Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5789

Large Languages Models in Brazilian Portuguese: A Chronological Survey

William Alberto Cruz-Castañeda [**Amadeus AI** | ***william@amadeus-ai.com***]
Marcellus Amadeus [**Amadeus AI** | ***marcellus@amadeus-ai.com***]

Av. Paulista, 2006 - CJ. 1110, São Paulo, SP, 01310-200, Brazil.

Received: 30 March 2025 • **Accepted:** 30 July 2025 • **Published:** 24 October 2025

Abstract The era of Large Language Models (LLMs) started with OpenAI's GPT-3 model, and the popularity of LLMs has increased exponentially after the introduction of models like ChatGPT and GPT4 that demonstrated remarkable capabilities in natural language processing tasks. LLMs are a special class of pre-trained language models (PLMs) obtained by scaling model size, pretraining corpus, and use of computational power. Large PLMs can be valuable assets, especially for languages such as Portuguese to capture the cultural and knowledge richness inherent in the language. In this sense, this survey encompasses, based on the existing scientific literature, an overview of the research development with LLMs on Brazilian Portuguese (PT-BR-LLMs). The objective is to bring a self-contained, comprehensive overview of PT-BR-LLMs advancements, architectures, and resources. This survey is intended not only to provide a systematic survey but also a quick, comprehensive reference for the research community and practitioners to draw insights from extensive informative summaries of the existing scientific works to advance and progress in the PT-BR-LLMs research field. Considering the emergence of new literature on PT-BR-LLMs, future updates will be made and regularly maintained in the project repository https://github.com/Amadeus-AI-Official/pt-br-llms

Keywords: Brazilian Portuguese, Large Language Models, Architectures, Configuration

1 Introduction

Large Language Models (LLMs) have been demonstrating emergent abilities extending beyond their core functions, showing proficiency in tasks like commonsense reasoning, code generation, and arithmetic Wei *et al.* [2022]. These capabilities have been attracting interest in both academic and industrial fields due to their ability to solve diverse tasks, contrasting with previous models limited to solving specific tasks Liu *et al.* [2023] Dong *et al.* [2024] Huang and Chang [2023].

For example, OpenAI's GPT-4 can be used not only for Natural Language Processing (NLP) but also as a general task solver, for instance, can follow human instructions for complex new tasks performing multi-step reasoning when needed Bubeck *et al.* [2023]. To track and explain this progress, as well as to spread emergent abilities or techniques of LLMs, surveys such as Naveed *et al.* [2024] and Zhao *et al.* [2023] provide informative summaries by both academia and industry with useful resources on pre-training, adaptation tuning, utilization, and capacity assessment.

However, the vast majority of foundational LLM research currently relies on the English language, and North American sociocultural preferences are dominant in its design, outcomes, and behavior Naous *et al.* [2024]. In this sense, Min *et al.* [2023] presents advances using pre-training and fine-tuning, prompting, or text generation approaches in Pretrained Foundation Language Models (PFLMs) to solve NLP tasks. Practical guidelines for understanding PFLMs were proposed by Qiu *et al.* [2020] and Li *et al.* [2024a] to pretrain Transformer models on large-scale corpora and multiple adapted tasks through fine-tuning, few-shot learning, or even zero-shot learning for small-scale tasks Zhuang *et al.* [2021], Zhou *et al.* [2023].

Also Chowdhery *et al.* [2024], Touvron *et al.* [2023a], Touvron *et al.* [2023b], Brown *et al.* [2020], Ouyang *et al.* [2024], OpenAI [2022], and Kalyan [2024] provide examples of PFLMs as PaLM, LLAMA, LLAMA2, GPT-3, InstrucGPT and ChatGPT fine-tuned from GPT-3.5. In this case, ChatGPT represents one of the most exciting LLM systems developed recently to showcase impressive language generation abilities and attract audience attention. Nevertheless, given their wide adoption, a natural question is, if can also be applied in other languages or whether further languege-specific technologies need to be developed. The work of Lai *et al.* [2023] fills this gap by evaluating Chat-GPT and similar LLMs, covering 37 languages with high, medium, low, and extremely low resources on seven multilingual NLP tasks. Their results show that ChatGPT exhibits worse performance for the considered NLP tasks. Furthermore, recent results presented by Qin *et al.* [2025] and Yue *et al.* [2025] show that the effectiveness of state-of-the-art (SOTA) supervised multilingual models in the same NLP tasks is lower in other languages due to the limited availability of linguistic resources.

Therefore, this indicates that there is still a knowledge gap for research in foundational LLMs that cover other languages, and those that do so generally do not perform as well in comparison to English Li *et al.* [2024b]. This knowledge gap in other languages affects the development of LLMs as a result of several factors. For example, the unknown use and outcomes of pre-training and post-training techniques, the dissemination of high-quality text-based datasets, the adaptation know-how of the architecture, the optimization process, the use of GPUs/TPUs, as well as the implemented distributed computing frameworks to train models with billions, and even trillions, of parameters Miranda *et al*. [2024], Marion *et al*. [2023], Matarazzo and Torlone [2025]. Together, these factors would enable LLMs to capture nuanced linguistic patterns, cultural context, and domain-specific knowledge, enhancing their ability to generate coherent, contextually appropriate, and highly versatile outputs.

In the case of Brazilian Portuguese, a low-resource language, the current explosion in LLMs poses challenges for foundational research. Despite these challenges, variants of LLMs in Brazilian Portuguese (PT-BR-LLMs) are beginning to emerge, and several NLP resources are available in the scientific literature, but they are scattered. On the one hand, it often leads to situations where it is difficult to know where to start when trying to understand and learn about the PT-BR-LLMs' progress. On the other hand, there is a lack of systematic organization regarding the differences, types, domains, etc., among various existing PT-BR-LLMs. In this context, to shorten the learning curve, promote foundational research for PT-BR-LLMs and technological innovation, and enhance academic and industrial awareness, we conducted a survey to summarize the progress on PT-BR-LLMs.

Thus, the contribution of this paper is to provide researchers and practitioners with a comprehensive and concise overview of the direction of the PT-BR-LLMs field, facilitating a better understanding of the distribution and their role, thereby advancing collective knowledge and their evolution. The organization of this article is as follows. Section II presents the methodology carried out to identify studies on LLMs and Brazilian Portuguese. Section III presents the existing LLM variants in Brazilian Portuguese, in chronological order, describing their architecture, features, datasets, and availability. Section IV presents some configuration and parameters findings derived from each PT-BR-LLM that play a crucial role in the functioning of those models. The discussion and conclusion are presented in section V.

2 Methodology

This survey was carried out to identify studies on LLMs developed for Brazilian Portuguese. Figure 1 depicts the flow of the search strategy followed in this article. IEEE Xplore Digital Library, ScienceDirect, PubMed, SciELO, ACM Digital Library, Springer Link, arXiv databases as well as Google Scholar were extensively searched to include articles published between 2020 and 2025. The list of keywords was created around the combinations of generic terms. For generic terms, it was used the following Boolean search string: (1) "Brazilian Portuguese" AND "Large Language Models".

Exploring titles and abstracts, the search strategy returned 108 articles. Descriptions in articles including only datasets, a specific NLP task, speech recognition, and applications of commercial LLMs in Brazilian Portuguese were omitted, and the search was refined, returning 76 articles. Those seventy-six articles were retrieved for comprehensive analysis, in other words, they were read completely. In this comprehensive analysis, it was found that twenty-nine articles were related to evaluation, corpus, embedding generation, semantic knowledge, and data augmentation quality. Because this article aims to comprehensive and concise overview of the direction of the PT-BR-LLMs research, they are not included. Thus, forty-seven articles were analyzed and divided into five areas, each one per year.

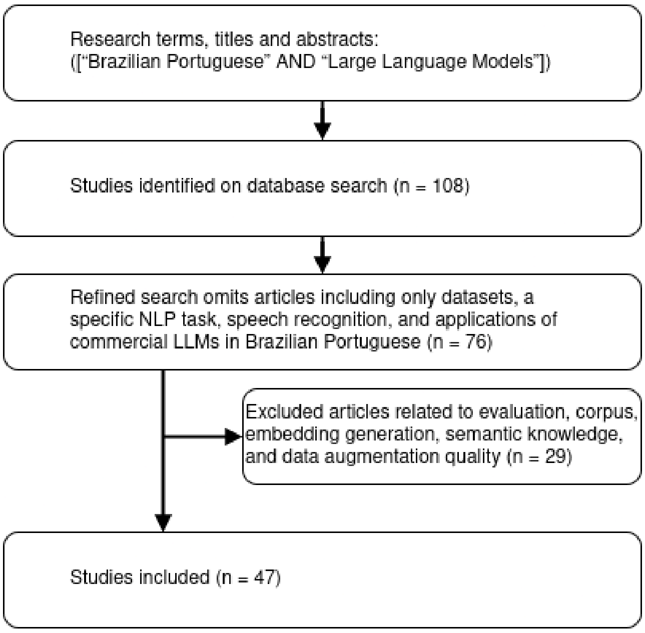
Research terms, titles and abstracts:
(["Brazilian Portuguese" AND "Large Language Models"])
Studies identified on database search (n = 108)
Refined search omits articles including only datasets, a specific NLP task, speech recognition, and applications of commercial LLMs in Brazilian Portuguese (n = 76)
Excluded articles related to evaluation, corpus, embedding generation, semantic knowledge, and data augmentation quality (n = 29)
Studies included (n = 47)

Figure 1. Methodological search strategy followed in this survey to include research studies.

3 Variants of LLMs in Brazilian Portuguese

This section provides a synopsis of pre-trained SOTA models in Brazilian Portuguese, classified by years starting from 2020. Summaries include the base model used, pre-training and/or fine-tuning strategy, and dataset used. Figure 2 depicts these years with the number of published and unpublished models, and Figure 3 shows the evolution of Brazilian Portuguese models over time from 2020 onwards.

3.1 PT-BR-LLMs produced in 2020

Combining the transfer capabilities of the BERT model with the structured conditional random field (CRF) predictions, Souza *et al*. [2020b] compares feature-based and fine-tuning strategies in a Portuguese BERT model using a **BERT-CRF** architecture with the named entity recognition (NER) task. In the feature-based strategy, the weights of the BERT model are kept frozen and only the classifier model and the CRF layer are trained. For the fine-tuning strategy, a linear classifier is used, and all weights are jointly updated during training. For pre-training, the brWaC corpus was used. First HAREM is used as a training set and MiniHAREM as a testing set.

Carmo *et al*. [2020] refers to **PTT5**[1] as the T5 model pre-trained using three Brazilian Portuguese datasets: BrWac for

[1] https://github.com/unicamp-dl/PTT5

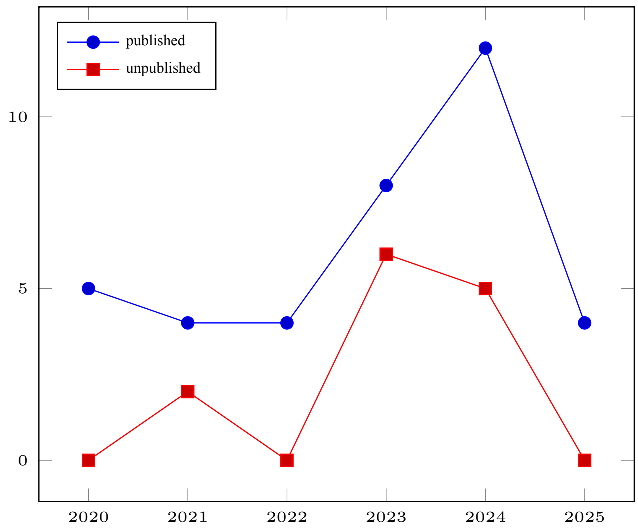

Figure 2. Number of published and unpublished models starting from 2020 to 2025.

pretraining, and ASSIN2 and HAREM for fine-tuning and evaluation. PTT5 uses the same control tokens and vocabulary size as the original T5 to start pre-training from the original checkpoint. PTT5 improves the original T5 model in Portuguese language tasks of sentence prediction and NER.

The Bonifacio *et al.* [2020]'s research considers **ELMo** and **BERT** language models pre-trained on three general domain corpora: the 104 largest Wikipedias, the Portuguese Wikipedia, and the brWaC corpus. It also considers the fine-tuning of these models on a legal domain on the Acórdãos-TCU corpus[2]. For evaluation, the NER corpora used were HAREM, LeNER-Br, DrugSeizures-Br. The AllenNLP library was used to train and evaluate the model.

Souza *et al.* [2020a] trained BERT-base and BERT-large models using the brWaC dataset and evaluated in ASSIN2, First HAREM, and MiniHARE datasets. These models are called **BERTimbau**[3]. The BERTimbau-base weights are initialized with the multilingual BERT-base checkpoint. BERTimbau-large weights are initialized with the checkpoint of English BERT-large. Early stopping is implemented.

