Machine Learning with Applications 16 (2024) 100558

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

TeenyTinyLlama: Open-source *tiny* language models trained in Brazilian Portuguese☆

Nicholas Kluge Corrêa [a,b,*], Sophia Falk [a], Shiza Fatimah [a], Aniket Sen [a], Nythamar De Oliveira [b]

[a] University of Bonn, Germany
[b] Pontifical Catholic University of Rio Grande do Sul, Brazil

ARTICLE INFO

Dataset link: https://huggingface.co/collections/nicholasKluge/teenytinyllama-6582ea8129e72d1ea4d384f1

Keywords:
Large language models
Portuguese
Text generation
Low-resource settings
Low-resource languages

ABSTRACT

Large language models (LLMs) have significantly advanced natural language processing, but their progress has yet to be equal across languages. While most LLMs are trained in high-resource languages like English, multilingual models generally underperform monolingual ones. Additionally, aspects of their multilingual foundation sometimes restrict the byproducts they produce, like computational demands and licensing regimes. In this study, we document the development of open-foundation models tailored for use in low-resource settings, their limitations, and their benefits. This is the *TeenyTinyLlama* pair: two compact models for Brazilian Portuguese text generation. We release them under the permissive Apache 2.0 license on GitHub and HuggingFace for community use and further development.

1. Introduction

Large language models have radically changed the field of natural language processing (NLP) with their exceptional ability to perform downstream tasks after being trained on vast amounts of data in a self-supervised learning regime. Under this paradigm, transformer-based models like BERT (Devlin, Chang, Lee, & Toutanova, 2018), RoBERTa (Liu et al., 2019), mT5 (Xue et al., 2020), and the whole family of GPT-style models (Almazrouei et al., 2023; Biderman et al., 2023; Black et al., 2022; Gunasekar et al., 2023; Jiang et al., 2024; Luo et al., 2023; Radford et al., 2019; Touvron et al., 2023; Workshop et al., 2022), have become the foundation for many NLP applications and research areas. While part of the field is still pushing the search for new architectures, with innovations like the RWKV (Peng et al., 2023) and state space models (Gu & Dao, 2023) promising new directions for research, the majority of LLM research remains focused on the scaling of model size, training data, and the general efficiency and capabilities of transformer-based LLMs (Vaswani et al., 2017).

Despite the tremendous success of the field, progress has yet to be made equally regarding all languages. Hence, another trend in current NLP research involves the expansion of language domains with which such systems can interact. Current practices to tackle our linguistic multitude involve either training singular models in multiple languages (Conneau et al., 2019; Lin et al., 2021; Shliazhko et al., 2022; Workshop et al., 2022) or fine-tuning foundational models trained on multi-linguistic corpora to become monolingual or more proficient when working with low-resource languages (Alabi, Adelani, Mosbach, & Klakow, 2022; Eisenschlos et al., 2019; Guillou, 2020; Lankford, Afli, & Way, 2023; Nguyen, Aljunied, Mahani and Joty & and Bing, 2023; Pires, Abonizio, Rogério, & Nogueira, 2023; Zhao, Zhang, Zhang, Gui, & Huang, 2024).

However, most multilingual models available today still have a disproportional performance across languages due to the imbalance of training data, where usually high-resource languages, like English, represent the majority of such corpora, which creates user dissatisfaction with multilingual model's capabilities on non-English languages. Meanwhile, fine-tuned byproducts of multilingual models sometimes end up being restricted by the conditions imposed by the foundation used, like high computational costs for training and inference, which restrain adoption in low-resource settings, besides licensing regimes that prevent free use and open-source development. Factors like these highlight the necessity of building the foundations for monolingual LLMs for low-resource languages (Antoun, Baly, & Hajj, 2021; Gutiérrez-Fandiño et al., 2021; Ko et al., 2023; Martin et al., 2020; Nagoudi, Chen, Abdul-Mageed, & Cavusogl, 2021; Rodrigues et al., 2023; Scheible, Thomczyk, Tippmann, Jaravine, & Boeker, 2020; Souza, Nogueira, & Lotufo, 2020).

☆ This research was funded by RAIES (Rede de Inteligência Artificial Ética e Segura). RAIES is a project supported by FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul), Brazil and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

* Corresponding author at: University of Bonn, Germany.
E-mail address: kluge@uni-bonn.de (N.K. Corrêa).

https://doi.org/10.1016/j.mlwa.2024.100558
Received 29 February 2024; Received in revised form 23 April 2024; Accepted 9 May 2024
Available online 10 May 2024

This study follows the trend of developing LLMs tailored for low-resource regimes (Gunasekar et al., 2023; Stability AI Language Team, 2024; Zhang, Zeng, Wang, & Lu, 2024). In it, we sought to explore the challenges of developing LLMs in such settings. For this, we chose Brazilian Portuguese as our target language. While writing this paper, only a few LLMs for text generation were trained or fine-tuned to be proficient in Brazilian Portuguese. Meanwhile, even fewer were available to the general public with permissive licenses and open-source code development. The models we developed, the *TeenyTinyLlama* pair (TTL), were created to help democratize LLMs for low-resource languages and the open-source community in general, providing a simple and extensible implementation for LLM pre-training and fine-tuning at a small (< 2B parameters) scale. Finally, to our knowledge, the TTL pair is the first text-generation, auto-regressive transformer to be trained natively in Brazilian Portuguese.

2. Related works

As already stated, multilingual models like BLOOM (Workshop et al., 2022), mGPT (Shliazhko et al., 2022), and XGLM (Lin, Mihaylov, et al., 2021) usually do not have satisfactory performance on low-resource languages, especially when compared to monolingual models. To overcome this, much of the community repurposes models trained on multilingual corpora to create mono-linguistic models for a target language via supervised fine-tuning (SFT), or, as some call it, "extended pre-training" (Larcher et al., 2023; Pires et al., 2023).

Sometimes, this strategy even works on models not trained on multilingual datasets. For example, one of the first text-generation models for Brazilian Portuguese to appear to the general public was Pierre Guillou GPorTuguese-2 in 2020 (Guillou, 2020), which, to our knowledge, coincided with the release of the first LLM natively trained in Brazilian Portuguese (BERTimbau) (Souza et al., 2020). GPorTuguese-2 is a byproduct of fine-tuning GPT-2 small (Radford et al., 2019) on the Portuguese portion of the Wikipedia dataset (Wikimedia Foundation, 2024) while also modifying the structure of the original model, giving it a new byte-pair encoding (BPE) tokenizer and repurposing the joint embeddings from the original model. While capable of generating fluent text in Brazilian Portuguese, GPorTuguese-2 fails to achieve a perplexity score on par with GPT-2 small, as documented by OpenAI, which is not surprising, given the limited hours of training (≈ 30 h of training) it received. Regardless, this pioneering work is available to all users under a permissive MIT License (as all GPT-2 models).

However, since the proposition of the GPT-2 architecture, many advances in GPT-style models have been made, and currently, more modern models are the default choice for engineers (Black et al., 2022; Jiang et al., 2023; Shoeybi et al., 2019). Perhaps one of the most significant contributions to the open source community in 2023 was the release of the Llama 2 architecture (Touvron, Martin, et al., 2023). Being the successor of Llama (Touvron et al., 2023), it brings many improvements that make training and inference of transformer-based language models more efficient, like the use of grouped query attention (Ainslie et al., 2023), better sub-layer normalization techniques (Zhang & Sennrich, 2019), changing ReLU activation's by SwiGLU (Shazeer, 2020), and the use of rotary positional embeddings instead of positional ones (Su, Lu, Pan, Wen, & Liu, 2021), besides being trained on a massive pre-training corpus, even beyond what is estimated to be optimal by the scaling Chinchilla laws (Hoffmann et al., 2022). Its current successor, Llama 3,[1] although bigger and trained on a much larger dataset, still inherits the same basic architecture.

Currently, there is an entire ecosystem of Llama-based models being released on a Cambrian explosion rate (Bi et al., 2024; Geng & Liu, 2023; Luo et al., 2023; NousResearch, 2024; Roziere et al., 2023; Zhang et al., 2024), with other open-architectures like Mistral (Jiang et al., 2023) also being heavily used by the open source community. Yet, it is from the LLama models that most of the currently fine-tuned models in Brazilian Portuguese come from. Three byproducts of fine-tuning Llama 1/2 with Brazilian Portuguese corpora are Bode (Gabriel Lino Garcia et al., 2024), Sabiá (Pires et al., 2023; Sales Almeida, Abonizio, Nogueira, & Pires, 2024), and Canarim (Domingues, 2023a).

