. It is important to highlight that when comparing the Jabuticaba methodological pipeline with the CulturaX, OSCAR, and CC-100 pipelines, Jabuticaba covers the same foundational steps but also improves upon them to create a high-quality dataset exclusively in Portuguese. Appendix B presents the pipeline summarization of these datasets for comparison purposes but the full papers of each pipeline step can be consulted in [25] [27][26].

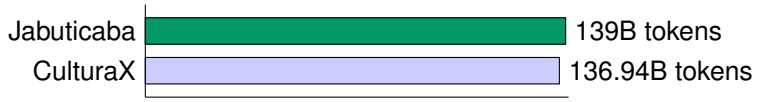

Figure 3: Jabuticaba tokens comparison between the corresponding Portuguese language part of the multilingual CulturaX dataset.

Thereafter, with the corpora previously compiled, each corpus was divided into documents and saved in JSON files of around 500 MB. All corpora were concatenated and divided into JSON Lines files up to 200 MB each. The methodological pipeline steps were applied to the JSON files. Table 3 describes the structure of the JSON files of the dataset.

Dataset	Description
Jabuticaba	90 billion words, 175 million lines, ≈ 3.5k JSON lines (.jsonl) files with up to 200MB each, 669 GB total. Token count: ≈ 139B (tiktoken).

Table 3: Description of JSON file structure of the Jabuticaba dataset.

Each instance of the dataset has the data fields shown in Table 4. A JSON representation is in Listing 1, which shows an example of an instance in JSON format.

Key/Field	Type	Note
source	str	The Identifier of the original corpus
domain	str	Domain type defined in the Domain taxonomy
text	str	The actual text
char_count	int	Character count using len(str)
word_count	int	Word count using str.split()

Table 4: Data fields for each instance of the Jabuticaba dataset.

5 Discussion

5.1 Social Impact

The Jabuticaba corpus holds significant potential for advancing Portuguese language models, especially for LLMs aimed at Brazilian markets. By providing access to a large, clean, and deduplicated corpus, Jabuticaba can foster innovation in a myriad of applications, such as conversational AI, language translation, and text generation, benefiting industries, education, and technology development. However, it is essential to consider the societal consequences of

[20]https://oscar-project.github.io/documentation/versions/oscar-2201/

```
{
   'source': 'ALC4',
   'domain': 'Internet / Web scraping',
   'text': 'Chevrolet Equinox\nAproveite que o novo carro SUV esportivo da Chevrolet está à venda
   ↪  na concessionária Palazzo.',
   'char_count': 112,
   'word_count': 17
}
```

Listing 1: JSON data instances example of the Jabuticaba dataset.

large-scale language models trained on such data. While there is the opportunity to improve access to information and resources in Portuguese, there is also a risk of amplifying existing biases present in the source texts (particularly considering domain and corpus distribution).

Furthermore, the Jabuticaba corpus has the potential to democratize technological advancements in underserved Portuguese-speaking regions, particularly in Brazil, where access to localized AI tools has been limited. Offering an extensive and diverse dataset collected from real language use can empower developers and researchers to create more inclusive language models that reflect the rich linguistic and cultural diversity of the Portuguese language. This could significantly reduce the language barriers often encountered in global technology platforms, allowing more individuals and organizations to access advanced digital tools in their native language. Additionally, by fostering the development of AI systems that understand and process Portuguese, Jabuticaba can enable governments, non-profits, and local businesses to deploy AI-driven solutions that address specific social, educational, and healthcare challenges within Portuguese-speaking communities. However, achieving these positive outcomes requires that both developers and stakeholders actively consider the societal implications of these technologies, such as the impact on job displacement and the digital divide.

5.2 Biases

Given that the Jabuticaba corpus is composed of diverse datasets scraped from various domains such as news, legal documents, user-generated content, and academic papers, it may inherit preexisting biases from these sources. Biases related to gender, race, socio-economic status, and regional disparities may be reflected in the data. For instance, media sources might prioritize certain viewpoints, and user-generated content may display social biases prevalent in the contributing population. Mitigating such biases is an ongoing challenge. Users of this dataset should be aware of the potential for unintentional reinforcement of stereotypes and discriminatory language, particularly in tasks such as sentiment analysis or text generation. Applying bias detection methods and fairness evaluation tools is recommended when making use of this corpus for model training.

To effectively address the biases that might arise from the Jabuticaba corpus, it is crucial to implement bias mitigation techniques throughout the entire model training and deployment process. This could include the use of debiasing algorithms, the diversification and/or normalization of training data, and regular audits of AI outputs to ensure fairness across different demographic groups. Additionally, the transparent documentation of the dataset's sources and its potential biases can help users make informed decisions when utilizing the data for NLP tasks. Encouraging interdisciplinary collaborations between data scientists, sociologists, and linguists could also prove beneficial in identifying subtle biases that may not be immediately evident but could have long-term societal impacts.