Schneider *et al.* [2020] developed **BioBERTpt**[4], which is a collection of three BERT-based models fine-tuned in Brazilian Portuguese, initialized with multilingual BERT weights, and using clinical and biomedical corpora. The clinical corpus disidentified clinical notes from Brazilian hospitals with multi-specialty information. In total, the clinical notes contain 3.8 million sentences with 27.7 million words. The biomedical corpus contains titles and abstracts of Portuguese scientific articles on biology and health published in the PubMed and Scielo databases, resulting in 16.4 million words. The preprocessing steps involve splitting the corpus into sentences and tokenizing them with the standard BERT text tokenizer.

Thus, advancements in PT-BR-LLMs during 2020 show BERT-based models consistently achieving state-of-the-art performance across various tasks like NER and sentence entailment, often outperforming multilingual equivalents. Intra domain finetuning proves crucial for domain-specific tasks like legal NER, though language style within a domain can impact effectiveness. While Portuguese-specific vocabularies and full model pretraining enhance T5 model performance, BERTimbau Large currently remains the top performer. The development of models like BioBERTpt further demonstrates the significant benefits of domain transfer learning for clinical NLP in Portuguese.

3.2 PT-BR-LLMs produced in 2021

GPT2-Bio-Pt[5] was developed by Schneider *et al.* [2021] based on a Generative Pre-trained Transformer 2 (GPT-2) language model for Portuguese to support clinical and biomedical NLP tasks. It was fine-tuned using transfer learning in GPorTuguese-2 (Guillou [2020]) with biomedical corpora containing titles and abstracts of Portuguese scientific articles published in Pubmed and Scielo databases. The preprocessing step splits the corpora into sentences and tokenizes them. All sentences were truncated at 1,024 tokens due to the length supported by GPT-2-base.

José and Cozman [2021] developed **mRAT-SQL+GAP**[6], which is a fine-tuned Portuguese-to-SQL query translator of BART multilingual model. The choice was mBART-50 because it covers Portuguese and English to allow multilingual processing. To handle and support Portuguese lemmatization, the original code on the Stanford CoreNLP lemmatization tool was replaced by Simplemma[7].

In the financial domain, Finardi *et al.* [2021] developed **BERTaú**, a model trained from scratch in an uncased BERT-base model with data from the Itaú virtual assistant chatbot (AVI). BERTaú's smaller and lighter models achieve SOTA performance on three NLP tasks. The training setup follows the BERT paper guidelines. The vocabulary has 14,5 GB of conversation data (approximately 22,500,000 words) in BERT-uncased format. Also, the model was quantized in $int8$.

BertBR is a BERT model trained in Portuguese, developed by Ciurlino [2021], with a pre-training and fine-tuning processes using legal texts. The BERTimbau model was imported from Huggingface and trained. All layers of the model were trained using the methods described in BERT guidelines.

Advancements in PT-BR-LLMs during 2021 highlight the effectiveness of domain-specific model training for specialized tasks. New models like Portuguese GPT-2 for clinical NLP and BertBR for legal NER demonstrate state-of-the-art performance by leveraging in-domain data, even outperforming general-purpose models. Specialized BERT models like BERTaú significantly improve digital customer service tasks, proving more efficient than multilingual alternatives. Additionally, for Portuguese-to-SQL translation, a multilingual approach that combines English and Portuguese data with models like mBART-50 is found to be essential for optimal results.

[2] part of the Federal Court of Accounts in Brazil
[3] https://github.com/neuralmind-ai/portuguese-bert
[4] https://github.com/HAILab-PUCPR/BioBERTpt
[5] https://github.com/HAILab-PUCPR/gpt2-bio-pt
[6] https://github.com/C4AI/gap-text2sql
[7] https://github.com/adbar/simplemma

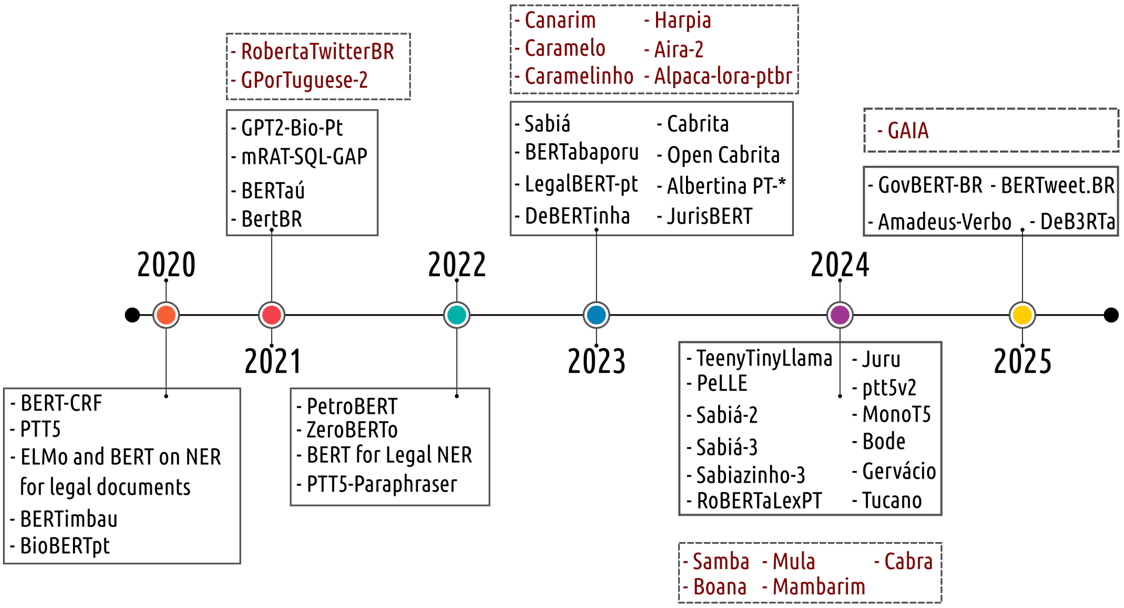

Figure 3. Chronological evolution of PT-BR-LLM releases. Square cards represent pre-trained models with publication, while dotted square cards correspond to unpublished models.

3.3 PT-BR-LLMs produced in 2022

PetroBERT is the result of the research of Rodrigues *et al.* [2022], which adapted the BERT-based model to the oil and gas exploration domain. PetroBERT was pre-trained using the Petrolês corpus and a private corpus of daily drilling report (DDR-Corpus) on top of multilingual BERT and BERTimbau-based models. The fine-tuning step is performed on GeoCorpus (an open dataset of geoscience scientific articles, and reports) and DDR-Corpus for NER and text classification tasks. Four models were derived. $PetroBERT_{multi-pet}$ with BERT Multilingual Cased-based model adapted with Petrolês. $PetroBERT_{multi-ddr}$ with BERT Multilingual Cased-based model adapted with DDR-corpus. $PetroBERT_{pt-pet}$ with BERTimbau-based model adapted with Petrolês. $PetroBERT_{pt-ddr}$ with BERTimbau-based model adapted with DDR-corpus.

Developed by Alcoforado *et al.* [2022], **ZeroBERTo** is a BERT-based model for topic modeling and zero-shot multi-class text classification tasks. ZeroBERTo instead of processing the entire document in the model, learns a compressed data representation in an unsupervised manner and processes only that representation. Thus, it is possible to obtain better performance with large inputs and less total time than the standard model, even considering the training time added by the unsupervised step. To learn this representation, ZeroBERTo uses a statistical topic modeling model to examine documents and discover hidden topics and semantic structures.

The research of Zanuz and Rigo [2022] presents a fine-tuned **BERT models for Legal NER**. BERTimbau-base and large were fine-tuning on LeNER-Br[8] dataset with a maximum sentence length of 512 tokens and a vocabulary size of 29794 words. The main difference between the base and large versions lies in the training and evaluation batch sizes.

PTT5-Paraphraser[9] is a paraphraser for Portuguese based on PTT5 model and fine-tuned with TaPaCo[10] dataset by Pellicer *et al.* [2022]. The fine-tuning process involved only TaPaCo's Portuguese paraphrases. The standard maximum likelihood loss function and the AdaFactor optimizer were used, as well as a prefix paraphrase was created for training.

Advancements in PT-BR-LLMs in 2022 showcase several specialized models. PetroBERT demonstrates promising results in the petroleum domain, with Portuguese-only models outperforming multilingual ones despite resource constraints. For zero-shot text classification in low-resource settings, ZeroBERTo significantly improves performance and speed, though it can overfit new data and requires more memory. In the legal domain, the first BERT models fine-tuned exclusively for Brazilian Portuguese Legal NER achieved state-of-the-art results on the LeNER-Br dataset. Additionally, PTT5-Paraphraser offers a valuable tool for data augmentation in low-resource scenarios, balancing meaning fidelity with lexical diversity.

3.4 PT-BR-LLMs produced in 2023

Sabiá[11], developed by Pires *et al.* [2023], is a set of models pre-trained on GPT-J (Wang and Komatsuzaki [2021]) and LLaMA (Touvron *et al.* [2023a]) with data coming from the Portuguese subset of the ClueWeb 2022[12] dataset. GPT-J and LLaMA tokenizers are used for tokenization. Three models were trained using LLaMA 7B, LLaMA 65B, and GPT-J. The pre-trained LLaMA models are called Sabiá, while the one derived from GPT-J is referred to as Sabiá-J. The Sabiá models were trained with LLaMA weights using the t5x and seqio frameworks. Sabiá-J was trained using the mesh-transformer-jax framework.

BERTabaporu[13] is a BERT language model pre-trained

[8] https://github.com/peluz/lener-br

[9] https://github.com/lucasfaop/ptt5_paraphraser

[10] https://doi.org/10.5281/zenodo.3707949

[11] https://huggingface.co/maritaca-ai/sabia-7b

[12] https://www.lemurproject.org/clueweb22/

[13] https://huggingface.co/pablocosta/bertabaporu-base-uncased

by Costa *et al*. [2023] on a dataset of 238 million tweets in Brazilian Portuguese. A monolingual BERT model was pre-trained from scratch using BERT-base and BERT-large architectures. In both cases, the vocabulary is initialized with 64K tokens.

LegalBert-pt is a specialized model pre-trained and fine-tuned on a large and diverse corpus from the Brazilian legal domain by Silveira *et al*. [2023]. Two versions of the model were created: one as a complement to the BERTimbau model (*LegalBert-pt FP*)[14], and the other from scratch (*LegalBert-pt SC*)[15]. The data were obtained from the Codex system of the Brazilian National Council of Justice (CNJ), which maintains the largest and most diverse set of legal texts. Both models were pre-trained as case-sensitive, using the masked language modeling (MLM) task. The *LegalBert-pt SC* model was pre-trained for 7.5 million steps and the *LegalBert-pt FP* model was initialized with the weights from the pre-trained BERTimbau-Base checkpoint and performed additional pre-training up to 2.4 million steps.

Adapted by Campiotti *et al*. [2023], **DeBERTinha**[16] is a model for Brazilian Portuguese from the DebertaV3 XSmall model. Carolina and BrWac served as baseline datasets for pre-training. The model was initialized using random embeddings, and for pre-training, it was used a combination of MLM and Replaced Token Detection (RTD) losses with a hyperparameter λ. DeBERTinha, demonstrates effectiveness on downstream tasks like NER, sentiment analysis, and determining sentence relatedness, outperforming BERTimbau-Large in two tasks.

Cabrita is a methodology created by Larcher *et al*. [2023] that addresses the performance and efficient tokenization problem with continuous pre-training using Portuguese corpus on OpenLLaMA 3B (Geng and Liu [2023]) (Touvron *et al*. [2023a]). The result is **OpenCabrita**[17], models that undergo additional pre-training, the adaptation of the tokenizer is inspired by the work of Cui *et al*. [2024], and knowledge of the original language is preserved. OpenCabrita involves three steps to result in a bilingual tokenizer: training a new Portuguese tokenizer, merging the original tokenizer with the Portuguese tokenizer, and resizing the model. The Portuguese mC4 subset was used as the pre-training corpus. The Pires *et al*. [2023] recipe was followed to apply MassiveText-based quality filters and ensure model training. The continuous pre-training step was performed on an additional 7 billion tokens using the EasyLM framework, OpenLLaMA hyperparameter, and weights.

Albertina PT-*, developed by Rodrigues *et al*. [2023], has two variants, European Portuguese from Portugal (PT-PT) and American Portuguese from Brazil (PT-BR)[18]. Albertina PT-* is based on DeBERTa, and its pre-training was done on Portuguese datasets. In the training of Albertina PT-BR, the brWaC dataset was tokenized with the original DeBERTa tokenizer and dynamic padding. The Albertina PT-BR No-brWaC model variant was trained using a curated selection of documents from the OSCAR dataset.