Bode is a low-rank adaptation (LoRA) of Llama 2 fine-tuned with a translated version of the Alpaca dataset (Taori et al., 2023), which contains 52 K instruction-following demonstrations generated by text-davinci-003 (Ouyang et al., 2022). Bode is offered as LoRA adapters for Llama 7B and 13B while having a fine-tuned version of Llama 7B without using LoRA or other parameter-efficient fine-tuning techniques.[2] In short, Bode only went through a fine-tuning phase, and the results of the model capacity are documented in their paper (Gabriel Lino Garcia et al., 2024). According to these, Bode stays within the performance of models that underwent a similar training regime, like Cabrita (Larcher et al., 2023). It is also on par with its base model and even surpasses it on specific evaluations performed by the authors. However, like all derivatives of Llama models, Bode is licensed under the Llama 2 Community License Agreement, which is not as permissive as Apache 2.0, MIT, or commercial versions of the Creative Commons licenses.

Meanwhile, Sabiá models are fine-tuned versions of GPT-J (Wang & Komatsuzaki, 2021), and Llama trained on a filtered portion of the ClueWeb 2022 dataset (Overwijk, Xiong, & Callan, 2022), which equates to 7.3–7.8 billion tokens, according to GPT-J and Llama tokenizers, respectively. The outcomes of this fine-tuning process are Sabiá-7B, 65B (both derivatives of Llama), and Sabiá-J (using GPT-J as a base). According to the authors, their evaluations show that Sabiá-65B outperforms Llama 2-65B and GPT-3.5-turbo on the Portuguese Evaluation Tasks (Poeta) benchmark (a set of tests gathered by the authors of the Sabiá paper). However, Sabiá-65B and Sabiá-J are not available to the public,[3] while Sabiá-7B, just like the Bode models, was released with a Llama 2 license. Also, none of these models' training, evaluation, or fine-tuning source codes have been released to the community. Even less is known about Sabiá 2 (Sales Almeida et al., 2024), in which, in its development, the authors have chosen to limit substantially more the sharing of information that could help reproduce their results. These types of practices, we argue, only help to aggravate the current reproducibility crisis in ML (Kapoor & Narayanan, 2022), where closed-source development and lack of reproducibility hinder the scientific status of ML research.

Regarding Canarim, there is not much to be said except that it is also a Llama 2 licensed model that underwent a fine-tuning process on 16 billion tokens from a Portuguese subset of Common Crawl (2023) and was further fine-tuned on two datasets, one for instruction-tuning (Domingues, 2023b) and the other for open-ended question answering with a focus on the ENEM exams. Again, the models are also under the licensing regime of the Llama model's community license. For the interested reader, there are many other examples of these types of models on repositories like Hugging Face (de Camaret, 2024; Henrique, 2024a, 2024b; Souza, 2024).[4]

Finally, there are the Cabrita models (Larcher et al., 2023). Unlike the work done by Gabriel Lino Garcia et al. (2024), Pires et al. (2023), and Domingues (2023a), Cabrita is a byproduct of fine-tuning OpenLLaMA 3B (Geng & Liu, 2023), which comes with an Apache 2.0 license. Also, Cabrita models have a modified tokenizer, unlike Sabiá, Bode, and Canarim, which directly repurposed the original Llama tokenizers.

[1] Released by Meta AI in the following blog post: https://ai.meta.com/blog/meta-llama-3/.

[2] It could also be a result of merging LoRA weights with the original Llama 2 weights.

[3] Even though GPT-J is available under an Apache 2.0 License.

[4] That is, LLMs repurposed for other languages via full or LoRA/PEFT fine-tuning not accompanied by a paper or report.

Cabrita extended training was performed on a subset of the mC4 dataset (Raffel et al., 2019), using ≈ 7 billion tokens. Cabrita 3B, like its base model, is available under an Apache 2.0 license.[5]

To our knowledge, at the writing of this paper and release of our models, open-source LLMs for text generation pre-trained natively in Brazilian Portuguese were nonexistent, nor have the above projects open-sourced their training and evaluation methods for reproducibility and further community development. At the same time, all models cited, based on billion-sized transformers, require non-trivial computational resources to use, adapt, and reproduce in low-resource settings. Hence, this study aimed to produce a pair of compact LLMs in Brazilian Portuguese by pre-training them from scratch (the TTL pair), tailored for (and produced by) a low-resource environment. The rest of this work documents our development, experiments, and results.

3. Pre-training

In this section, we will describe how we designed the training of our TTL models. This study was performed with a closed budget of 50,000 USD, forcing us to make many developmental decisions to reduce costs and optimize our pre-training runs. This low-resource setting also influenced the size of our models that, as already shown by Pires et al. (2023), can range from 9000 to 80,000 USD when just fine-tuning billion parameter-sized models on a dataset similar to what we have gathered. Meanwhile, training models with trillion tokens, even in small settings, is beyond what we could finance and what is possible to accumulate with available Brazilian Portuguese text datasets. However, scaling down allowed us to choose a range of sizes where, according to scaling laws (Hoffmann et al., 2022; Kaplan et al., 2020), our limited budget was enough to pay for the computing necessary to pre-train our models and evaluate them.

By providing a straightforward code implementation and emphasizing scaling laws, we aim to offer a blueprint that researchers can customize for their specific linguistic contexts. Hence, the underlying methodology opens doors for exploration and adaptation across various settings, conditioned that researchers have access to a minimal level of know-how, data, and compute (50,000 USD worth of A100 h).

3.1. Sizing up models and datasets

While empirical evidence seems to point to the fact that existing scaling laws (Hoffmann et al., 2022) may not provide accurate predictions in situations where models are trained for extended (i.e., past the Chinchilla estimations of optimal scaling) periods (Touvron, Martin, et al., 2023; Zhang et al., 2024), in this study, we choose to use the Hoffmann et al. (2022) scaling laws, like done by Dey et al. (2023), to estimate the size of models. Even though extrapolating such boundaries might benefit models of different sizes, we did not have the budget (or tokens) to sustain longer runs.

According to Hoffmann et al. (2022), we can model language modeling loss, L, as a function of model size N (the number of parameters) and training dataset size D (the number of tokens):

$$L(N,D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E.$$

Where $A = 406.4$, $B = 410.7$, $E = 1.69$, $\alpha = 0.32$, and $\beta = 0.28$ are parameters estimated by the authors after fitting a regression model to a dataset of 400 language model training runs. In their paper, Hoffmann et al. (2022) present estimations for dataset size for many model sizes. With 70B parameters, Chinchilla requires 1.4T tokens according to these laws, which equates to roughly 20 tokens per parameter. Based on this average, we estimated an optimal dataset size for two models: 3.5 and 9.5 billion tokens for 160 and 460 million parameter models, respectively. We considered these to be fair sizes for this project, given that we would be able to train them without requiring much computing while, at the same time, the token count was still within a manageable range, i.e., something we could gather by relying on open-source datasets.[6]

3.2. Pre-training dataset

We consider English to be a high-resource language because, with datasets like the Pile (Gao et al., 2020), RedPajama (Together Computer, 2023), the ROOTS corpus (Laurençon et al., 2022), the Stack (Kocetkov et al., 2022), UltraChat Ding et al. (2023), etc., given that one can easily have access to trillions of high-quality, domain-specific tokens. Currently, most available tokens in Brazilian Portuguese come from datasets like BrWaC (Wagner Filho, Wilkens, Idiart, & Villavicencio, 2018), ClueWeb22 (Overwijk et al., 2022), Wikipedia (Wikimedia Foundation, 2024), OSCAR (Abadji, Suarez, Romary, & Sagot, 2022), and other byproducts of massive web crawling that require considerable filtering and pre-processing (Conneau et al., 2020; Ortiz Su'arez, Romary, & Sagot, 2020; Ortiz Su'arez, Sagot, & Romary, 2019; Wenzek, Lachaux, Conneau, Chaudhary, Guzmán, Joulin, & Grave, 2020).[7]

Studies like the ones performed by Xue, Fu, Zhou, Zheng, and You (2023) and Muennighoff et al. (2023) explore the challenges of training data-constrained LLMs, i.e., settings where the amount of data available is constrained. Both studies mainly focus on the downsides of repeating data during training runs, given that training language models with fresh data seem to have beneficial outcomes (Lee et al., 2021). According to Muennighoff et al. (2023), in data-constrained scenarios, training with up to 4 epochs of repeated data yields minor changes to loss compared to unique data. After this mark, increased repetition yields less performance, eventually decaying to zero. With this in mind, we aimed to build a dataset allowing training runs to be extended up and pass the optimal range without reaching the 4-epoch mark.