5.3 Limitations

While Jabuticaba aims to provide the largest Portuguese corpus for LLMs, several limitations must be acknowledged. First, the dataset's reliance on publicly available sources may result in incomplete coverage of certain language varieties, including regional dialects and informal speech prevalent in oral communication, which would be key for Sociolinguistic goals. Regional varieties of register variation (oral transcription or originally written data) may not be easily identified, as entries are not labeled as such. Current literary style is also not widely covered, as most publications are still not in the public domain, and we were not prone to including earlier literary works, considering language varies significantly across time. Moreover, deduplication efforts may not eliminate all redundant content, and some noise from web-scraped

data could persist. Researchers and developers should also consider the variability in tokenization results depending on the tokenizer used, as different models may produce varying token counts. Lastly, the corpus predominantly reflects written language, and its applicability to tasks involving spoken language may be constrained.

The Jabuticaba corpus, while extensive, also highlights the challenge of ensuring linguistic diversity in AI development. To address the lack of representation of informal speech and regional dialects, future iterations of the corpus could benefit from targeted data collection initiatives that focus on these underrepresented forms of language. Additionally, techniques such as active learning or data augmentation could be used to enrich the dataset with syntactically and semantically diverse samples. Given the dynamic nature of the internet and written language, maintaining an up-to-date corpus will require ongoing efforts to update and refine the dataset. Developers should also be mindful of the varying levels of noise and inconsistencies in web-scraped data, using sophisticated filtering techniques to preserve data quality while minimizing redundancies and errors.

Acknowledgments

We would like to thank our internal reviewers: Alfredo J. G. Martinez, Pedro A. Santos Neto, Raimundo Moura, Rafael T. Anchiêta, and Rogério F. de Sousa for taking the time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions that helped us improve the quality of the manuscript. In addition, we truly appreciate the support from the Piauí Institute of Technology (PIT) and the Piauí government for making this publication possible.

Declarations

Authors' Contributions

All authors contributed equally to the conceptualization, formal analysis, investigation, methodology, and writing of this study. All authors read and approved the final manuscript.

Competing interests.

The authors declare no conflicts of interest.

Appendices

A Complete List of corpora included in Jabuticaba Dataset

Name	Identifier	Domain	Words	Percentage	License
C4	ALC4	Internet / Web scraping	68,238,423,781	75.6%	
Oscar	OSCR	Internet / Web scraping	14,173,696,374	15.7%	
Opus	OPUS	Internet / Web scraping	5,705,375,172	6.3%	
Blogset	BGST	Internet / User-Generated Content	1,525,281,151	1.7%	
Wikipedia PT Dump	WKPT	Internet / Informational	307,431,718	0.34%	
XLent	XLNT	Research / NER	207,888,796	0.23%	
Acórdãos STF	ASTF	Documents / Legal	76,388,660	0.08%	
Brazilian Legal Proceedings	BRLP	Documents / Legal	27,026,247	0.03%	
Gutenberg Project PT	GPPT	Arts / Literature	16,928,363	0.02%	
Wikibooks	WKBK	Arts / Literature	7,190,289	0.008%	
Brazilian Portuguese Literature	BRLT	Arts / Literature	3,370,103	0.004%	
How2	HOW2	Internet / Informational	3,018,834	0.003%	
Fernando Pessoa	FEPE	Arts / Literature	808,530	0.001%	
CorpusTCC	CTCC	Documents / Academic	52,223	0.00006%	
OpiSums	OPSU	Internet / User-Generated Content	6,328	0.00001%	
Total	-	-	**90,292,834,398.223**	**100%**	-

Table 5: Summary of the complete list of corpora contained in Jabuticaba dataset.

B CulturaX, ROOTS, and CC-100 Dataset Pipeline

B.1 CulturaX

CulturaX [25] pipeline involves two major steps of cleaning and deduplication to produce an enormous and high-quality dataset for multilingual LLMs.

1. Data Cleaning
   (a) Language Identification
   (b) URL-based Filtering
   (c) Metric-based Cleaning
   (d) Threshold Selection
   (e) Document Refinement
2. Data Deduplication
   (a) MinHash Deduplication
   (b) URL-based Deduplication

B.1.1 Data cleaning

Given the combination of the mC4 and OSCAR datasets, CulturaX performs a comprehensive data cleaning procedure to remove noisy and bad content from the data, including language identification, ULR-based filtering, metric-based cleaning, and document refinement.