Developed by Viegas *et al*. [2023], **JurisBERT**[19] [20] is a model trained from scratch using the BERT model. Is specific for semantic textual similarity of the ementas of acordaos (Brazilian legal field domain-specific texts) with an sBERT network. For the study, it was necessary to construct two corpora (for training and fine-tuning) with data available on the court websites. Experiments showed JurisBERT is better than other models such as multilingual BERT and BERTimbau, five times reduced training time, and using accessible hardware.

Thus, strides in PT-BR-LLMs in 2023 emphasize the power of domain-specific and specialized models to achieve state-of-the-art results. Models like Sabiá (for Next Phrase Prediction), BERTabaporu (for Twitter tasks), and LegalBert-pt (for legal NLP) consistently outperform general-purpose or multilingual baselines by leveraging specialized pretraining and data. Even smaller, lightweight models like DeBERTinha are demonstrating competitive to state-of-the-art performance, highlighting efficient strategies for resource-constrained scenarios. Innovations in tokenization, seen with openCabrita3b, are improving inference times. Furthermore, new foundation models like Albertina PT- are setting new benchmarks for both European and Brazilian Portuguese, and models like JurisBERT are showing the benefits of domain-specific pre-training in legal contexts, often with reduced computational costs.

3.5 PT-BR-LLMs produced in 2024

The **TeenyTinyLlama**[21] pair is a two-compact models created by Corrêa *et al*. [2024a], which provides a simple and extensible implementation for pre-training and fine-tuning small-scale LLM. To estimate the model and dataset size, Hoffmann *et al*. [2024]'s scaling law was chosen. The GPorTuguese-2 tokenizer and embedding layer were used as replacements for the originals. Llama 2 was the adopted architecture, and the dataset used was Pt-Corpus-Instruct, which consists of the concatenation of two parts: a) open-source Brazilian Portuguese datasets (Wikipedia, CulturaX, OSCAR, Common Crawl, and ROOTS), and b) translated versions of native English datasets (Instruct-PTBR, Gpt4all-J, Bactrian-X, Dolly 15K, and CosmosQA). As a filtering step, Rae *et al*. [2022] was followed, in addition to using a fine-tuned BERTimbau to exclude samples classified above a pre-defined toxicity threshold.

de Mello *et al*. [2024] developed **PeLLE**, a family of models (*pPeLLE*, *xPeLLE*, *mPeLLE*) based on the RoBERTa architecture (Liu *et al*. [2019], XML-RConneau *et al*. [2020], mBERT Devlin *et al*. [2019], respectively) and trained on a curated version of the Carolina corpus. Those models were trained from scratch using the Masked Language Modeling (MLM) objective. NFKD Unicode normalization was used as the only text preprocessing step.

Sabiá-2 is a family of models (Sabiá 65B, Sabiá-2 Small, and Sabiá-2 Medium) created by Almeida *et al*. [2024]. The architecture, hyperparameters, and other techniques used are

[14] https://huggingface.co/raquelsilveira/legalbertpt_fp

[15] https://huggingface.co/raquelsilveira/legalbertpt_sc

[16] https://huggingface.co/sagui-nlp/debertinha-ptbr-xsmall

[17] https://huggingface.co/22h/open-cabrita3b

[18] https://huggingface.co/PORTULAN/albertina-900m-portuguese-ptbr-encoder-brwac

[19] https://github.com/alfaneo-ai/brazilian-legal-text-dataset

[20] https://huggingface.co/alfaneo

[21] https://github.com/Nkluge-correa/TeenyTinyLlama

not specified, but following the predecessor model Sabiá, based on Llama-1 architecture, it is assumed that Sabiá-2 could be based on Llama-2 architecture, released at the time.

Bode[22] is a fine-tuned LLaMA 2 model with two versions: 7B and 13B. The models were developed by Garcia *et al.* [2024b] to adapt and improve the prompt-based tasks for instruction-following responses in Portuguese. In the fine-tuning procedure, the LLaMA 2 models were trained on the same Portuguese dataset assembled by Larcher *et al.* [2023], used the LowRank Adaptation (LoRA) method with the following hyperparameter: LoRA alpha = 32, and dropout rate = 0.05. Furthermore, zero-shot and in-context learning approaches were adopted.

Juru is a specialized version of the Sabiá-2 Small model developed by Junior *et al.* [2024] with 1.9 billion unique tokens from Brazilian legal sources. Juru demonstrates the benefits of domain specialization with a reduced amount of pre-training data. The pre-training data involves a web scraping of academic articles prioritizing data with educational value and a subset of Brazilian federal laws. The documents were curated using the filter proposed by Rae *et al.* [2022]. With the Sabiá-2 Small tokenizer, each document was divided into sequences of 4,096 tokens. Seqio and t5x frameworks were used. The pre-training hyperparameters are based on Pires *et al.* [2023].

ptt5v2 and **MonoT5**[23] are continued pretraining models of Google-T5 developed by Piau *et al.* [2024] with up to 3 billion parameters. As pretraining data, the Portuguese segment of the mC4 dataset adopts the vocabulary from ptt5-v1. The experiments implement Google's original checkpoints for pretraining and finetuning using t5 and seqio frameworks. On the other hand, MonoPTT5 is an adaptation of the ptt5-v2 model for information retrieval tasks. Training data comes from the mMARCO dataset and a Portuguese-English bilingual dataset.

Created by Santos *et al.* [2024], **Gervácio** is based on the LLaMa 2 model with 7 billion parameters, for both Portuguese, European, and Brazilian[24] variants. Each variant implements a supervised fine-tuning and zero-out technique during the fine-tuning process. The translation of the datasets STS-B and WNLI, from GLUE, and BoolQ, CB, and MultiRC, from SuperGLUE, were used for training.

Developed by Abonizio *et al.* [2024], **Sabiá-3** and **Sabiazinho-3** are pre-trained and fine-tuned models that the main goal is to specialize in linguistic nuances, social norms, and regional variations unique to the country. The architecture is not specified and only a technical report explains the evaluations in several professional and academic benchmarks in tasks related to Brazilian Portuguese. The development consists of two phases, pre-training with specialized data following a self-supervised learning strategy, and post-training to follow instructions and align with human preferences. The quality of the pre-training data was improved using a mixture of heuristic and model-based methods to filter out low-quality data.

Developed by Corrêa *et al.* [2024b], **Tucano**[25] is a set of pre-trained models based on Llama architecture and trained on the GigaVerbo dataset. All models were trained using a causal language modeling objective and cross-entropy as its loss.

Garcia *et al.* [2024a] works on **RoBERTaLexPT**[26], which emphasize the importance of adapting pre-trained models, such as RoBERTa, from specialized corpora in the legal domain. Two corpora for pre-training were used: LegalPT and CrawlPT. Additionally, it was created the PortuLex benchmark, composed of a set of legal supervised tasks designed to evaluate the language models.

Researchs significantly advances in PT-BR-LLMs in 2024 with several new models. TeenyTinyLlama and PeLLe models offer strong performance in low-resource settings, often competing with larger multilingual models. The Sabiá-2 and Sabiá-3 families represent major leaps, matching or surpassing GPT-4 on many benchmarks, demonstrating the power of linguistic and cultural specialization. Bode shows promise for various Portuguese NLP applications, while Juru highlights the benefits of domain-specific LLMs for legal tasks, despite performance trade-offs. Additionally, PTT5-v2 achieves state-of-the-art results in several Portuguese datasets, emphasizing the superiority of monolingual models. Gervásio demonstrates the effectiveness of adapting LLMs to specific languages with small datasets, and RoBERTaLexPT sets new benchmarks in Portuguese legal NLP by leveraging high-quality pretraining data. The Tucano series further contributes by providing open-source, highly reproducible Portuguese language models, with plans to expand datasets and model scales.

3.6 PT-BR-LLMs produced in 2025

Introduced by Silva *et al.* [2025], **GovBERT-BR** [27] is a pre-trained language model covering legal and administrative domains, addressing the challenges of accurately interpreting the legal and bureaucratic terminology prevalent in governmental documents.

BERTweet.BR[28], developed by Carneiro *et al.* [2025], is a pre-trained model for the tweets domain with the same architecture of BERTweet$_{base}$. It was trained following the RoBERTa pre-training procedure on a corpus of approximately 9 GB containing 100 M Portuguese tweets.

Developed by Pires *et al.* [2025], **DeB3RTa**[29] is a DeBERTa-v2-based model for the financial domain. The corpus incorporates four datasets (OFFCOMBR-3, FAKE.BR, CAROSIA, BBRC) and external sources such as relevant facts, patents, Scielo, Wikipedia, and news.

The technical report of Cruz-Castañeda and Amadeus [2025] presents **Amadeus-Verbo**[30], a post-trained family of LLMs for Brazilian Portuguese that includes base-tuned, merged, and instruction-tuned models in parameters sizes of

[22] https://huggingface.co/collections/recogna-nlp/bode-llm-em-portugues-65b97aa411162bf34f8da221

[23] https://huggingface.co/unicamp-dl

[24] https://huggingface.co/PORTULAN/gervasio-7b-portuguese-ptbr-decoder

[25] https://huggingface.co/TucanoBR

[26] https://github.com/eduagarcia/roberta-legal-portuguese

[27] https://huggingface.co/dccmpmgfinalisticas/GovBERT-BR

[28] https://huggingface.co/melll-uff/bertweetbr

[29] https://huggingface.co/higopires/DeB3RTa-base

[30] https://huggingface.co/collections/amadeusai/amadeus-verbo-qwen25-pt-br-powered-by-aws-67cf2e7aae69ce2b3bcdcfda

0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B. The main objective is to show how easy it is to fine-tune foundation models and to democratize the open-source development of PT-BR-LLMs when data and resources are available. Amadeus-Verbo family is based on the Qwen2.5 open-weights model series and the fine-tuning constitutes a cornerstone of adapting pre-trained LLMs to specialized tasks, refining their capabilities through targeted parameter adjustments.

Advancements in PT-BR-LLMs in 2025 showcase several new specialized models. GovBERT-BR, is an initiative covering the for governmental context. BERTweet.BR is the first large-scale pre-trained model for Portuguese tweets, consistently outperforming competitors in sentiment classification, validating the effectiveness of domain-specific models. Similarly, DeB3RTa, a Portuguese financial domain-specific model, consistently outperformed general and multilingual models in financial tasks due to specialized pre-training. Finally, the Amadeus-Verbo series demonstrates comparable performance to original instructional language models on Brazilian Portuguese tasks, with future efforts focusing on data refinement and reasoning capabilities to advance Brazilian Portuguese LLMs.

3.7 PT-BR-LLMs without publication

The following models have repurposed LLMs for PT-BR and other Portuguese variants through full fine-tuning or LoRA/PEFT, mentioned in some references but not accompanied by a paper or technical report.

RobertaTwitterBR[31] is a RoBERTa model trained on 7M tweets approximately. **GPorTuguese-2**[32], developed by Guillou [2020], is based on the GPT-2-small model. It was trained on Portuguese Wikipedia using transfer learning and fine-tuning techniques in just over a day, on a 32GB NVIDIA V100 GPU and with approximately 1GB of training data. It was fine-tuned from the pre-trained GPT-2-small in English using the HuggingFace libraries integrated with the Fastai framework.

Canarim[33] is a byproduct of the fine-tuning of Llama 1/2, which extends the training process to 16 billion tokens of the Portuguese subset of Common Crawl. Canarim-7b is initialized with the weights of LLaMA2-7B. Canarim-7B-Instruct is initialized from Canarim-7B and trained on a variety of publicly available instruction datasets. Canarim-7b-vestibulaide is specifically designed to deal with questions, exercises, and answers from Brazilian university entrance exams. The models are also under a Llama 2 license.

Caramelo, Caramelinho, and Harpia[34] are models adapted from Falcon-7b, created with the PEFT library and fine-tuned on alpaca-data-pt-br, Canarim, and open assistant-guanaco using the method QLoRA.

Aira-2[35] is the second version of the instruction-tuned Aira series. Aira-2-portuguese-124M is an instruction-tuned model based on GPT-2. Aira-2-portuguese-560M and Aira-2-portuguese-1B7 are instruction-tuned models based on BLOOM. The models were trained with a dataset composed of prompts, and synthetically generated completions by prompting already-tuned models.

Alpaca-lora-ptbr-7b[36] is a model created and made available exclusively for research purposes. It involves using a low-rank (LoRa) adapter for LLaMA-7b fine-tuned on the Stanford Alpaca dataset translated into Brazilian Portuguese using the Helsinki-NLP/opus-mt-tc-big-en-pt model.

Cabra[37] are open-source models fine-tuned on the PortugueseDolly and Cabra 10k datasets to deepen understanding of the Portuguese language and Brazilian culture. All the models are available for demonstration and research purposes only. Commercial use is prohibited. Furthermore, the models require further training and may generate lies or untruths.

Samba[38] is a model based on TinyLlama-1.1B, developed by Zhang *et al.* [2024], which is a 1.1B parameters version of LLaMA-2.