Hence, the first portion of our dataset comprises a concatenation of open-source Brazilian Portuguese datasets. These include: Wikipedia (Wikimedia Foundation, 2024), CulturaX (Nguyen et al., 2023), OSCAR (Abadji et al., 2022; Ortiz Su'arez et al., 2020; Ortiz Su'arez et al., 2019), Common Crawl (Conneau et al., 2020; Wenzek et al., 2020), and ROOTS (Laurençon et al., 2022) datasets. As a filtering step, we also utilized some of the filters used in Rae et al. (2021), besides using a fine-tuned BERTimbau (Souza et al., 2020) to exclude samples classified above a pre-defined toxicity threshold.[8] This first portion equates to 4.1 billion tokens of text. We call this first portion Pt-Corpus.

[5] Although out of the Brazilian Portuguese context, we could also mention models like Glória (Lopes, Magalhães, & Semedo, 2024), a European Portuguese decoder language model based on GPT-Neo (Black et al., 2022) and trained natively on 35B tokens of European Portuguese, and Gervásio PT, another Llama 2 byproduct of fine-tuning it.

[6] We also experimented with extended training and embedding transplant, like in other works (Domingues, 2023a; Gabriel Lino Garcia et al., 2024; Pires et al., 2023). In one of our initial explorations, we recycled the GPorTuguese-2 tokenizer and embedding layer, using them as a replacement for the original tokenizer and embedding weights (which are of the same dimension, i.e., 768) used by Pythia-160 m (Biderman et al., 2023). We then performed a test training run of 100,000 steps using the same hyperparameters and settings later described in our work. In our experiments, the loss curves exhibit significant variance, with sudden increases in loss, suggesting a lack of smooth convergence during the first quarter of training. This has led us further into favoring the idea of pre-training a model from scratch, which, in the end, allowed us to train our TTL pair smoothly from beginning to end.

[7] We did not include European Portuguese sources in the creation of our dataset. We argue that Brazilian Portuguese and European Portuguese are substantially different in aspects like vocabulary (e.g., the way words are written, the differences in common expressions) and grammar (e.g., the conjugation of verbs), which could hinder the performance of a monolingual text-generation model.

[8] This model is available on Hugging Face: https://huggingface.co/nicholasKluge/ToxicityModelPT.

Table 1
Tokenizer efficiency.

Model tokenizer	n° of tokens	Vocabulary size
TTL	**9937**	32,000
GPorTuguese-2	9959	50,257
BERTimbau	11,006	29,794
Cabrita-3B	11,488	52,000
Sabiá-7B	14,813	32,000

The second portion of our dataset was inspired by the many studies that show that models fine-tuned with demonstrations of instruction-following behavior perform better in many downstream tasks Askell et al. (2021), Bai et al. (2022), Chung et al. (2022), Ouyang et al. (2022), Shen et al. (2023), Touvron, Martin, et al. (2023), leading us to experiment with the inclusion of such type of data as part of a pre-training corpus. For this, we utilized the following datasets: Instruct-PTBR (Moro, 2024), Gpt4all-J (Moreira, 2024), Bactrian-X (Li, Koto, Wu, Aji, & Baldwin, 2023), Dolly 15 K (Conover et al., 2023), and CosmosQA (Huang, Le Bras, Bhagavatula, & Choi, 2019), many of which are translated versions of English native datasets. The concatenation of both portions is what we call Pt-Corpus-Instruct. 60% of this corpus is plain Brazilian Portuguese text (e.g., books, articles, blogs, etc.), while 40% demonstrate instruction-following behavior. Pt-Corpus-Instruct equates to approximately 6.2 billion tokens.

3.3. Tokenization

As pointed out by Cui, Yang, and Yao (2023) and Larcher et al. (2023), one of the obstacles related to adapting LLMs to new low-resource languages is the recycling of the tokenizer. For example, since Llama 2 was trained on a primarily English corpus, its tokenizer requires more tokens to encode non-English languages, shattering the information that could be already encoded into words or sub-units of words, as it does with much of the English vocabulary. However, one cannot merely exchange the tokenizer of a language model without some surgical adaptation of the embedding layer, as done by Guillou (2020), given that this exchange would break the learned mapping between tokens and embeddings.

Since the model architecture we adopt in this study is Llama 2, we trained a Sentencepiece tokenizer (Kudo & Richardson, 2018) to make our model compatible with the ever-growing Llama ecosystem. We trained our tokenizer on 2 million text samples from our dataset, with a vocabulary size of 32 K tokens. To test its efficiency, we performed the same test used by Larcher et al. (2023) to access their tokenizer, where we counted the number of tokens required to encode 7400 words. According to the results in Table 1, our tokenizer shows a 66% improvement in efficiency compared to the original Llama 2 tokenizer, allowing for a more efficient way to encode Brazilian Portuguese text.

We utilized our tokenizer to encode our dataset into sequences of 2048 tokens. The raw text datasets and their tokenized versions are available for download on HuggingFace.

3.4. Architecture

As done by Zhang et al. (2024), we used a decoder-only Transformer model (Vaswani et al., 2017) based on Llama 2 (Touvron, Martin, et al., 2023) as the basis for our models. The dimensions of our models are documented in Table 2.

Our models have all the implementations that the Llama 2 architecture benefits from, i.e., grouped query attention (Ainslie et al., 2023), root mean square layer normalization (Zhang & Sennrich, 2019), SwiGLU activation's (Shazeer, 2020), and RoPE embeddings (Su et al., 2021). While our 160 million parameter model uses 12 attention heads paired with 12 key–value heads, the 460 million parameter version uses 16 attention and key–value heads.

3.5. Training

We created all of our code implementations using the libraries tied to the Hugging Face ecosystem, i.e., Transformers (Wolf et al., 2020), Datasets (Lhoest et al., 2021), Tokenizers (HuggingFace, 2019), and Accelerate (Gugger et al., 2022), which allow for easy reproducibility, adaptation, and further scaling. Our training and evaluation scripts follow a standard PyTorch structure (Paszke et al., 2019), while we utilized CodeCarbon (CodeCarbon, 2019) and Weights & Biases (Weights&Biases, 2017) for tracking our experiments.

Regarding hardware, we were limited to using a single NVIDIA A100-SXM4-40 GB. To optimize its use, we performed several experiments to find the least costly training configuration regarding computing time and resource consumption, mainly regarding GPU memory utilization and tokens per second throughput. In these experiments, we explored the use of different mixed precision strategies and math modes (fp32, fp16, bf16, tf32), gradient accumulation steps, gradient checkpointing (Chen, Xu, Zhang, & Guestrin, 2016), the use of FlashAttention (Dao, 2023; Dao, Fu, Ermon, Rudra, & Ré, 2022), different types of optimizers (Dettmers, Lewis, Shleifer, & Zettlemoyer, 2022; Kingma & Ba, 2014; Loshchilov & Hutter, 2017; Shazeer & Stern, 2018), among other implementations choices. Ultimately, we arrived at the following training configurations, showcased in Table 3, which produced a throughput of up to 29,491 tokens per second during training and 3× that during inference on an Ampere GPU. The hyperparameters related to the optimizer and learning rate scheduler were based on the documentation of other open-source LLMs of similar size (Biderman et al., 2023; Workshop et al., 2022; Zhang et al., 2022).

This setting increased token throughput during training three times and in evaluation six times compared to using float32 precision, no tf32 mode, and a vanilla attention mechanism. The training of TTL-160 m took approximately 36 h (1.5 days), while the training of our 460 million parameter version took 280 h (11.5 days).[9]

During training, we saved several checkpoints for each model between an interval of 22,000 steps for TTL-160 and 25,000 for TTL-460 m, resulting in 20 and 48 intermediate checkpoints, respectively. All checkpoints were saved along with the current state of their optimizer and scheduler, allowing our models to resume training at any checkpoint desired or for others to use these checkpoints as a starting point for further training or fine-tuning. At the same time, we measured, for each checkpoint, their estimated energy consumption and carbon emissions, which we used to compare with our model evaluation scores. We evaluated our models every 100,000 steps with a sample size corresponding to approximately 1% of the training dataset. All models trained and checkpoints are available on HuggingFace.