Language Identification

A particular issue concerns the use of two different language identification tools, i.e., cld3 and FastText, for mC4 and OSCAR (respectively). It has been shown in previous studies that cld3 is significantly worse than FastText, causing substantially more language detection errors for mC4. FastText has demonstrated state-of-the-art performance over benchmark datasets.[21]

URL-based Filtering

Eliminates pages from known toxic and harmful sources to reduce relevant risks in the data. In particular, it used the latest UT1 blacklist of URLs and domains provided by the University of Toulouse. This list involves sites from different topics, including pornography, grumbling, and hacking, which should be discarded for LLM training.

Metric-based Cleaning

To enhance the dataset's quality, motivated by the data processing pipeline from BigScience's ROOTS corpus for BLOOM, it was used the distributions for various dataset metrics to identify and filter outlying documents. Each metric provides a singular value for every document within the dataset, quantifying specific attributes such as number of words, character repetition ratio, word repetition ratio, special character ratio, stop word ratio, flagged word ratio, language identification confidence, perplexity score, document length (number of characters), number of lines, short line length ratio, short line ratio for each document.

Threshold Selection

In the BigScience ROOTS project, the selection process is carried out by native speakers of 13 languages. The resulting thresholds are employed for the rest of their 46 languages. However, this process cannot be easily extended to different languages due to the requirement of experienced native speakers, which incurs significant costs. Furthermore, the limited sample sizes hinder the representativeness of the chosen thresholds for the full datasets. In CulturaX, it was observed that some selected thresholds for certain languages within BigScience ROOTS almost fall outside the value ranges for the entire dataset, leading to the deactivation of the corresponding metrics. To address these issues, a variant of the Interquartile Range (IQR) method was used to select appropriate thresholds for the filtering metrics for our dataset.

Document Refinement

The previous cleaning steps are done at the dataset level, aiming to remove low-quality documents from the dataset. In this step, it was cleaned the retained documents to improve the quality. It is important to note that the prior metric-based filtering step plays a vital role in eliminating highly noisy documents, which, in turn, streamlines the process of developing effective document-cleaning rules during this step.

Data Deduplication

Despite thorough data cleaning, the remaining dataset might still contain a substantial amount of repeated data due to various reasons, including information being reposted on the web, multiple references to the same articles, boilerplate content, and plagiarism. The duplicated data can thus cause memorization and significantly hinder generalization for LLMs. Data deduplication is thus considered a crucial step to guarantee the highest quality of data for training LLMs. To this end, it was performed a comprehensive deduplication procedure for the CulturaX dataset, utilizing MinHash and URLs. This deduplication process is carried out independently for each language. Furthermore, it restricted deduplication to languages that retain over 100K documents following the data cleaning procedures, aiming to promote smaller languages within the dataset.

MinHash Deduplication and URL-based Deduplication

For each language's dataset, the MinHashLSH method was applied to filter similar documents in the dataset. Finally, all documents that share identical URLs with other documents in the dataset were eliminated.

B.2 ROOTS

ROOTS [27] include the BigScience Catalogue and the Masader repository to start obtaining text from identified sources, which contain both existing NLP datasets and collections of documents of various compositions. To that end, was established a two-phase approach: first, collect as many data sources as possible in an easily accessible location; second, map all of them to a common format to ease further processing. In the first phase, was organized an open hackathon to start gathering identified sources on the HuggingFace Datasets hub. In the second phase, the collected datasets were further processed via (1) Language segmentation, whereby data sources were split using metadata for each covered language to obtain monolingual datasets, and the use of (2) a Uniform interface whereby a document

[21] https://modelpredict.com/language-identification-survey

consists of two fields: "text" for the actual text content, and "meta" with a JSON representation of metadata for a given document, containing sufficient information to trace documents back to their sources.

In the ROOTS dataset was decided to design the pipeline based on “pseudo-crawling”: that is, rather than crawling the websites ourselves, was retrieved pages corresponding to the target domain names from 18 snapshots archived by Common Crawl in 2020 and 2021 in Web ARChive (WARC) format. After gathering and processing language data via the pipeline, was manually inspected, deduplicated, and made a further selection of the sources. First, the dataset overlaps by looking through the sources. Dataset deduplication removed a high incidence of documents that were not fully in natural language, as well as very small datasets in the higher-resourced languages.