Boana[39] is a model based on LLaMA-2-7B. The Boana project aims to provide LLM options in the Portuguese language while providing a less complex model so that users with less computational power can benefit from LLMs.

Mula is a series of Sparse Mixture of Experts (SMoE) language models, all trained natively in Brazilian Portuguese, and designed to help democratize LLMs for low-resource languages. Mula-4x160-v0.1 is the first pre-training experiment on an SMoE using the Pt-Corpus-Instruct dataset. The model consists of 4 experts per layer and activates 2 for each token. The SMoE architecture is pre-trained through causal language modeling. It contains 407,820,288 parameters (only 237,950,976 parameters activated during runtime) and a context length of 2048 tokens.

Mambarim-110M[40] is a model based on a state-space model (Mamba) architecture, not a transformer. It contains 119,930,880 parameters, a context length of 2048 tokens, and was trained using the Pt-Corpus Instruct dataset with 6.2B tokens.

GAIA[41] is an open model, built on Gemma 3, designed to improve the understanding of the Brazilian Portuguese language. It was developed through continuous pre-training on a vast dataset of Portuguese content. The training dynamically balanced data from various sources and processed approximately 13 billion tokens on NVIDIA DGX infrastructure with H100 GPUs, allowing to follow instructions without traditional fine-tuning.

As a brief analysis of this section, numerous LLMs have been repurposed and fine-tuned by the industry to generate Portuguese variants, utilizing methods like full fine-tuning or LoRA/PEFT. Notable examples include RobertaTwitterBR for social media, GPorTuguese-2 (GPT-2 based), and the Canarim series (Llama 1/2 based) with specialized versions for Brazilian university exams. Other adapted models like Caramelo (Falcon-7b based), Aira-2 (GPT-2/BLOOM based), and Alpaca-lora-ptbr-7b (LLaMA-7b based) leverage instruc-

[31] https://huggingface.co/verissimomanoel/RobertaTwitterBR

[32] https://huggingface.co/pierreguillou/gpt2-small-portuguese

[33] https://huggingface.co/collections/dominguesm/canarim-models-657efdf4f5eacd4bda0cb77b

[34] https://huggingface.co/Bruno

[35] https://huggingface.co/nicholasKluge

[36] https://huggingface.co/dominguesm/alpaca-lora-ptbr-7b

[37] https://huggingface.co/botbot-ai/Cabra

[38] https://huggingface.co/lrds-code/samba-1.1B

[39] https://huggingface.co/lrds-code/boana-7b-instruct

[40] https://huggingface.co/dominguesm/mambarim-110m

[41] https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it

tion tuning and translated datasets. Initiatives like Cabra, Boana (LLaMA-2-7B based), and Mula (Sparse Mixture of Experts) focus on open-source contributions and accessibility for lower-resource environments. Additionally, Mambarim-110M introduces a non-transformer, state-space model architecture for Portuguese.

GAIA, based on Gemma 3, represents a new approach with continuous pre-training. Most of these models are for research only, and commercial use is prohibited, and may still produce inaccuracies, highlighting the ongoing development in this specialized LLM field.

4 Configurations in PT-BR-LLMs

This section focuses on describing training, architectural, and optimization details about how PT-BR-LLMs works. A complete list of all the published models is found in Tables 1, 2, and 3. Models with articles or reports discussed individually in the previous sections are presented. Figure 4 depicts the total number of encoder, encoder-decoder, and decoder architectures detected in the PT-BR-LLMs.

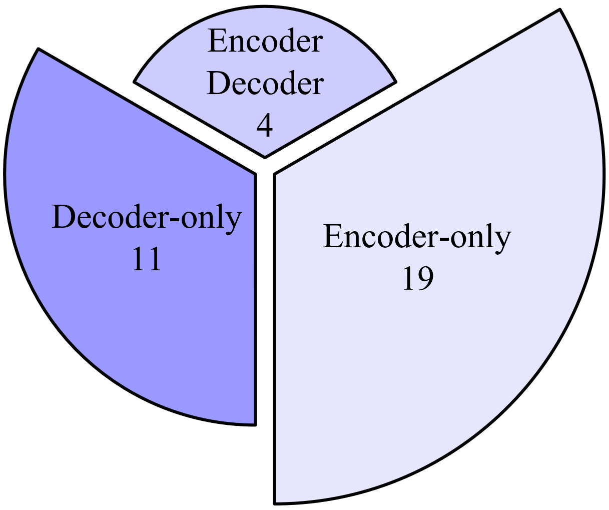

Figure 4. Number of encoder, encoder-decoder, and decoder architectures present in the PT-BR-LLMs.

4.1 Encoder-only PT-BR-LLMs

As shown in Figure 4, from 33 models found, 19 are encoder-only, reflecting the greater focus on PT-BR-LLM development currently. From these 19 models, as illustrated in Figure 5, 14 are based on the BERT architecture, 3 on DeBERTa, and 2 on RoBERTa.

From Table 1, we extract the lists of these models and information related to each. Model contain the name of the specific model. Publication venue is the venue or conference where the model was published (e.g., arXiv'20, BRACIS'20, Clinical NLP'20). The number after the apostrophe indicates the year of publication. License type is the type of license under which the model is distributed. A dash (-) indicates that license information is not available in the table. Model creators identify the entity(ies) or company(ies) that created the model. The last column indicates whether the model can be used for commercial purposes. A check mark (✓) means yes, an ”×” means no, and a dash (-) means the information is not available. Many of the models listed are created by Google.

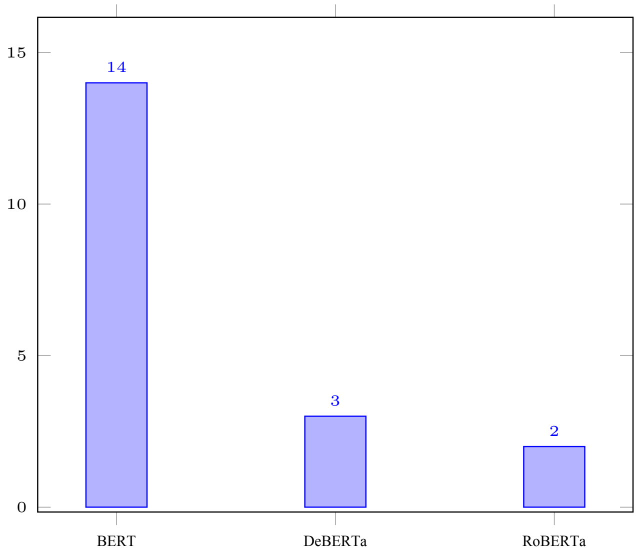

Figure 5. Number of BERT, DeBERTa, and RoBERTa architectures founded in the decoder-only models.

Also, Table 1 shows a comparison of the encoder-only PT-BR-LLMs, focusing on their training characteristics. The number of parameters in the models, indicates its size and complexity. ”M” stands for millions, ”B” for billions. Some models list two numbers, representing different model sizes (e.g., base and large). Steps represents the number of training iterations that the model underwent. ”M” stands for millions, ”K” for thousands. The duration it took to train the model, is expressed in days (d), hours (h), or a combination. Finally, epochs represents the number of epochs that the model was trained for. An epoch represents one complete pass through the entire training dataset. Some models list multiple epoch numbers, possibly indicating training until a certain performance metric was reached or for different configurations.

Figure 6 shows the energy consumption (in kilowatt-hours, kWh) ranges produced by the hardware runtimes, GPUs or TPUs, to train encoder-only models. The varying sizes of the colored segments roughly correspond to the magnitude of the energy consumption ranges, with A100 showing the widest and highest range, and RTX 2080 and TITAN XP showing the lowest consumption.

The data presented in Figure 6 reveals that models using A100 GPUs consumes between 33.2 and 960 kWh (represented by a large red section). Using V100 GPUs consumes between 604.8 and 720 kWh (represented by a smaller red/pink section). Using RTX 3080 GPUs consumes 53.76 kWh (represented by a yellow section). TPUv3-8 consumes between 27.17 and 47.54 kWh (represented by a light yellow section). Models using TPUv2-8 consumes 26.52 kWh (represented by a light green section). Models using TITAN XP GPUs consumes 24 kWh (represented by a darker green section). RTX 2080 GPUs consumes between 1.25 and 12 kWh (represented by a light green section). Thus, the overall impression of this chart is bright a comparison of the energy efficiency or total energy usage of different GPUs/TPUs used to train encoder-only models.

On the other hand, Figure 7 depicts the estimated carbon footprint in kilograms of CO_2 (kg CO_2) produced by the hardware runtimes, GPUs or TPUs. Similar to the energy

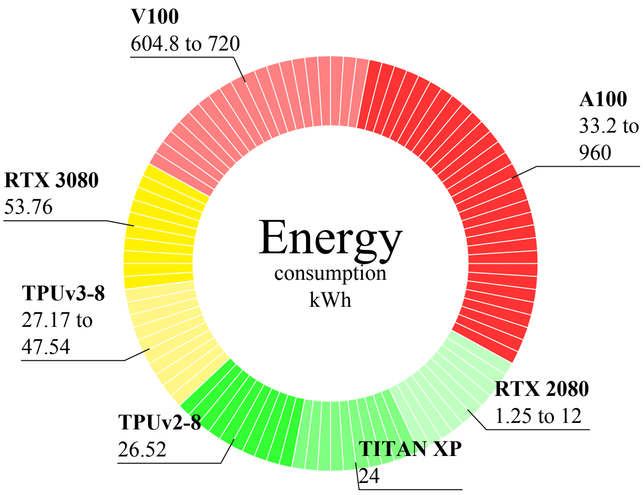

Figure 6. Energy consumption ranges produced by GPUs or TPUs to train encoder-only models.

consumption chart, the sizes of the colored segments in this chart roughly correspond to the magnitude of the carbon footprint ranges. The chart provides a visual comparison of the environmental impact, in terms of carbon emissions, associated with the use of different hardware to train encoder-only models.

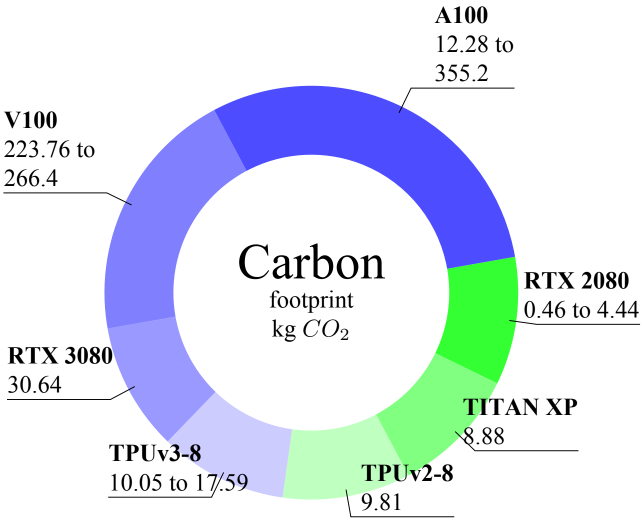

Figure 7. Estimated carbon footprint produced by GPUs or TPUs to train encoder-only models.

The data presented reveals that models using A100 GPUs has a carbon footprint ranging from 12.28 to 355.2 kg CO_2 (represented by a large, dark blue section). This is the widest range and potentially the highest overall carbon footprint among the listed devices. Using V100 GPUs has a carbon footprint ranging from 223.76 to 266.4 kg CO_2 (represented by a lighter blue section). RTX 3080 GPUs has a carbon footprint of 30.64 kg CO_2 (represented by a light purple-blue section). TPUv3-8 has a carbon footprint ranging from 10.05 to 17.59 kg CO_2 (represented by a light purple section).TPUv2-8 has a carbon footprint of 9.81 kg CO_2 (represented by a very light green section). TITAN XP GPUs has a carbon footprint of 8.88 kg CO_2 (represented by a light green section). RTX 2080 GPUs has a carbon footprint ranging from 0.46 to 4.44 kg CO_2 (represented by a darker green section). This appears to be the lowest carbon footprint among the listed devices.

Table 2 dives deep into the internal transformer encoder architectural dimensions of various pre-trained language models, focusing on components like vocabulary size, tokenization, layers, attention heads, and hidden dimensions. Figures 8 and 9 shows that WordPiece and SentencePiece are predominant tokenization methods, with vocabulary sizes typically ranging from 29k to 136k.