4. Results

4.1. Evaluations

During our training runs, both models showed consistent convergence. At no point did our evaluation curves show signs of overfitting or saturation.[10] In the case of our 460 m parameter model, we intentionally trained past the optimal point by approximately 75,000 steps to assess if there were any signs of saturation, but our evaluations consistently gave better results. We hypothesize that our models are

[9] Even though not used in this study, given the limitation of the number of GPUs we were able to use, our code implementation supports all parallel features from Accelerate, like distributed training on multiple GPUs, multiple nodes, and plugins to other distributed training libraries, like DeepSpeed (Rajbhandari, Rasley, Ruwase, & He, 2020), FSDP (Ott et al., 2021), and Megatron-LM (Shoeybi et al., 2019).

[10] By saturation, we refer to the phenomenon where a model stops improving after a certain threshold of ingested tokens, probably due to the model size itself (Biderman et al., 2023).

Table 2
TTL model's architecture.

Size	Hidden size	Intermediate size	Context length	Heads	Layers	Vocab size
160M	768	3072	2048	12	12	32,000
460M	1024	4096	2048	16	24	32,000

Table 3
TTL model's training configuration.

TTL-160m	
Key	Value
tokens per batch	8192
total training steps	458,000
gradient accumulation steps	1
optimizer	AdamW
learning rate	$6.0 \times e^{-4}$
adam epsilon	$1.0 \times e^{-8}$
adam beta 1	0.9
adam beta 2	0.999
weight decay	0.01
scheduler type	cosine
warmup steps	5000
gradient checkpointing	False
mixed precision	bfloat16
tf32	True
flash attention 2	True

TTL-460m	
Key	Value
tokens per batch	8192
total training steps	1,200,000
gradient accumulation steps	2
optimizer	AdamW
learning rate	$3.0 \times e^{-4}$
adam epsilon	$1.0 \times e^{-8}$
adam beta 1	0.9
adam beta 2	0.999
weight decay	0.01
scheduler type	cosine
warmup steps	10,000
gradient checkpointing	False
mixed precision	bfloat16
tf32	True
flash attention 2	True

The full details are available in our GitHub repository.

Table 4
Performance on the Language Model Evaluation Harness (Gao et al., 2021).

	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Pythia-410m	24.83*	**41.29***	25.99*	40.95*	33.26
TTL-460m	**29.40**	33.00	**28.55**	41.10	33.01
Bloom-560m	24.74*	37.15*	24.22*	42.44*	32.13
Xglm-564M	25.56	34.64*	25.18*	**42.53**	31.97
OPT-350m	23.55*	36.73*	26.02*	40.83*	31.78
TTL-160m	26.15	29.29	28.11	41.12	31.16
Pythia-160m	24.06*	31.39*	24.86*	44.34*	31.16
OPT-125m	22.87*	31.47*	26.02*	42.87*	30.80
GPorTuguese-2	22.48	29.62	27.36	41.44	30.22
Gpt2-small	21.48*	31.60*	25.79*	40.65*	29.97
Multilingual GPT	23.81	26.37*	25.17*	39.62	28.73

All evaluations used the Language Model Evaluation Harness standard settings. Given our constrained budget, we could only evaluate some models in the Brazilian Portuguese version of the used benchmarks. Unfortunately, HellaSwag and the MMLU benchmarks require considerable time to run, even more so as the model size increases. Hence, results marked with an "*" were extracted from the Open LLM Leaderboard (Beeching et al., 2023), which uses the same evaluation method, using their English version. Thus, these results may vary if conducted in different languages, especially given that some models were pre-trained on mainly English text. Regardless, these results show that our models can perform as well and, in some instances (ARC and MMLU), surpass models trained on a much larger dataset and with many more resources. We speculate that this may result from our mixed dataset, which contains many demonstrations of instruction following and Q&A (i.e., the type of prompt usually used behind many of these benchmarks).

under-trained but can improve if further trained to pass the Chinchilla optimal range. As suggested by Touvron, Martin, et al. (2023) and Zhang et al. (2024), perhaps the scaling laws proposed Hoffmann et al. (2022) are indeed ill-suited to estimate the performance of language models outside of a closed budget (i.e., when one has access to ample amounts of tokens and compute). In Fig. 1, we present the learning curves of our TTL pair.

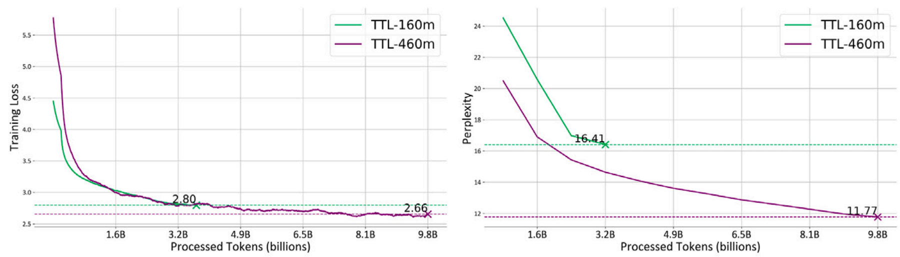
To test the performance of our models, we resort to evaluation benchmarks like the Language Model Evaluation Harness (Gao et al., 2021), which permit a common way to test language models on few-shot evaluations. We first utilized translated evaluation benchmarks, i.e., benchmarks constructed in English and later translated into Brazilian Portuguese, made available by the work of Lai, Ngo, Veyseh, Dernoncourt, and Nguyen (2023), which translated four benchmarks from the original EleutherAI evaluation harness to 29 languages in a commendable effort. These are:

- ARC-Challenge: a multiple-choice question-answering dataset containing questions from early grades science exams (Clark et al., 2018).
- HellaSwag: a multiple choice dataset for evaluating grounded commonsense inference (Zellers, Holtzman, Bisk, Farhadi, & Choi, 2019).
- MMLU: a benchmark that covers 57 subjects across STEM, humanities, social sciences, and more, measuring the performance of models on various natural language tasks (Hendrycks et al., 2020).
- TruthfulQA: a benchmark comprised of several questions, spanning 38 topics, that assess the model's tendency to replicate commonly believed falsehoods (Lin, Hilton, & Evans, 2021).

For comparison purposes, we evaluated models categorized as within the same size range as the TTL pair on these benchmarks. Our results are in Table 4.

We also evaluated our models on a natively Brazilian Portuguese implementation of the EleutherAI LM Evaluation Harness, made available by Garcia (2024). These evaluations seek to replicate, reproducibly, the Poeta benchmark (Pires et al., 2023), giving a standard evaluation procedure for Brazilian Portuguese language models on tasks involving the Brazilian context and its language. The evaluations we used for this harness are:

- ENEM: the Exame Nacional do Ensino Médio (ENEM) is an advanced High-School level exam widely applied annually by the Brazilian government to students wishing to undertake a University degree. This dataset contains 1430 questions that do not require image understanding of the exams from 2010 to 2018, 2022, and 2023 (Silveira & Mauá, 2017).
- BLUEX: a multimodal dataset consisting of Brazil's two leading university entrance exams, Convest (Unicamp) and Fuvest (USP), spanning from 2018 to 2024. The benchmark comprises 724 questions (Almeida, Laitz, Bonás, & Nogueira, 2023).
- OAB Exams: a dataset of more than 2000 questions from the Brazilian Bar Association's exams from 2010 to 2018 (Delfino, Cuconato, Haeusler, & Rademaker, 2017).
- ASSIN2 RTE: an Avaliação de Similaridade Semântica e Inferência Textual is the second edition of ASSIN, an evaluation that involves predicting if a given text (premise) entails (implies) another text (hypothesis) (Real, Fonseca, & Oliveira, 2020).