Pseudo-crawled sources were further processed to remove menus and pages that had a high incidence of character n-gram repetition, low language identification confidence, or a low proportion of closed-class words. Was removed entire domains whose size was less than 2MB after this step, yielding 147 pseudo-crawl-based datasets, and a total of 517 datasets. To improve the quality of that text, was removed noisy data from the dataset applying a processing pipeline for each dataset consisting of a sequence of functions. Document-scoped functions are operations that modify a document independently of other documents, and dataset-scoped functions are operations that take into account the whole dataset. Orthogonal to this scope, functions were also separated into cleaning and filtering functions.

The ROOTS dataset to process the OSCAR dataset first defines quality indicators for web content. These can then be used to filter out specific pages by defining cutoff thresholds. Was filtered out documents with character repetition or word repetition, special characters, closed class words, flagged words, perplexity, and number of words. Deduplication removes near-duplicate documents that were initially used by SimHash. Personally identifiable information is used as a rule-based approach leveraging regular expressions.

B.3 CC-100

CC-100 dataset [26] fetch, deduplicate, and filter the Common Crawl data. The focus is on preprocessing the text format of the common crawl snapshots. The pre-processing pipeline consists of the following steps

1. Preprocessing
2. Deduplication
3. Language Identification
4. LM filtering

Preprocessing

Each snapshot contains between 20 and 30 TB of uncompressed plain text, corresponding to approximately 3 billion web pages. Each one was downloaded and processed each snapshot independently. For each snapshot, regroup WET files into shards of 5 GB each. This makes up for 1600 shards for the February 2019 crawl. These shards are saved into a JSON file where one entry corresponds to one web page.

Deduplication

This step consists of removing duplicated paragraphs across the different web pages in a snapshot, as they represent 70% of the text. Each paragraph was normalized by lower-casing all characters, replacing numbers with a placeholder, and removing all Unicode punctuation and accent marks. Then, the deduplication is done in two independent steps. First, for every shard, a hash code was computed for each paragraph and saved into a binary file. The first 64 bits of SHA-1 digits were normalized paragraphs as the key. Then, was deduplicated every shard by comparing it with either a subset or all of the binary files.

Language identification The second step of the pipeline consists of splitting data per language. We use the language classifier from fastText.

LM filtering At this step, there are still documents with low-quality content. A way to filter out these samples is to compute a score of similarity of a web page with a targeted domain such as Wikipedia. CC-100 dataset proposes to use the perplexity of a language model trained on the targeted domain as the quality score.

References

[1] Jiya Manchanda, Laura Boettcher, Matheus Westphalen, and Jasser Jasser. The open source advantage in large language models (llms), 2025.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[4] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874, 2024.

[5] Toufique Ahmed, Christian Bird, Premkumar Devanbu, and Saikat Chakraborty. Studying llm performance on closed- and open-source data, 2024.

[6] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

[7] together.ai. Redpajama-data-v2: An open dataset with 30 trillion tokens for training large language models, 2023.

[8] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[10] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023.

[11] Douglas Trajano, Rafael H. Bordini, and Renata Vieira. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289, Dec 2024.

[12] Cássia C. S. Rosa, Fábio V. Martinez, and Renato Ishii. Natural language processing techniques for hate speech evaluation for brazilian portuguese. In Osvaldo Gervasi, Beniamino Murgante, Ana Maria A. C. Rocha, Chiara Garau, Francesco Scorza, Yeliz Karaca, and Carmelo M. Torre, editors, *Computational Science and Its Applications – ICCSA 2023 Workshops*, pages 104–117, Cham, 2023. Springer Nature Switzerland.

[13] Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, Felipe Ribas Serras, Guilherme Lamartine de Mello, Aline Silva Costa, Mayara Feliciano Palma, Renata Morais Mesquita, Raquel de Paula Guets, Mariana Marques da Silva, Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti, and Vanessa Martins do Monte. Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information, 2023.

[14] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France, June 2022. European Language Resources Association.

[15] Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In Vládia Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina Scarton, Fernando Batista, Diego Silva, Catarina Magro, and

Hugo Pinto, editors, *Computational Processing of the Portuguese Language*, pages 3–14, Cham, 2022. Springer International Publishing.

[16] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC corpus: A new open resource for Brazilian Portuguese. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[17] João A. Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis, 2020.

[18] Matheus Sanches, Jader de Sá, Henrique Foerste, Rafael Souza, Julio Dos Reis, and Leandro Villas. Textual datasets for portuguese-brazilian language models. In *Anais do IV Dataset Showcase Workshop*, pages 1–12, Porto Alegre, RS, Brasil, 2022. SBC.

[19] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. Tucano: Advancing neural text generation for portuguese, 2024.

[20] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[22] Timothy Nguyen. Understanding transformers via n-gram statistics, 2024.

[23] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

[24] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey, 2024.

[25] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023.

[26] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[27] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.