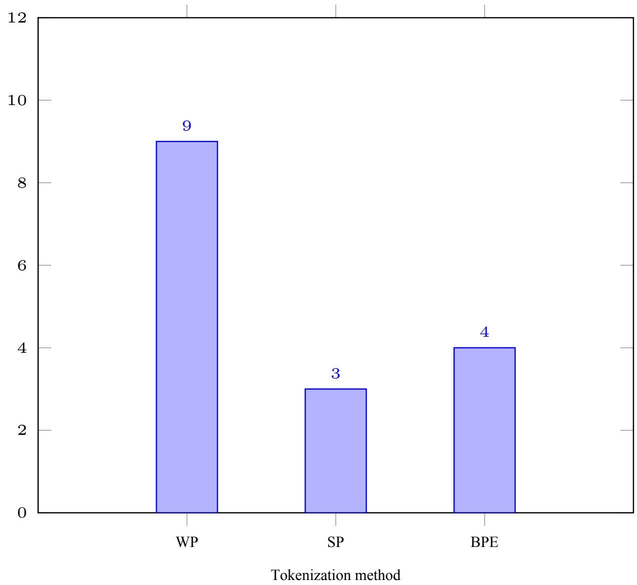

Figure 8. Type and number of tokenization methods detected in encoder-only models.

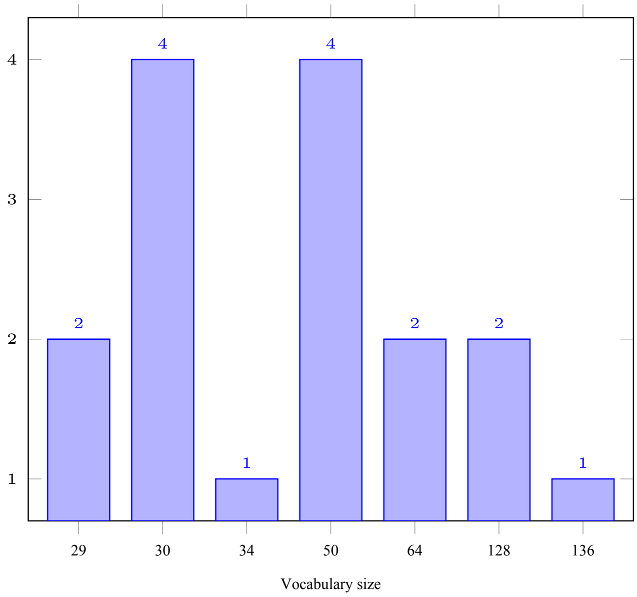

Figure 9. Vocabulary size detected in encoder-only models.

While many models share similar BERT-like encoder structures, there are variations in the number of layers (nL), attention heads (nH), and hidden sizes (hs), suggesting different scales and complexities. For instance, some models have a ”nL” of 12 or 24, and ”nH” of 12. Many fields have a dash (”-”), indicating that this specific detail is either not relevant to that model or not provided in the source from which this table was compiled. This is particularly true for ”Training objective,” ”Attention,” ”Norm,” and ”PE” for many models.

Table 3 summarizes the training configurations and hyperparameters for the encoder-only models. Provides a valuable overview of the training methodologies for encoder-only models, highlighting common practices such as linear learning rate decay and the use of an optimizer, alongside model-specific choices in batch size, sequence length, and other regularization techniques like weight decay, gradient clipping, and dropout. Specific batch size used during training varies significantly across models, from single values like 4 (BioBERTpt) or 256 (BERTaú) to multiple values indicating different phases or configurations, such as 128/256 (BERT CRF) or 2, 4, 8, 12 (PetroBERT, PeLLe). Sequence length values vary, with common ones being 128, 256, and 512. Most models use a learning rate around $1e^{-5}$ or $1e^{-4}$, sometimes with multiple values, suggesting different learning rate schedules or ranges. Most models utilize warm-up. Linear is a common learning rate decay schedule strategy. For weight decay regularization parameter, common values are 0.01. Figure 10 shows the type and number of optimizers used in many models.

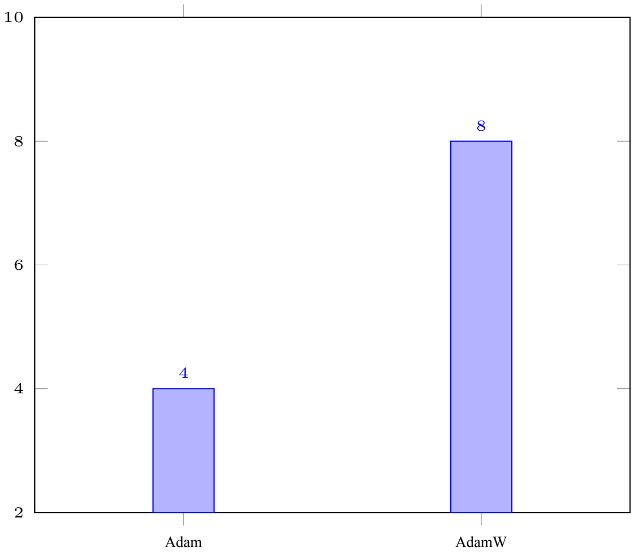

Figure 10. Type and number of optimizers used for encoder-only models.

4.2 Encoder-Decoder PT-BR-LLMs

As detected in Figure 4, 4 models are encoder-decoder (sequence-to-sequence). From these 4 models, as illustrated in Figure 11, 3 are primarily based on T5 architecture, with one model exception based on BART architecture.

Table 1 shows a concise overview of several models, their publication details, licensing, creators, and commercial usability. It highlights that Google has created multiple models under the MIT license, while Facebook AI created a model under the Apache 2.0 license. Only three listed models are being available for commercial use.

Also, the table provides a comparative overview of the training characteristics for the four models and gives a glimpse into the scale and training depth of these models, particularly highlighting the varying sizes. Number of parameter show multiple parameter counts (60M, 220M, 740M), suggesting different scales of the model. Ptt5v2 MonoPTT5 also has two distinct sizes (60M, 3B), indicating a very large variant. The parameter counts from mRAT SQL+GAP and PTT5 Paraphraser are not provided. Epochs show different intervals, from 4 (for PTT5), 6 (to PTT5 Paraphraser), and Ptt5v2 MonoPTT5 stands out with 100K training steps and multiple epoch counts (1, 2, 100), implying more extensive or varied training regimes compared to the others.

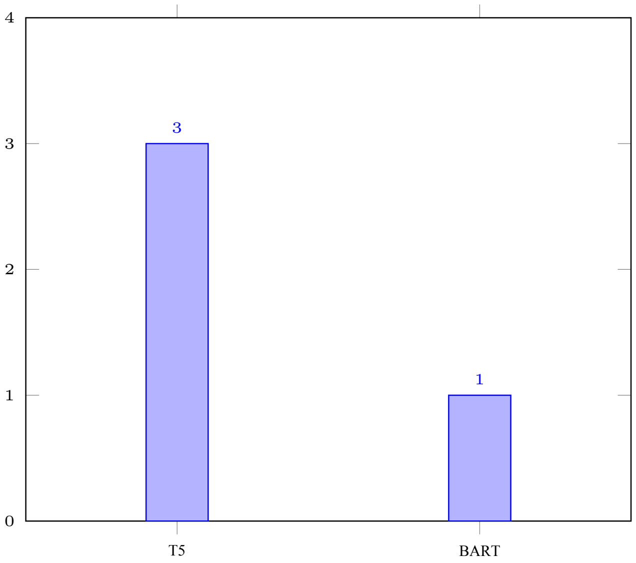

Figure 11. Predominant architectural types for encoder-decoder models.

Notably, training time, energy consumption (Ec kWh), and carbon footprint (Cf kg CO2) are consistently absent for all models, which are important metrics for evaluating the environmental impact and resource efficiency of model training.

Table 2 compares different encoder-decorder models based on various architectural and training characteristics. Provides a concise comparison, highlighting that PTT5 and Pt5v2 share similar characteristics (Encoder-Decoder, Denoising objective, 32k vocabulary, SP tokenizer), with PTT5 specifically mentioning Sigmoid activation and a cross for Bias. PTT5-Paraphraser is also Enc-Dec, but has a larger 36k vocabulary. For mRATSQL+GAP, and MonoPTT5, very little specific information is provided beyond their encoder-decoder architecture type.

Table 3 xx focuses on the training configurations of the encoder-decoder models. PTT5, PTT5-Paraphraser, and Pt5v2 show more detailed information about their training setups compared to mRAT, SQL+GAP, and MonoPTT5. AdaFactor is a common optimizer for PTT5, PTT5-Paraphraser, and Pt5v2. PTT5 also explicitly uses AdamW. PTT5 and Pt5v2 have specified learning rates. PTT5 and PTT5-Paraphraser did not use warm-up, while Pt5v2 did. Specific batch sizes, and sequence lengths are provided for PTT5, PTT5-Paraphraser, and Pt5v2, often with multiple values for PTT5 indicating flexible configurations.

4.3 Decoder-only PT-BR-LLMs

As depicted in Figure 4, 11 models are decoder-only. From these 11 models, as illustrated in Figure 12, two are based on GPT architecture, five are based on Llama architecture, one based on Qwen architecture, and three not mentioned their architecture.

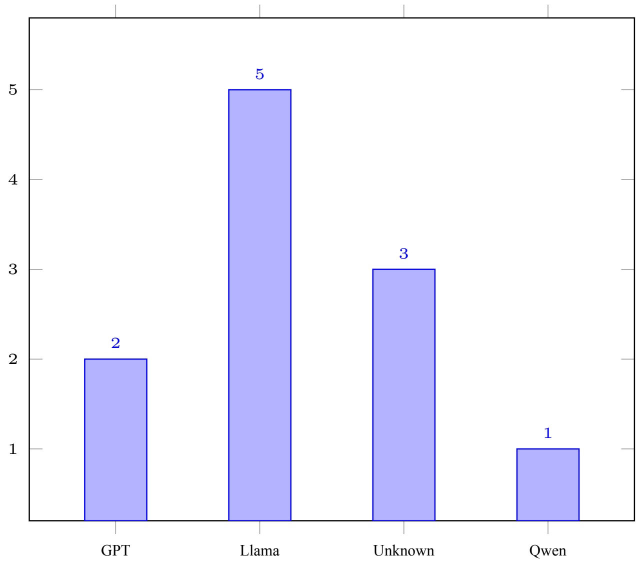

Figure 12. Number and type of architectures founded in the decoder-only models.

Table 1 provides a comprehensive overview of decorder-only models, detailing their publication, licensing, creators, and commercial use, with a noticeable presence of models from Meta AI and a prevalence of arXiv as a publication venue. This indicates a trend towards pre-print sharing of research in this field. The models are released under various licenses, including open-source (Llama2), permissive (Apache 2.0, MIT), and proprietary (Qwen). Several models lack specified license information. Commercial use varies among the models. Some explicitly allow commercial use, while others, such as Sabiá, Sabiá 2, Bode, Juru, and Sabiá 3, explicitly prohibit it. Information is missing for GPT2-Bio-Pt, Tucano, and Amadeus-Verbo. A checkmark ("✓") signifies commercial use is permitted, while an "x" indicates it is not. A dash ("-") means the information is not provided. Many of the models appear to be recently developed or announced.

Also, Table 1 reveals technical specifications, wide range in model sizes (from 124M to 72B parameters) and training efforts. Some models are relatively small and quick to train, while others are massive and require significant computational resources. There's also considerable variability in the completeness of the data provided, with many entries having missing information for parameters, steps, training time, or epochs. This suggests that some models might be in earlier stages of development, or that detailed training logs are not yet publicly available for all of them. The presence of multiple parameter sizes for models like Teeny-TinyLlama, Tucano, and Amadeus-Verbo indicates a common practice of releasing different scales of the same model.

Figure 13 shows the energy consumption (in kilowatt-hours, kWh) ranges produced by the hardware runtimes to train decoder-only models. The chart visually compares the energy consumption of different GPUs or TPUs, specifically NVIDIA's A100, H200, and H100 GPUs, and Google's TPUv2 and TPUv3. It highlights that the NVIDIA GPUs (A100, H200, H100) generally consume significantly more energy over their specified operational periods, with the A100 showing the widest and highest range of energy consumption. The TPUs, particularly TPUv2, demonstrate much lower energy consumption figures in comparison, suggesting they might be more energy-efficient for certain workloads. The ranges for A100, H200, and H100 suggest that the energy consumption can vary greatly depending on factors like the duration of use, and workload intensity.

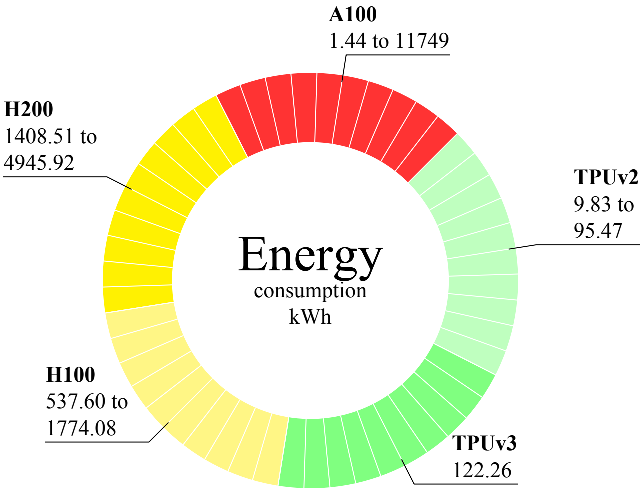

Figure 13. Energy consumption ranges produced by GPUs/TPUs to train decoder-only models.