(a) (b)

Fig. 1. Learning Curves for the TTL pair.
Fig. 1: Plot (a) shows the training loss of our TTL pair over their runs. TTL-460 was trained on 75,000 steps over our estimations (based on Hoffmann et al. (2022). Both models show consistent convergence, and we speculate that they could be trained for longer with a significant increase in improvements, as demonstrated by Biderman et al. (2023), on the training of even smaller models using over 300B tokens. Plot (b) shows the perplexity scores of our TTL pair measured at every 100,000 steps (8.1 million tokens). We were surprised that, even when trained on a comparatively smaller dataset, our models achieved a perplexity score similar to the results shown by Radford et al. (2019) for models of similar size. We attribute this to the superiority of the Llama 2 architecture compared to GPT-2, which, given the constant advances in the field, comes with several improvements that make training large neural networks more efficient.

Table 5
Performance on the Portuguese Language Model Evaluation Harness (Garcia, 2024).

	ASSIN2 RTE	ASSIN2 STS	BLUEX	ENEM	FAQUAD NLI	HateBR	OAB Exams	Avg.
Qwen-1.8B	64.83	19.53	26.15	30.23	43.97	33.33	27.20	35.03
TinyLlama-1.1B	58.93	13.57	22.81	22.25	43.97	36.92	23.64	31.72
TTL-460m	53.93	12.66	22.81	19.87	49.01	33.59	27.06	31.27
XGLM-564m	49.61	22.91	19.61	19.38	43.97	33.99	23.42	30.41
Bloom-1b7	53.60	4.81	21.42	18.96	43.97	34.89	23.05	28.67
TTL-160m	53.36	2.58	21.84	18.75	43.97	36.88	22.60	28.56
OPT-125m	39.77	2.00	21.84	17.42	43.97	47.04	22.78	27.83
Pythia-160	33.33	12.81	16.13	16.66	50.36	41.09	22.82	27.60
OLMo-1b	34.12	9.28	18.92	20.29	43.97	41.33	22.96	27.26
TTL-460m-Chat	43.39	4.84	23.23	19.38	33.98	33.49	26.97	26.46
Bloom-560m	33.33	8.48	18.92	19.03	43.97	37.07	23.05	26.26
Pythia-410m	33.33	4.80	19.47	19.45	43.97	33.33	23.01	25.33
OPT-350m	33.33	3.65	20.72	17.35	44.71	33.33	23.01	25.15
GPT-2 small	33.26	0.00	10.43	11.20	43.52	33.68	13.12	20.74
GPorTuguese	33.33	3.85	14.74	3.01	28.81	33.33	21.23	19.75
Samba-1.1B	33.33	1.30	8.07	10.22	17.72	35.79	15.03	17.35
GlorIA-1.3B	0.0	2.32	3.2	1.89	0.26	0.28	5.19	8.70

According to the results, TTL-460m was able to outperform multilingual models 3x its sizes (like Bloom 17b) while being on par with models like Qwen (Bai et al., 2023) and OLMo (Groeneveld et al., 2024), which were initially trained on many more tokens while also being more extensive regarding parameter count. This may be an indication that, for example, (1) we used a similar token count (in terms of Portuguese tokens) for the training of our model as specific multilingual models, which directly correlates with how well such models go against these benchmarks, or that (2) at small parameter counts (i.e., small model sizes), the scores models archives on these benchmarks are pretty standard, independent of how many tokens they were trained. Furthermore, the performance of GlorIA-1.3B (Lopes et al., 2024), an LLM natively trained in European Portuguese, suggests that despite its extensive training on 30 billion tokens, its language modeling capabilities may not effectively translate to the tasks within our evaluation harness.

- ASSIN2 STS: similar to the ASSIN2 RTE, the Semantic Textual Similarity (STS) test measures the degree of semantic equivalence between two sentences (Real et al., 2020).
- FAQUAD NLI: a Portuguese reading comprehension dataset that follows the Stanford Question Answering Dataset (SQuAD) format. It consists of 900 questions about 249 reading passages taken from 18 official documents of a computer science college from a Brazilian federal university and 21 Wikipedia articles related to the Brazilian higher education system (Rodrigues, 2024).
- HateBR: the first large-scale expert annotated dataset of Brazilian Instagram comments for abusive language detection on the web and social media. It comprises 7000 documents annotated with a binary classification (offensive versus non-offensive comments) (Vargas, Carvalho, de Góes, Pardo, & Benevenuto, 2022).

Again, we evaluated models of the same size range as the TTL pair on these benchmarks for comparison purposes. Results are in Table 5.

To further evaluate the downstream capabilities of our models, we decided to employ a basic fine-tuning procedure of our TTL pair on a subset of tasks of the Poeta benchmark (Pires et al., 2023). Those are tasks involving toxicity detection (Vargas et al., 2022), textual entailment (Real et al., 2020; Rodrigues, 2023), sentiment analysis (Maas et al., 2011), and a text classification (Zhang, Zhao, & LeCun, 2015). We apply the same procedure for comparison purposes on both BERTimbau models (Souza et al., 2020), given that they are also LLM trained from scratch in Brazilian Portuguese and have a similar size range to our models.[11] Given their bidirectional nature, encoder-only transformers are usually superior in tasks like text classification, SQUAD-style Q&A, and named entity recognition. However, we argue that we can still use these comparisons, especially if they are made in a standardized fashion, to assess if our pre-training runs produced LLM capable of producing good results ("good" here means "close to BERTimbau") when utilized for downstream applications. In this round of evaluations, we fine-tuned all models considered using the same setting and compared their final performance in Table 6.

We also measured the efficiency of our model in terms of its throughput capabilities. Given that generative language models in real-time applications are critically tied to their throughput and memory footprint, we estimated how many tokens our models can generate

[11] We choose not to include GlorIA-1.3B (Lopes et al., 2024) in this analysis because (1) it provided poor results in our Portuguese evaluation benchmark, and (2) it is substantially larger than the models we trained.

Table 6
Downstream performance on different tasks.

Models	IMDB	FaQuAD-NLI	HateBr	Assin2	AgNews	Average
BERTimbau-large	**93.58**	92.26	91.57	**88.97**	94.11	92.10
BERTimbau-small	92.22	**93.07**	91.28	87.45	94.19	91.64
TTL-460m	91.64	91.18	**92.28**	86.43	**94.42**	91.19
TTL-160m	91.14	90.00	90.71	85.78	94.05	90.34

All the shown results are the higher accuracy scores achieved on the respective task test sets after fine-tuning the models on the training sets. All fine-tuning runs used the same hyperparameters: 3 epochs, batch size of 16, AdamW as the optimizer ($\alpha = 4\mathrm{e}^{-5}$, $\epsilon = 1\mathrm{e}^{-8}$), and a weight decay rate of 0.01. Even though bidirectional encoder-only models usually perform better on the types of tasks under consideration, our larger model (TTL-410) can, in this simple fine-tuning setting, outperform BERTimbau-large on tasks involving toxicity detection and general text classification. Given that all results from our models present an over 90% average accuracy score across tasks (even without pushing for optional hyperparameter settings), we argue that this shows the potential for our language models to be performative in many types of downstream tasks. All fine-tuned versions of TTL and their respective fine-tuning scripts are available on Hugging Face.

per second (t/s). According to our test, on average, TTL-460 m can generate up to $12_{t/s}$ on a Tesla V4 GPU. Applying a 4-bit activation-aware weight quantization (Lin et al., 2023) increases the throughput to $25_{t/s}$, and reduces the model's memory footprint to 340 MB. On an A100, we increase throughput to $60_{t/s}$, approximating 80 words generated per second. Further improvements can be achieved using inference frameworks that utilize more high-performance languages, like C or C++ (e.g., Llama.cpp).

Lastly, we licensed all artifacts created by this project (i.e., models, datasets, and code) under an Apache 2.0 License.

4.2. Energy consumption and carbon emissions

Given the consensus that tracking energy consumption, estimating carbon emissions, and reporting these results should be standard practice in the field of deep learning (Desislavov, Martínez-Plumed, & Hernández-Orallo, 2021; Falk & van Wynsberghe, 2023; García-Martín, Rodrigues, Riley, & Grahn, 2019; Lacoste, Luccioni, Schmidt, & Dandres, 2019; Lottick, Susai, Friedler, & Wilson, 2019; Luccioni, Viguier, & Ligozat, 2022; Strubell, Ganesh, & McCallum, 2019), we logged our energy consumption during training and evaluation runs by using CodeCarbon (CodeCarbon, 2019). Besides achieving a count for the total energy consumption of our project, this allowed us to monitor performance and energy consumption increases coupled during a training run of an LLM. Table 7 shows the logs for TTL-460 m. According to it, performance improvements diminish midway through our training run while energy consumption and emissions rates remain constant.

This observation underscores that nearly half of the energy consumed during our training runs corresponds to a marginal uptick in the model's performance. Meanwhile, this uptick becomes significantly tiny as the model approximates the optimal training point. Although it is unclear how long we can push training runs for smaller models, it is evident that the cost related to training large neural networks is directly proportional to model size, training time, and the hardware used.

According to the estimations proposed by Lottick et al. (2019), implemented in CodeCarbon, in total, the 36 h of compute time to train TTL-160 m consumed 15.5 kWh (≈ 5.7 KgCO2eq), while the 280 h used to train TTL-460 m consumed 113.0 kWh (≈ 41.3 KgCO2eq). Summing all up, these emissions equate to a 185-kilometer car ride.[12],[13]

[12] Calculations were made using the region of North Rhine-Westphalia (Germany) as the region of computing.

[13] GPU stats indicate that our training runs kept a steady allocation of GPU memory utilization (between 70%–85% of its maximum capacity), power usage (83%), and thermal output (≈ 60 °C).

Table 7
Energy consumption during training (TTL-460m).

Processed tokens	Perplexity	Energy consumption (kWh)	Emissions (KgCO2eq)
8.1M	20.49	9.40	3.34
1.6B	16.90	18.82	6.70
2.4B	15.43	28.59	10.16
3.2B	14.64	38.20	13.57
4.0B	14.08	48.04	17.07
4.9B	13.61	57.74	20.52
5.7B	13.25	67.32	23.92
6.5B	12.87	76.84	27.30
7.3B	12.57	86.40	30.70
8.1B	12.27	96.19	34.18
9.0B	11.96	106.06	37.70
9.8B	11.77	115.69	41.31

Here, we display how the perplexity score of our model diminishes at every 100,000 steps (8.1 million tokens) and what the energy consumption (kWh) and estimated carbon emissions (CO2eq) are related to this process. As one can see, after half of our training run, our performance increase slows down as the rate of consumption and emissions keeps following a linear trend. This shows that, on our training runs, almost half of the energy we consumed was tied to a marginal increase in the model's performance (≈ 1.84). By analyzing the training loss logs of other models (Black et al., 2022; Touvron, Martin, et al., 2023; Zhang et al., 2024), we argue that this is a common reality in the training of large-scale neural networks, i.e., *convergence is slow and costly*.

4.3. Alignment

With the release of ChatGPT in November 2022, there has been an increase in interest in models that went through an alignment process (e.g., instruction tuning, preference modeling, etc.), making them more attuned to follow the commands of people without the need for sophisticated prompting or further fine-tuning, becoming, in general, more helpful tools (a.k.a. assistants) to their users. Nowadays, there are many assistant models like ChatGPT (Conover et al., 2023; Corrêa, 2023a; Geng et al., 2023; Jiang et al., 2023; Köpf et al., 2023; Taori et al., 2023; Touvron, Martin, et al., 2023), which, besides being an object of interest to the general public, have become one of the most used laboratories for alignment research (Askell et al., 2021).

Our base models can follow instructions with minimal prompting, given that they were already exposed to millions of instructions during training. To further expand these capabilities, we fine-tuned the 460 m parameter version of TTL on an instructional dataset to create a chat version of our larger base model, TTL-460m-Chat. Like in the Alpaca study (Taori et al., 2023), we trained our chat model via SFT on a synthetically generated dataset. This dataset contains a collection of single-turn conversations between an assistant and a user, generated by prompting models that already went through an alignment process (ChatGPT, Vicuna, LLama 2, Open-Assistant, etc.). The dataset is available in Brazilian Portuguese and English and contains approximately 81 K samples (Corrêa, 2023b).[14]

For the SFT, we used the same software stack utilized to pre-train our models. TTL-460m-Chat was trained for three epochs using almost the same configurations documented in Table 3. The only modifications are in the number of warm-up steps (1000) and learning rate (1×10^{-5}). The full details are available in our GitHub repository.[15] We also created an open demo of our Chat model, which allows users to have conversations with TTL-460m-Chat.[16] In the demo, we also implemented a simple vector search engine that allows users to explore the fine-tuning dataset quickly, allowing them to evaluate the

[14] The Brazilian Portuguese version is a translated version of the English one, generated via the Google Translate API.

[15] The additional training TTL-460m-Chat equated to an energy consumption of 5.6 kWh and an emission of 2.5 KgCO2eq. The procedure took approximately 13.5 h.

[16] Available in https://huggingface.co/spaces/nicholasKluge/TeenyTinyLlama-Chat.

😃: Escreva uma história sobre uma menina chamada Sophia - a domadora de gnomos.

🤖: Sophia sempre foi fascinada por animais e desde pequena adorava explorar o mundo ao seu redor. Ela passava horas brincando com seus amigos peludos, explorando as florestas e rios próximos e até mesmo montando em jipes para passear. Mas à medida que ela crescia, sua paixão só se intensificava.

Um dia, enquanto caminhava pela floresta, Sophia tropeçou em um pequeno riacho. Ao dar um passo à frente, sentiu uma onda repentina de energia percorrendo suas veias. De repente, ela percebeu que estava sendo puxada pelo ar!

A princípio, Sophia ficou assustada. Ela nunca tinha visto nada parecido antes. Mas quando olhou mais de perto, viu algo incrível: os gnomos estavam voando, cada um deles girando e dançando no ar. E então, tão rapidamente quanto apareceu, eles desapareceram.

Sophia correu freneticamente pelas árvores, tentando encontrar o caminho de volta para casa. Mas não havia como voltar atrás. Os gnomos continuaram a voar e, eventualmente, chegaram a uma clareira onde Sophia finalmente encontrou o caminho de volta para casa.

Enquanto observava os gnomos desaparecerem, Sophia sabia que havia encontrado sua verdadeira paixão. Ela passou inúmeras noites estudando os animais, aprendendo tudo o que pôde sobre eles e os incríveis poderes que possuíam. E embora tenha sido difícil, ela também descobriu algo muito valioso – que às vezes as coisas mais bonitas da vida são aquelas que nos deixam sem fôlego e inspiradas.

Fig. 2. Sample generated by TTL-460m-Chat.
Fig. 2: Above is a user prompt requesting the model to generate a story; the English translation equates to: "*Write a story about a girl called Sophia — the gnome tamer.*" Model Output: "*Sophia has always been fascinated by animals, and from an early age, she loved exploring the world around her. She would spend hours playing with her furry friends, exploring the nearby forests and rivers, and even riding in jeeps for rides. But as she grew older, her passion only intensified. One day, while walking through the forest, Sophia stumbled across a small stream. As she stepped forward, she felt a sudden surge of energy coursing through her veins. Suddenly, she realized that she was being pulled through the air! At first, Sophia was frightened. She had never seen anything like it before. But when she looked closer, she saw something incredible: the gnomes were flying, each spinning and dancing in the air. And then, as quickly as they had appeared, they disappeared. Sophia ran frantically through the trees, trying to find her way back home. But there was no turning back. The gnomes continued to fly and eventually reached a clearing, where Sophia finally found her way home. As she watched the gnomes disappear, Sophia knew she had found her true passion. She spent countless nights studying the animals, learning everything about them and their incredible powers. And although it was difficult, she also discovered something precious — that sometimes the most beautiful things in life leave you breathless and inspired.*".

model's capabilities regarding "how much the model can go beyond its fine-tuning distribution." We also made available a 4-bit quantized version of this model for faster inference with almost no loss in performance. Fig. 2 shows a sample of TTL-460m-Chat capabilities in story generation.

5. Limitations

Our work documents the process of developing text-generation language models for low-resource languages while being under a condition of low resources. With a budget of 500 USD, there is much that we could not perform, like multi-GPU distributed training with larger batch sizes and prolonged training runs. Even though our results indicate that we can achieve reasonable performance by limiting a training run to the stipulations made by specific scaling laws (Hoffmann et al., 2022), regardless, that evidence points out that our models are still undertrained. However, aware of the limitations of our work, we logged and documented every step in a reproducible manner, allowing others to push our work beyond what we currently can or begin new projects from our source.