Figure 14 visually represents the carbon emissions associated with the use of different types of hardware accelerators: NVIDIA's A100, H200, and H100 GPUs, and Google's TPUv2 and TPUv3. NVIDIA GPUs (A100, H200, H100) generally have a significantly higher carbon footprint** compared to Google's TPUs for the scenarios presented. The A100 shows the largest potential range of emissions, reaching up to 4475 kg CO2. Google TPUs (TPUv2, TPUv3) appear to be more carbon-efficient, with their carbon footprints remaining in the tens of kilograms of CO2, which is orders of magnitude lower than the upper bounds for the NVIDIA GPUs. The use of ranges for the NVIDIA GPUs and TPUv2 (as opposed to single values for TPUv3) suggests that the carbon emissions are highly dependent on factors like the duration of use, and workload intensity.

Table 2 provide a comparative analysis of various decoder-only language models, focusing on their architectural and training specifics. The table provides a detailed comparison of decoder-only models, highlighting their diverse choices in training objectives (primarily CLM), attention mechanisms (including grouped query), vocabulary sizes, tokenizers, normalization, positional encodings, and activation functions. Describes the attention mechanism. "Parallel" for Sabiá, "Grouped query" for TeenyTinyLlama, Bode, Tucano, and Amadeus-Verbo.

Figure 15 depicts the specific tokenizer used. "GPT-2 GPT-J SP" for Sabiá, "SP" for Cabrita and TeenyTinyLlama,

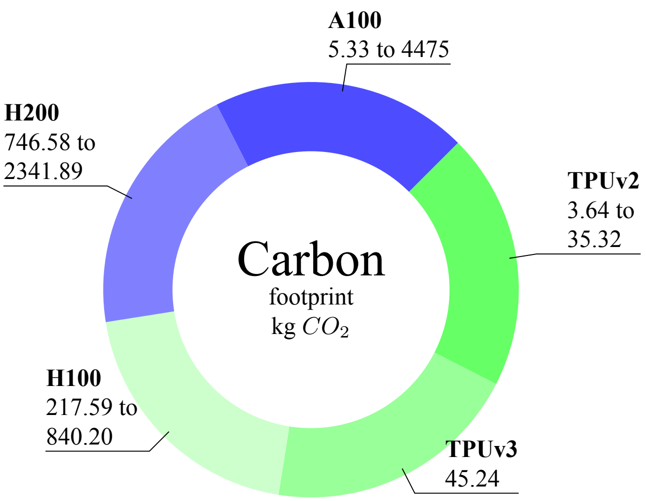

Figure 14. Estimated carbon emissions associated with GPUs/TPUs to train decoder-only models.

”SP+BPE” for Gervacio, and ”BBPE” for Amadeus-Verbo. Figures 16 depicts the vocabulary size. ”1T/4T” for Sabiá (potentially referring to tokens processed/total tokens), ”52k” for Cabrita, ”32k” for TeenyTinyLlama, Gervacio, and Tucano, and ”150k” for Amadeus-Verbo.

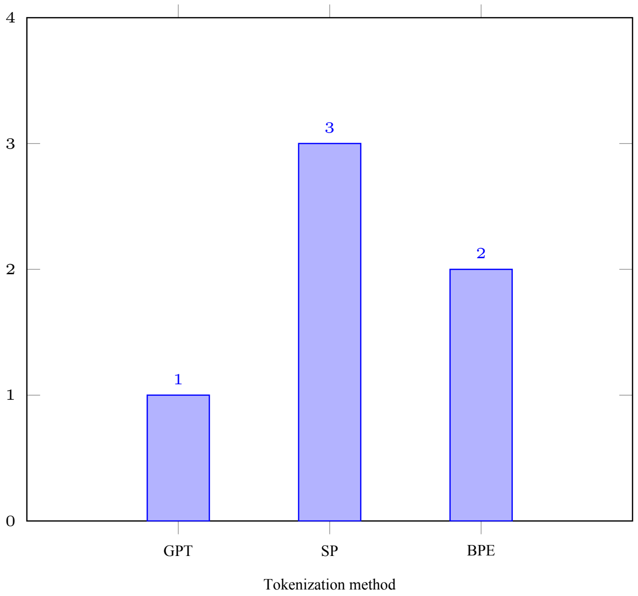

Figure 15. Type and number of tokenizer methods used in decoder-only models.

Also, Table 2 indicates the normalization method. ”RMS” is used by TeenyTinyLlama, Bode, Tucano, and Amadeus-Verbo. Denotes the Positional Encoding method. ”RoPE” (Rotary Positional Embedding) is common, seen in Sabiá, TeenyTinyLlama, Bode, Tucano, and Amadeus-Verbo. ”Rotary” is listed for Cabrita. Lists the activation function. ”GELU” for GPT2-Bio-Pt, ”SwiGlu” for Cabrita, TeenyTinyLlama, Bode, and Amadeus-Verbo, and ”Silu” for Tucano. It also details structural parameters like the number of layers, attention heads, and hidden sizes, often showing multiple configurations for the same model.

Table 3 provides a detailed comparison of training configurations for decoder-only models, highlighting diverse choices in batch sizes, sequence lengths (including very long ones), learning rates and their schedules (linear vs. cosine decay), and optimizers (AdaFactor and AdamW being prominent). It also details the application of regularization techniques such as weight decay, gradient clipping, and dropout. The batch size used during training varies, with some models using fixed sizes and others indicating multiple sizes, possibly for different training stages or configurations. Sequence length, indicates the maximum input sequence length. Values include 1024, 2048, and 32k, indicating very long sequence handling for the latter two (Tucano, Amadeus-Verbo). Learning rate values range from $1e^{-6}$ to $1e^{-3}$, with some models specifying a single value and others a range (e.g., $1.2e^{-5}$, $2.4e^{-6}$ for Sabiá). Most models utilize warm-up and use a weight decay of 0.1 or 0.01. The common learning rate decay schedule is”linear”, while ”cosine” is used by Tucano and Amadeus-Verbo. Figure 17 depicts the most used optimizers in the models.

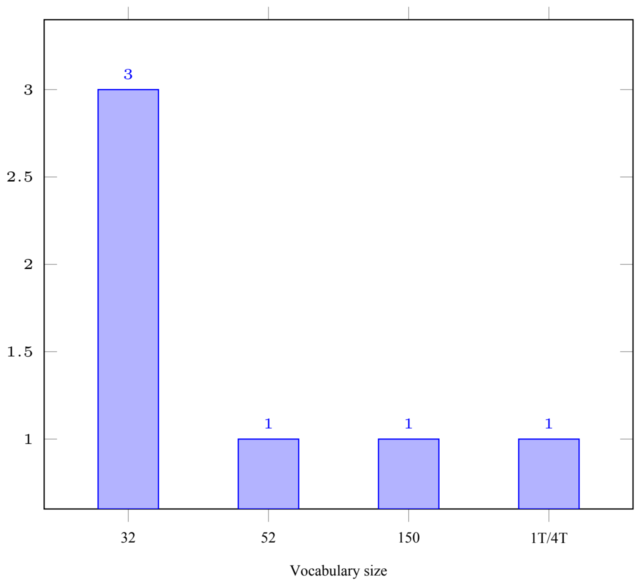

Figure 16. Vocabulary size used in decoder-only models.

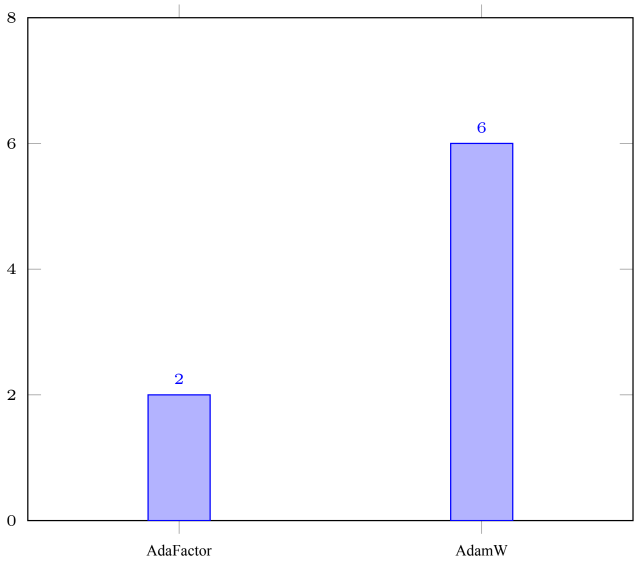

Figure 17. Optimizers used in decoder-only models.

Table 1. Summary of current PT-BR-LLMs settings for the functioning of the models. Here ”Comm. use” represents Commercial use, ”Ec” represents Energy consumption in kWh, and ”Cf” represents Carbon footprint in kg CO_2.

Model	Publication venue	License type	Model creators	No. of parameters	Comm. use	Steps trained	Processing unit type	Training time	Epochs	Ec kWh	Cf kg CO_2
BERT CRF	arXiv’20	MIT	Google	110M, 340M	✓	1M	TPUv3-8	4d, 7d	8	27.17 47.54	10.05 17.59
PTT5	arXiv’20	MIT	Google	60M, 220M 740M	✓	-	TPUv3	-	4	-	-
ELMO-BERT	BRACIS’20	Apache 2.0	Google AllenAI	-	✓	-	-	-	1, 5 10, 75	-	-
BERTimbau	BRACIS’20	MIT	Google	110M, 330M	✓	1M	TPUv3-8	4d, 7d	6, 8 10	27.17 47.54	10.05 17.59
BioBERTpt	Clinical NLP’20	-	Google	-	-	-	GTX2080Ti Titan 12GB	-	5	-	-
GPT2-Bio-Pt	CBMS’21	-	OpenAI	124M	-	-	GTX2080Ti Titan 12GB	-	3	-	-
mRAT SQL+GAP	BRACIS’21	Apache 2.0	Facebook AI	-	✓	-	GeForce RTX 3090 24GB	-	-	-	-
BERTaú	arXiv’21	-	Google	110M	×	1M	V100	-	1, 2 3, 5	-	-
BertBR	UnB’21	CC BY NC-ND	Google	335M	×	1M	GP102 TITAN XP	96h	3	24	8.88
PetroBERT	PROPOR’22	-	Google	-	-	-	GeForce RTX 2080 Ti	5h, 48h	1, 4 8	1.25 12	0.46 4.44
ZeroBERTo	PROPOR’22	-	Google	-	-	-	Tesla K80 12GB	-	-	-	-
BERT legal NER	PROPOR’22	-	Google	110M, 330M	×	1M	Tesla P100 PCIE 16GB	-	10	-	-
PTT5-Paraphraser	PROPOR’22	MIT	Google	-	✓	-	Colab GPU	-	6	-	-
Sabiá	BRACIS’23	Open-source Llama2	EleutherAI Meta AI	-	×	10K	TPUv2-512 TPUv3-8	18d	1.52	95.47 122.26	35.32 45.24
BERTabaporu	RANLP’23	-	Google	-	-	1M	TPUv2-8	120h	3	26.52	9.81
LegalBert-pt	BRACIS’23	OpenRAIL	Google	110M	-	2.4M	-	-	-	-	-
DeBERTinha	arXiv’23	-	Google	-	-	40M	8 A100 80GB 8 V100 32GB	12.5d	1, 3 20	960 720	355.2 266.4
Cabrita	arXiv’23	Apache 2.0	Meta AI	3B	✓	128	TPUv3-8	-	-	-	-
Albertina	EPIA’23	MIT	Google	100M, 900M 1.5B	✓	200K	GCP A2 16 A100	1d11h	5, 25	224	82.88
JurisBERT	ICSSA’23	-	Google	110M	-	220K	2 GeForce RTX 3080 12GB	7d	20	53.76	30.64
Teeny TinyLlama	ArXiv’24 MLWA’24	Apache 2.0	Meta AI	160M, 460M	✓	100K	A100-SXM4 40GB	1.5d, 11.5d	3	1.44 11.04	5.33 40.85
PeLLE	arXiv’24	-	Meta AI Google	125M, 279M	×	100K	-	-	5, 10 40	-	-
Sabiá 2	arXiv’24	-	-	-	×	-	-	-	-	-	-
Bode	arXiv’24	MIT	Meta AI	7B	×	-	4 V100	-	-	-	-
Juru	arXiv’24	-	-	-	×	2800	TPUsv2-128	44.22h	2.94	9.83	3.64
Ptt5v2 MonoPTT5	arXiv’24	MIT	Google	60M, 3B	-	100K	TPUv2-8 TPUv3-8	-	1, 2 100	-	-
Gervácio	LREC COLING’24	MIT	Meta AI	7B	✓	-	GCP A2 16 A100	2h	2	12.8	4.8
Sabiá 3	arXiv’24	-	-	-	×	-	TPUv5	-	-	-	-
Tucano	arXiv’24	Apache 2.0	Meta AI	160M, 630M 1.1B, 2.4B	✓	320K, 400K 480K, 1.9M	8 A100 80GB	44h, 170h 180h, 845h	1, 1.25 1.5, 4	235.54, 920 2524, 11749	89.73, 350 962, 4475
RoBERTaLexPT	PROPOR’24	CC-BY 4.0	Google	125M	✓	50k, 62.5k	2 A100 80GB	72h	8, 17 50	57.6	21.32
GovBERT-BR	BRACIS’24	-	-	-	-	-	-	-	-	-	-
BERTweet.BR	NC&A’25	Apache 2.0	Google	110M	-	7M	4 V100 32GB	504h	30	604.8	223.76
DeB3RTa	BDCC’25	MIT	Google	70M, 426M	✓	80K	A100	83h, 103h	50	33.2, 41.2	12.28, 15.24
Amadeus-Verbo	ArXiv’25	Qwen	Alibaba	0.5B, 1.5B, 3B, 7B 14B, 32B, 72B	✓	-	8 H100 8 H200	-	2	-	-

5 Discussion and Final Remarks

Based on the analysis of all the information extracted from the models and their training configurations, we can draw some final considerations and point to future directions, taking a critical approach and focusing on opportunities for PT-BR-LLMs. The survey presents an interesting overview of models, predominantly of the ”Decoder” type. However, some gaps and critical points deserve attention:

1. **Uniformity vs. Diversity in Configurations.** Although there is some standardization in some aspects (such as the predominance of AdamW or AdaFactor as optimizers and the use of warm-up), the wide variation in parameters such as batch size, sequence length (SL), and learning rate (LR) indicates that there is still no consensus or ideal ”recipe” for training LLMs in Portuguese. This can be both a strength (exploring different approaches) and a weakness (difficulty in replicating and comparing results fairly).
2. **Incomplete Optimization and Regularization.** Columns such as ”LR decay,” ”weight decay,” ”Grad clip,” and ”Dropout” often contain empty cells or sparse information. This may mean that not all models explicitly utilize these techniques consistently or report their values. There is a lack of standardization in the documentation and sharing of this crucial information. Model optimization and regularization are still areas with significant room for improvement and systematic research in Portuguese.
3. **Comparability and Replicability.** The lack of complete details in the tables makes direct comparison and replication of experiments challenging. To advance the field, it is essential that researchers publish all training

Table 2. Architectural details of PT-BR-LLMs. Here, ”PE” means positional embedding, ”nL” number of layers, ”nH” number of attention heads, and ”hs” size of hidden states.

Model	Type	Training objective	Attention	Vocab	Tokenizer	Norm	PE	Activation	Bias	nL	nH	hs
BERT CRF	Enc	-	-	30k	WP	-	-	-	✓	4	-	-
PTT5	Enc-Dec	Denoising	-	32k	SP	-	-	Sigmoid	×	-	-	-
ELMO-BERT	B-LSTM/Enc	-	-	136k	WP	-	-	-	-	-	-	-
BERTimbau	Enc	MLM, NSP	-	30k	SP/WP	-	-	-		12/24	12/16	768/1024
BioBERTpt	Enc	-	-	-	WP	-	-	-	-	-	-	-
GPT2-Bio-Pt	Dec	-	-	-	-	-	-	GELU	-	12	12	-
mRAT SQL+GAP	Enc-Dec	-	-	-	-	-	-	-	-	-	-	-
BERTaú	Enc	-	-	34k	WP	-	Dense	-	✓	4	-	-
BertBR	Enc	-	-	50k	-	-	Learned	ReLU	-	-	-	-
PetroBERT	Enc	-	-	29k	WP	-	-	-	-	-	-	-
ZeroBERTo	Enc	-	-	-	-	-	-	-	-	-	-	-
BERT legal NER	Enc	-	-	29k	WP	-	-	-	-	12/24	12/16	768/1024
PTT5-Paraphraser	Enc-Dec	-	-	36k	-	-	-	-	-	-	-	-
Sabiá	Dec	CLM	Parallel	1T/4T	GPT-2 GPT-J	-	RoPE	-	-	-	256	-
BERTabaporu	Enc	-	-	64k	-	-	-	Sigmoid ReLU Softmax	-	1/2	1/16	16/1280
LegalBert-pt	Enc	-	-	30k	WP SP+BPE	-	-	-	-	12	12	768
DeBERTinha	Enc	-	-	50k	SP	-	-	-	-	-	-	-
Cabrita	Dec	CLM	-	52k	SP	-	Rotary	SwiGlu	-	-	-	-
Albertina	Enc	-	-	128k	SP	-	Relative	-	-	24	-	1536
JurisBERT	Enc	-	-	128k	SP	-	Relative	-	-	24	-	1536
Teeny TinyLlama	Dec	-	Grouped query	32k	SP	RMS	RoPE	SwiGlu	-	12/24	12/16	768/1024
PeLLE	Enc	MLM	-	50k	BPE	-	-	-	-	-	-	-
Sabiá 2	Dec	-	-	-	-	-	-	-	×	-	-	-
Bode	Dec	-	Grouped query	-	-	RMS	Rotary	SwiGlu	×	-	-	-
Juru	Dec	-	-	-	-	-	-	-	-	-	-	-
Ptt5v2 MonoPTT5	Enc-Dec	Denoising	-	32k	SP	-	-	-	-	-	-	-
Gervacio	Dec	CLM	-	32k	SP+BPE	-	-	-	×	32	32	4096
Sabiá 3	Dec	Next-Token	-	-	-	-	-	-	✓	-	-	-
Tucano	Dec	CLM	Grouped query	32k	SP	RMS	RoPE	Silu	×	12/14 22/24	12/16 32	768/2048 2560
RoBERTaLexPT	Enc	MLM	-	50k	BPE	-	-	-	-	12	12	768
GovBERT-BR	Enc	-	-	-	-	-	-	-	-	-	-	-
BERTweet.BR	Enc	MLM	-	64k	fastBPE	-	-	-	-	12	12	768
DeB3RTa	Enc	-	-	128k	SP	-	Relative	Softmax	✓	12/12	6/12	384, 768
Amadeus-Verbo	Dec	-	Grouped query	150k	BBPE	RMS	RoPE	SwiGLU	×	24/28/36 48/64/80	12/14/16 28/40/64	-

hyperparameters explicitly and consistently.

4. **Attention to Attention Mechanisms.** The ”Attention” column points to the use of ”Grouped query” in some models, indicating the exploration of more efficient attention mechanisms. However, the lack of details for many other models suggests that research into attention optimization for Portuguese models still needs further development, especially to efficiently handle longer sequences.
5. **Positional Encoding (PE).** The predominance of ”RoPE” and ”Rotary” indicates a trend toward more robust PEs. However, the lack of information for several models is another gap that makes it difficult to analyze which PEs are most effective for Brazilian Portuguese across different architectures.
6. **Encoder-Decoder vs. Decoder-Only Models.** The predominance of decoder-only models suggests a strong interest in generative models. For applications that require complex understanding and robust representation (such as summarization, translation, or Q&A systems), encoder-decoder models still play a crucial role and may not be receiving the same attention in terms of training optimization for Portuguese.
7. **Encoder-Only vs. Decoder-Only.** This distinction is crucial, as each architecture is best suited for different types of tasks (comprehension vs. generation). The Brazilian community needs to continue developing and optimizing both types, ensuring robust and efficient models for the full spectrum of NLP applications.
8. **Domain-Specific. A Positive but Incipient Trend.** The existence of models like ”BERT legal NER” or ”RoBERTaLexPT” is excellent news, indicating the search for domain specialization. However, most of the listed models lack an explicit domain designation. There is a vast opportunity to create models adapted to other crucial sectors in Brazil (healthcare, economics, education, agriculture, etc.).

Considering the research gaps and context in PT-BR-LLMs, the following opportunities stand out:

- **Systematic Hyperparameter Optimization.** There is a significant opportunity to conduct systematic hyperparameter optimization studies (batch size, LR, decay, weight decay, etc.) specifically for Portuguese models. This could lead to the discovery of optimal configurations that maximize performance and efficiency.

Table 3. Optimization settings used in PT-BR-LLMs. Here ”SL” means sequence length, and ”LR” is Learning Rate.

Model	Batch size	SL	LR	warm up	LR decay	AdaFactor	Adam	AdamW	weight decay	Grad clip	Dropout
BERT CRF	128/256	128/512	$1e^{-4}$	✓	linear	-	✓	-	0.01	-	-
PTT5	64/128/256	128/512	$1e^{-4}, 2e^{-4}, 3e^{-4}$	×	-	✓	-	✓	-	-	-
ELMO-BERT	8/16/32	256	$1e^{-3}, 5e^{-3}, 5e^{-5}$	×	-	-	-	-	-	-	✓
BERTimbau	32/128/256	512	$1e^{-4}, 1e^{-5}, 4e^{-5}$	✓	linear	-	-	✓	0.01	-	-
BioBERTpt	4	256	$2e^{-5}, 3e^{-5}$	✓	linear	-	-	✓	0.01	-	-
GPT2-Bio-Pt	32/64	1024	$5e^{-5}$	✓	-	-	-	✓	-	-	✓
mRAT SQL+GAP	-	-	-	×	-	-	-	-	-	-	-
BERTaú	256	512	$5e^{-5}$	✓	linear	-	-	✓	-	-	-
BertBR	-	-	$2e^{-5}$	×	-	-	-	✓	0.01	-	-
PetroBERT	2, 4, 8, 12	256,512	$2e^{-5}, 4e^{-5}, 5e^{-5}$	×	linear	-	-	-	-	✓	-
ZeroBERTo	20	-	-	✓	-	-	-	-	-	-	-
BERT legal NER	4, 8	512	$2e^{-5}$	-	linear	-	✓	-	0.01	✓	-
PTT5-Paraphraser	8	-	-	×	-	✓	-	-	-	-	-
Sabiá	32/512	2048	$1.2e^{-5}, 2.4e^{-6}$	✓	linear	✓	-	✓	0.1	-	-
BERTabaporu	512	128/512	-	✓	-	-	-	-	-	-	✓
LegalBert-pt	-	-	$1e^{-4}$	-	-	-	-	✓	-	-	-
DeBERTinha	288, 1664	-	$5e^{-5}$	-	-	-	-	✓	-	-	-
Cabrita	16	2048	$3e^{-4}$	-	linear	-	-	✓	0.1	✓	-
Albertina	128, 256	128	$1e^{-5}$	✓	linear	-	-	-	-	-	✓
JurisBert	128	384	$1e^{-4}$	×	linear	-	✓	-	0.01	-	✓
Teeny TinyLlama	16	2048	$6e^{-4}, 3e^{-4}$	✓	linear	-	-	✓	0.01	✓	✓
PeLLE	2, 4, 8, 16	-	$1e^{-3}, 1e^{-4}$	×	-	-	-	-	-	-	×
Sabiá 2	-	-	-	×	-	-	-	-	-	-	×
Bode	-	-	-	×	-	-	-	-	-	-	✓
Juru	-	-	$1e^{-3}$	✓	-	✓	-	-	✓	✓	×
Ptt5v2 MonoPTT5	128	512	$1e^{-3}$	✓	-	✓	-	-	-	-	-
Gervacio	16	512	$2e^{-5}$	×	-	-	-	-	0.1	-	-
Sabiá 3	-	-	-	-	-	-	-	-	-	-	-
Tucano	32, 512	32k	$1e^{-6}$	✓	cosine	-	-	✓	0.1	✓	×
RoBERTaLexPT	16, 32 2048	128, 512	$7.5e^{-6}, 1e^{-5}, 2.5e^{-5}$ 4e^{-4}, $5e^{-5}$	✓	constant linear	-	-	✓	0.01	✓	✓
GovBERT-BR	-	-	-	-	-	-	-	-	-	-	-
BERTweet.BR	96	128	$1e^{-4}$	✓	-	-	✓	-	-	✓	×
DeB3RTa	16, 32, 64	128	$1e^{-4}$	✓	linear	-	-	✓	0.1	-	-
Amadeus-Verbo	16	32k	$1e^{-5}$	✓	cosine	-	-	✓	0.1	✓	×

- **Attention to Efficiency and Scale.** With the emergence of models with very long context sequences, research into more efficient attention mechanisms, such as Grouped Query Attention and its variants, is crucial. For Portuguese, which has a rich morphology and can generate longer sentences, optimizing efficiency in extended sequences is a key differentiator.
- **Tool Development and Benchmarks.** The lack of detail in several columns suggests the need for more comprehensive benchmarks and standardized tools for evaluating and comparing models in Portuguese. This would allow for a more rigorous assessment of progress and identify areas most in need of research.
- **Data Diversity and Tasks.** Although the table does not detail the training data, the quality and diversity of Brazilian Portuguese corpora are crucial. There is an ongoing opportunity to collect and curate datasets that represent the richness of the language across different domains (legal, health, news, conversational, etc.), which would lead to more robust and versatile models.
- **Specific Applications for Brazilian Portuguese - Text Generation with Regional Nuances.** Develop models that capture the nuances and regionalisms of Brazilian Portuguese, going beyond the standard norm. **Legal and Medical Natural Language Processing.** There is a growing demand for LLMs specialized in domains specific to Brazil, such as law (with its legislative peculiarities) and medicine (with its own terminology and jargon). **Enhanced Conversational Interaction.** Models capable of maintaining more natural and contextually relevant dialogues for Brazilian users, incorporating slang, idiomatic expressions, and a deeper cultural understanding. **Support for Indigenous Languages and Regional Variants.** A more ambitious area, but with great social impact, would be the exploration of LLMs that can assist in the preservation and processing of minority languages and regional variants of Portuguese in Brazil.
- **Multimodal Modeling Research for Brazilian Portuguese.** Although not explicitly mentioned in the table, the combination of text with other modalities (image, audio, video) is an important frontier. For Brazilian Portuguese, this may mean developing models that understand and generate multimodal content relevant to the local culture and context.
- **Systematic and Efficient Training Optimization. Hyperparameter Sensitivity Studies.** Conduct systematic research to understand the sensitivity of models to different batch sizes, SLs, LRs, and decay strategies for Portuguese, aiming to identify robust and efficient

configurations. **Hardware Utilization.** Research and develop distributed training, quantization, pruning, and distillation techniques optimized for Portuguese, allowing high-quality models to be trained and used with more modest computational resources, democratizing access.