Other limitations of this work are related to the TTL pair. Like almost all other language models trained on large text datasets scraped from the web, the TTL pair exhibited behavior that does not make them an out-of-the-box solution to many real-world applications, especially those requiring factual, reliable, nontoxic text generation. Our models are all subject to producing hallucinations (i.e., the generation of text that is incorrect, nonsensical, or not real),[17] reproducing historical biases or generating toxic language, being overly verbose or repetitive, and generally producing content that should not be taken as given without human moderation.[18] Hence, even though our models are released with a permissive license, we urge users to perform their risk analysis on these models if intending to use them for real-world applications and also have humans moderating the outputs of these models in applications where they will interact with an audience, guaranteeing users are always aware they are interacting with a language model.

Additionally, our evaluations observed that the SFT-aligned version of the TTL-460 performed worse than the unaligned version on the Portuguese language model evaluation harness. This discrepancy suggests that alignment may necessitate native design and contextual considerations. In other words, simply translating an English instruction dataset to language X may not universally solve the challenge, highlighting the importance of tailored approaches for different languages and contexts.

6. Future works

The utility of language models, coupled with their capabilities remaining largely under-explored for most languages, presents numerous research opportunities. In this endeavor, we aimed to document and highlight the inherent challenges in harnessing resources for such a project while sharing our methods and tools with the broader community.

Regarding the TTL pair and the development of LLMs on a smaller scale, there is ample room for exploration regarding their utility and application. Models tailored for low-resource scenarios hold potential for various uses, including edge computing, game development, real-time applications, and more. Additionally, such models serve as vital

[17] However, it is worth noting that such behavior may well be irreparable features of LLMs (Xu, Jain, & Kankanhalli, 2024).

[18] A limitation of our models is how they generate code. Given that much of our training came from translated conversations on subjects like coding, our model tends to create code with programmatic commands that are natively written in English (`import torch`) on Brazilian Portuguese (`importar torch`), given the unstructured way in which people translated these samples.

resources for research endeavors. The availability of models trained in the native languages of non-English speakers expands access to a broader population interested in research involving LLMs, whether in the realms of NLP, AI ethics, AI alignment, or related fields.

With this in mind, here are some possible avenues for future projects seeking to expand our work:

1. **Scaling to the billion-parameter range.** Accelerate permits a simple way to scale training on multiple GPUs, and at the billion-parameter range, we are bound to encounter the emergence of improved capabilities (Gunasekar et al., 2023). Given that our datasets, according to scaling laws for data-constrained models (Muennighoff et al., 2023), are sufficient to train models up to that range, future training runs on multiple GPUs might give us the first billion parameter text generation models trained on Brazilian Portuguese text.
2. **Scaling dataset size to the 1T tokens mark.** To our knowledge, there are still no trillion-sized datasets for Brazilian Portuguese. Collecting text datasets to create such a corpus would enable us to push the training of models to pass the Chinchilla scaling laws (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023), explore the saturation and training limits of smaller language models (Zhang et al., 2024).
3. **Expand the open-source development of language models for low-resource languages.** Our experiments could be reproduced for other languages if they can access a minimum amount of tokens. At the same time, open-source development should be encouraged (Dey et al., 2023; Geng & Liu, 2023; Köpf et al., 2023), so in the future, the whole concept of "*low-resource language*" will be a thing of the past.

7. Conclusion

Large language models spearhead a paradigm shift in natural language processing, while widespread applications keep catalyzing these technologies to push the field's boundaries further. However, it is crucial to acknowledge that the advancements in LLMs have not been universally distributed across all languages and remain unevenly accessible, underscoring the ongoing challenge of achieving linguistic inclusivity openly and equitably.

In this study, our primary objective was to meticulously document the obstacles encountered and insights gained while training language models for low-resource languages while navigating the constraints imposed by low-resource settings. As a result, we successfully crafted a pair of language models, the *TeenyTinyLlama* pair, trained in Brazilian Portuguese to the optimal range defined by the Chinchilla scaling laws. Remarkably, our findings indicate that these models exhibit comparable performance to other multilingual and monolingual language models of similar size in various linguistic tasks while still demonstrating signs of possible undertraining.

All models, datasets, and source code developed in this study have been released under a permissive license, fostering open access and encouraging collaborative research within the academic community.

CRediT authorship contribution statement

Nicholas Kluge Corrêa: Project's idealization, Software stack's implementation, Dataset creation, Training, Evaluation of the models, Writing the article, Documenting the repositories. **Sophia Falk:** Implementing the carbon tracking methodology, Monitoring training runs, Writing the article. **Shiza Fatimah:** Developing the datasets, Including deduplication and cleaning, Writing the article, Documenting the repositories. **Aniket Sen:** Optimization of the software stack, Training, Evaluation of the models, Article's writing. **Nythamar De Oliveira:** Coordinated this project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is available at the following URL: https://huggingface.co/collections/nicholasKluge/teenytinyllama-6582ea8129e72d1ea4d384f1.

References

Abadji, J., Suarez, P. O., Romary, L., & Sagot, B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. arXiv preprint arXiv:2201.06642.

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245.

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. arXiv preprint arXiv:2204.06487.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., et al. (2023). The falcon series of open language models. arXiv preprint arXiv:2311.16867.

Almeida, T. S., Laitz, T., Bonás, G. K., & Nogueira, R. (2023). BLUEX: A benchmark based on Brazilian leading universities entrance exams. arXiv:2307.05410.

Antoun, W., Baly, F., & Hajj, H. (2021). AraGPT2: Pre-trained transformer for arabic language generation. In N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, & S. Touileb (Eds.), *Proceedings of the sixth arabic natural language processing workshop* (pp. 196–207). Kyiv, Ukraine (Virtual): Association for Computational Linguistics.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al. (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al. (2023). Qwen technical report. arXiv preprint arXiv:2309.16609.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., et al. (2023). Open LLM leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., et al. (2024). DeepSeek LLM: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning* (pp. 2397–2430). PMLR.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745.

Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. (2022). Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., et al. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457.

CodeCarbon (2019). CodeCarbon: Track emissions from compute and recommend ways to reduce their impact on the environment. https://github.com/mlco2/codecarbon.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., et al. (2023). Free dolly: Introducing the world's first truly open instruction-tuned LLM. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Corrêa, N. K. (2023a). Aira.

Corrêa, N. K. (2023b). Instruct-aira dataset version 2.0. https://huggingface.co/datasets/nicholasKluge/instruct-aira-dataset-v2.

Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177.

Dao, T. (2023). Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691.

Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, *35*, 16344–16359.

de Camaret, N. (2024). Cabra. https://huggingface.co/nicolasdec/Cabra.

Delfino, P., Cuconato, B., Haeusler, E. H., & Rademaker, A. (2017). Passing the Brazilian OAB exam: data preparation and some experiments. arXiv preprint arXiv:1712.05128.

Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2021). Compute and energy consumption trends in deep learning inference. arXiv preprint arXiv:2109.05472.

Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-Bit optimizers via block-wise quantization. In *9th International conference on learning representations*. ICLR.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dey, N., Gosal, G., Khachane, H., Marshall, W., Pathria, R., Tom, M., et al. (2023). Cerebras-GPT: Open compute-optimal language models trained on the cerebras wafer-scale cluster. arXiv preprint arXiv:2304.03208.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., et al. (2023). Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:2305.14233.

Domingues, M. (2023a). Canarim-7b (revision 08fdd2b). https://huggingface.co/dominguesm/canarim-7b.

Domingues, M. (2023b). Canarim-instruct-PTBR-dataset (revision c2de751). https://huggingface.co/datasets/dominguesm/Canarim-Instruct-PTBR-Dataset.

Eisenschlos, J. M., Ruder, S., Czapla, P., Kardas, M., Gugger, S., & Howard, J. (2019). MultiFiT: Efficient multi-lingual language model fine-tuning. arXiv preprint arXiv:1909.04761.

Falk, S., & van Wynsberghe, A. (2023). Challenging AI for sustainability: what ought it mean? *AI and Ethics*, 1–11.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., et al. (2021). A framework for few-shot language model evaluation. Version v0. 0.1. Sept.

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

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., et al. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based task. arXiv:2401.02909.

García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, *134*, 75–88.

Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., et al. (2023). Koala: A dialogue model for academic research. URL https://bair.berkeley.edu/blog/2023/04/03/koala/.

Geng, X., & Liu, H. (2023). OpenLLaMA: An open reproduction of llama. https://github.com/openlm-research/open_llama.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., et al. (2024). Olmo: Accelerating the science of language models. arXiv preprint arXiv:2402.00838.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., et al. (2022). Accelerate: Training and inference at scale made simple, efficient and adaptable.. https://github.com/huggingface/accelerate.