- **Corpora Expansion and Curation. Data Diversity and Quality.** Pre-training LLMs requires massive, high-quality corpora. There is an ongoing opportunity to collect and curate datasets that represent the richness of Brazilian Portuguese in terms of dialects, registers (formal, informal, colloquial), domains (scientific, legal, journalistic texts, social media), and even regional varieties. This is essential to avoid biases and improve model robustness. **Multimodal Data.** In the future, the focus should expand to multimodal corpora in Portuguese (text-image, text-audio), enabling the development of richer and more interactive LLMs.
- **Active Exploration of New Architectures and Mechanisms.** Beyond BERT and GPT: Although dominant, it is crucial that Brazilian research actively explore emerging architectures and innovative attention mechanisms (e.g., more efficient for long context, such as linear attention, recurrent neural networks with attention, etc.) that may be better suited to the specificities of Portuguese or more resource-efficient.
- **Multi-Agent Systems and Critical Models.** Building more complex systems with LLMs in Brazilian Portuguese, where models can interact, reason, and even critique their own outputs, striving for greater accuracy and reliability.

6 Conclusion

Advances in PT-BR-LLMs will depend on the existence of pre-trained datasets with diverse data types. Currently, technical obstacles arise due to the distinctive characteristics of such data.Text generation is the broadest task among PT-BR-LLMs, however, some representative domains emerge as applications, highlighting Twitter text classification, financial, code translation, clinical and biomedical domains, Legal, and Oil and Gas. However, the training process and dataset management often impose substantial computational demands. Optimizing model parameters represents another aspect of ensuring a correct evaluation of metrics. Furthermore, fine-tuning and inference require considerable computational resources, posing challenges for academia and industries with constrained hardware capabilities. Nonetheless, researchers and developers implement techniques that encompass working with data subsets and transfer learning to leverage insights gained from previous tasks. The Transformer architecture, in general, has become the basis for several PT-BR-LLMs. To improve efficiency, some PT-BR-LLMs aim to design emergent architectures, such as State Space Model (SSM) Mamba or with a sparse Mixture of Expert (MoE), in which a subset of neural network weights for each input are sparsely activated. The main merit of these architectures are a flexible way to increase the model parameter while maintaining a constant computational cost.

This survey presents a chronological order of the recent progress with pre-trained LLMs for Brazilian Portuguese. In particular, the survey introduces models, resources, datasets, project repositories, configurations, and architectures for understanding and utilizing LLMs for Brazilian Portuguese. Furthermore, were summarized some available configurations for developing, implementing, and reproducing those LLMs. This survey tries to cover the most recent literature about LLMs for Brazilian Portuguese and provides a good reference resource on this topic for both researchers in academia or industry and engineers. The LLMs landscape in Brazilian Portuguese is vibrant, but still maturing. The key to the future lies in greater standardization, transparency, and collaboration in research, combined with a strategic focus on optimization, data diversity, and application in specific domains to maximize the impact of these technologies on Brazilian society.

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.

Availability of data and materials

Papers analyzed during the current study are publicly available.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. DOI: 10.48550/arxiv.2410.12049.

Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., Siqueira, F. L., and Costa, A. H. R. (2022). Zeroberto: Leveraging zero-shot text classification by topic modeling. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 125–136, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-98305-5$_1$2.

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. DOI: 10.48550/arxiv.2403.09887.

Bonifacio, L. H., Vilela, P. A., Lobato, G. R., and Fernandes, E. R. (2020). A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 648–662, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8$_4$6.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. DOI: 10.48550/arxiv.2303.12712.

Campiotti, I., Rodrigues, M., Albuquerque, Y., Azevedo, R., and Andrade, A. (2023). Debertinha: A multistep approach to adapt debertav3 xsmall for brazilian portuguese natural language processing task. DOI: 10.48550/arxiv.2309.16844.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. DOI: 10.48550/arxiv.2008.09144.

Carneiro, F., Vianna, D., Carvalho, J., Plastino, A., and Paes, A. (2025). Bertweet.br: a pre-trained language model for tweets in portuguese. *Neural Computing and Applications*, 37(6):4363–4385. DOI: 10.1007/s00521-024-10711-3.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2024). Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1). DOI: 10.48550/arxiv.2204.02311.

Ciurlino, V. (2021). Bertbr : a pretrained language model for law texts. Available at:`https://bdm.unb.br/handle/10483/27824`.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.747.

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

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

Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., and Paraboni, I. (2023). BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria. Available at:`https://aclanthology.org/2023.ranlp-1.24/`.

Cruz-Castañeda, W. A. and Amadeus, M. (2025). Amadeus-verbo technical report: The powerful qwen2.5 family models trained in portuguese. DOI: h10.48550/arxiv.2506.00019.

Cui, Y., Yang, Z., and Yao, X. (2024). Efficient and effective text encoding for chinese llama and alpaca. DOI: 10.48550/arxiv.2304.08177.

de Mello, G. L., Finger, M., , Serras, F., de Mello Carpi, M., Jose, M. M., Domingues, P. H., and Cavalim, P. (2024). Pelle: Encoder-based language models for brazilian portuguese based on open data.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. (2024). A survey on in-context learning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.64.

Finardi, P., Viegas, J. D., Ferreira, G. T., Mansano, A. F., and Caridá, V. F. (2021). Bertaú: Itaú bert for digital customer service. DOI: 10.48550/arXiv.2101.12015.

Garcia, E. A. S., Silva, N. F. F., Siqueira, F., Albuquerque, H. O., Gomes, J. R. S., Souza, E., and Lima, E. A. (2024a). RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. Available at:`https://aclanthology.org/2024.propor-1.38.pdf`.

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

Geng, X. and Liu, H. (2023). Openllama: An open reproduction of llama. Available at:`https://github.com/openlm-research/open_llama`.

Guillou, P. (2020). Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...). Available at:`https://huggingface.co/pierreguillou/gpt2-small-portuguese`.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. (2024). Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.2203.15556.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-acl.67.

José, M. A. and Cozman, F. G. (2021). mrat-sql+gap: A portuguese text-to-sql transformer. In Britto, A. and Valdivia Delgado, K., editors, *Intelligent Systems*, pages 511–525, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-91699-2$_3$5.

Junior, R. M., Pires, R., Romero, R., and Nogueira, R. (2024). Juru: Legal brazilian large language model from reputable sources. DOI: 10.48550/arxiv.2403.18140.

Kalyan, K. S. (2024). A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048. DOI: 10.1016/j.nlp.2023.100048.

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.878.

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

Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. (2024a). Pre-trained language models for text generation: A survey. *ACM Comput. Surv.*, 56(9). DOI: 10.1145/3649449.

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., and Du, M. (2024b). Language ranker: A metric for quantifying llm performance across high and low-resource languages. DOI: 10.1609/aaai.v39i27.35038.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9). DOI: 10.1145/3560815.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. DOI: 10.48550/arxiv.1907.11692.

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. (2023). When less is more: Investigating data pruning for pretraining llms at scale. DOI: 10.48550/arxiv.2309.04564.

Matarazzo, A. and Torlone, R. (2025). A survey on large language models with some insights on their capabilities and limitations. DOI: 10.48550/arxiv.2501.04040.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2). DOI: 10.1145/3605943.

Miranda, B., Lee, A., Sundar, S., Casasola, A., and Koyejo, S. (2024). Beyond scale: The diversity coefficient as a data quality metric for variability in natural language data.

Naous, T., Ryan, M. J., Ritter, A., and Xu, W. (2024). Having beer after prayer? measuring cultural bias in large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.862.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2024). A comprehensive overview of large language models. DOI: 10.48550/arXiv.2307.06435.

OpenAI (2022). Introducing chatgpt. Available at:`https://openai.com/index/chatgpt/`.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2024). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.2203.02155.

Pellicer, L. F. A. O., Pirozelli, P., Costa, A. H. R., and Inoue, A. (2022). Ptt5-paraphraser: Diversity and meaning fidelity in automatic portuguese paraphrasing. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 299–309, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-98305-5$_2$8.

Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language. DOI: 10.1007/978-3-031-79032-4$_2$3.

Pires, H., Paucar, L., and Carvalho, J. P. (2025). Deb3rta: A transformer-based model for the portuguese financial domain. *Big Data and Cognitive Computing*, 9(3). DOI: 10.3390/bdcc9030051.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland. DOI:

10.1007/978-3-031-45392-$2_1$5.

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., and Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, 6(1). DOI: 10.1016/j.patter.2024.101118.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897. DOI: 10.1007/s11431-020-1647-3.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., *et al.* (2022). Scaling language models: Methods, analysis and insights from training gopher.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, page 441–453. Springer Nature Switzerland. DOI: 10.1007/978-3-031-49008-$8_3$5.

Rodrigues, R. B. M., Privatto, P. I. M., de Sousa, G. J., Murari, R. P., Afonso, L. C. S., Papa, J. P., Pedronette, D. C. G., Guilherme, I. R., Perrout, S. R., and Riente, A. F. (2022). Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 101–109, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-98305-$5_1$0.

Santos, R., Silva, J. R., Gomes, L., Rodrigues, J., and Branco, A. (2024). Advancing generative AI for Portuguese with open decoder gervásio PT*. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 16–26, Torino, Italia. ELRA and ICCL. DOI: 10.48550/arXiv.2402.18766.

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

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.clinicalnlp-1.7.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2025). Govbert-br: A bert-based language model for brazilian portuguese governmental data. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 19–32, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-79032-4_2.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 268–282, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45392-$2_1$8.

Souza, F., Nogueira, R., and Lotufo, R. (2020a). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-$8_2$8.

Souza, F., Nogueira, R., and Lotufo, R. (2020b). Portuguese named entity recognition using bert-crf. *arXiv*, (1909.10649). DOI: 10.48550/arxiv.1909.10649.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models. 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. DOI: 10.48550/arxiv.2307.09288.

Viegas, C. F. O., Costa, B. C., and Ishii, R. P. (2023). Jurisbert: A new approach that converts a classification corpus into an sts one. In Gervasi, O., Murgante, B., Taniar, D., Apduhan, B. O., Braga, A. C., Garau, C., and Stratigea, A., editors, *Computational Science and Its Applications – ICCSA 2023*, pages 349–365, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-36805-$9_2$4.

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

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. DOI: 10.48550/arxiv.2206.07682.

Yue, X., Song, Y., Asai, A., Kim, S., de Dieu Nyandwi, J., Khanuja, S., Kantharuban, A., Sutawika, L., Ramamoorthy, S., and Neubig, G. (2025). Pangea: A fully open multilingual multimodal llm for 39 languages. DOI: h10.48550/arxiv.2410.16153.

Zanuz, L. and Rigo, S. J. (2022). Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 219–229, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-98305-$5_2$1.

Zhang, P., Zeng, G., Wang, T., and Lu, W. (2024). Tinyllama: An open-source small language model. 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., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A survey of large language models. DOI: 10.48550/arXiv.2303.18223.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., and Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. DOI: h10.48550/arXiv.2302.09419.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76. DOI: 10.1109/JPROC.2020.3004555.