Guillou, P. (2020). GPorTuguese-2 (portuguese GPT-2 small): a language model for portuguese text generation (and more NLP tasks...). https://huggingface.co/pierreguillou/gpt2-small-portuguese.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., et al. (2023). Textbooks are all you need. arXiv preprint arXiv:2306.11644.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., et al. (2021). Maria: Spanish language models. arXiv preprint arXiv:2107.07253.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., et al. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Henrique, B. (2024a). Caramelo 7B. https://huggingface.co/Bruno/Caramelo_7B.

Henrique, B. (2024b). Harpia-7b-guanacolora. https://huggingface.co/Bruno/Harpia-7b-guanacoLora.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Huang, L., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* EMNLP-IJCNLP, (pp. 2391–2401). Hong Kong, China: Association for Computational Linguistics, URL https://www.aclweb.org/anthology/D19-1243.

HuggingFace (2019). Tokenizers: Fast state-of-the-art tokenizers optimized for research and production.. https://github.com/huggingface/tokenizers.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., et al. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. arXiv preprint arXiv:2207.07048.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ko, H., Yang, K., Ryu, M., Choi, T., Yang, S., Park, S., et al. (2023). A technical report for polyglot-ko: Open-source large-scale Korean language models. arXiv preprint arXiv:2306.02254.

Kocetkov, D., Li, R., Allal, L. B., Li, J., Mou, C., Ferrandis, C. M., et al. (2022). The stack: 3 tb of permissively licensed source code. arXiv preprint arXiv:2211.15533.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., et al. (2023). OpenAssistant conversations–democratizing large language model alignment. arXiv preprint arXiv:2304.07327.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

Lai, V., Ngo, N. T., Veyseh, A. P. B., Dernoncourt, F., & Nguyen, T. H. (2023). Open multilingual LLM evaluation leaderboard. https://huggingface.co/spaces/uonlp/open_multilingual_llm_leaderboard.

Lankford, S., Afli, H., & Way, A. (2023). adaptMLLM: Fine-tuning multilingual language models on low-resource languages with integrated LLM playgrounds. *Information*, *14*(12), 638.

Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., & Caridá, V. (2023). Cabrita: closing the gap for foreign languages. arXiv preprint arXiv:2308.11878.

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., et al. (2022). The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, *35*, 31809–31826.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., et al. (2021). Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., et al. (2021). Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations* (pp. 175–184). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, arXiv:2109.02846.

Li, H., Koto, F., Wu, M., Aji, A. F., & Baldwin, T. (2023). Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. arXiv:2305.15011.

Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., et al. (2021). Few-shot learning with multilingual language models. arXiv preprint arXiv:2112.10668.

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. arXiv preprint arXiv:2306.00978.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Lopes, R., Magalhães, J., & Semedo, D. (2024). Gl\'oria-a generative and open large language model for portuguese. arXiv preprint arXiv:2402.12969.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lottick, K., Susai, S., Friedler, S. A., & Wilson, J. P. (2019). Energy usage reports: Environmental awareness as part of algorithmic accountability. arXiv preprint arXiv:1911.08354.

Luccioni, A., Viguier, S., & Ligozat, A. (2022). Estimating the carbon footprint of BLOOM, a 176b parameter language model. http://dx.doi.org/10.48550, arXiv preprint arXiv:2211.02001.

Luo, Y., Kong, Q., Xu, N., Cao, J., Hao, B., Qu, B., et al. (2023). YAYI 2: Multilingual open-source large language models. arXiv preprint arXiv:2312.14862.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., et al. (2020). CamemBERT: a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7203–7219). Online: Association for Computational Linguistics.

Moreira, P. F. (2024). Gpt4all-J prompt generations PT. https://huggingface.co/datasets/pablo-moreira/gpt4all-j-prompt-generations-pt.

Moro, C. (2024). Instruct-PTBR-enUS-11m. https://huggingface.co/datasets/cnmoro/Instruct-PTBR-ENUS-11M.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., et al. (2023). Scaling data-constrained language models. arXiv preprint arXiv:2305.16264.

Nagoudi, E. M. B., Chen, W.-R., Abdul-Mageed, M., & Cavusogl, H. (2021). IndT5: a text-to-text transformer for 10 indigenous languages. arXiv preprint arXiv:2104.07483.

Nguyen, X.-P., Aljunied, S. M., Joty, S., & Bing, L. (2023). Democratizing LLMs for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. arXiv preprint arXiv:2306.11372.

Nguyen, T., Nguyen, C. V., Lai, V. D., Man, H., Ngo, N. T., Dernoncourt, F., et al. (2023). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. arXiv:2309.09400.

NousResearch (2024). Nous-hermes-2-Yi-34B. URL https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B/tree/main.

Ortiz Su'arez, P. J., Romary, L., & Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1703–1714). Online: Association for Computational Linguistics, URL https://www.aclweb.org/anthology/2020.acl-main.156.

Ortiz Su'arez, P. J., Sagot, B., & Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. L"ungen, & C. Iliadi (Eds.), *Proceedings of the workshop on challenges in the management of large corpora (CMLC-7) 2019. cardiff, 22nd July 2019* (pp. 9–16). Mannheim: Leibniz-Institut f"ur Deutsche Sprache.

Ott, M., Shleifer, S., Xu, M., Goyal, P., Duval, Q., & Caggiano, V. (2021). Fully sharded data parallel: faster ai training with fewer gpus.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Overwijk, A., Xiong, C., & Callan, J. (2022). ClueWeb22: 10 billion web documents with rich information. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 3360–3362).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., et al. (2023). RWKV: Reinventing RNNs for the transformer era. arXiv preprint arXiv:2305.13048.

Pires, R., Abonizio, H., Rogério, T., & Nogueira, R. (2023). Sabi\'a: Portuguese large language models. arXiv preprint arXiv:2304.07880.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910.10683,

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis* (pp. 1–16). IEEE.

Real, L., Fonseca, E., & Oliveira, H. G. (2020). The assin 2 shared task: a quick overview. In *International conference on computational processing of the portuguese language* (pp. 406–412). Springer.

Rodrigues, R. C. (2023). FaQuAD-NLI: a benchmark for textual entailment. https://huggingface.co/datasets/ruanchaves/faquad-nli.

Rodrigues, R. C. (2024). Faquad-NLI. https://huggingface.co/datasets/ruanchaves/faquad-nli.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., et al. (2023). Advancing neural encoding of portuguese with transformer albertina PT. arXiv preprint arXiv:2305.06721.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., et al. (2023). Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950.

Sales Almeida, T., Abonizio, H., Nogueira, R., & Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. arXiv e-prints, arXiv–2403.

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). Gottbert: a pure german language model. arXiv preprint arXiv:2012.02110.

Shazeer, N. (2020). Glu variants improve transformer. arXiv preprint arXiv:2002.05202.

Shazeer, N., & Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In *International conference on machine learning* (pp. 4596–4604). PMLR.

Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., et al. (2023). Mixture-of-experts meets instruction tuning: A winning combination for large language models. arXiv preprint arXiv:2305.14705.

Shliazhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., & Shavrina, T. (2022). Mgpt: Few-shot learners go multilingual. arXiv preprint arXiv:2204.07580.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.

Silveira, I. C., & Mauá, D. D. (2017). University entrance exam as a guiding test for artificial intelligence. In *BRACIS, Proceedings of the 6th Brazilian conference on intelligent systems* (pp. 426–431).

Souza, L. (2024). Samba. https://huggingface.co/lrds-code/samba-1.1B.

Souza, F., Nogueira, R., & Lotufo, R. (2020). Bertimbau: pretrained BERT models for Brazilian portuguese. In *Intelligent systems: 9th Brazilian conference, BRACIS 2020, rio grande, Brazil, October 20–23, 2020, proceedings, part i 9* (pp. 403–417). Springer.

Stability AI Language Team (2024). Stable LM 2 1.6b.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). Roformer: Enhanced transformer with rotary position embedding. CoRR abs/2104.09864 (2021). arXiv preprint arXiv:2104.09864.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., et al. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*, *3*(6), 7, https://crfm.stanford.edu/2023/03/13/alpaca.html.

Together Computer (2023). RedPajama: an open dataset for training large language models. https://github.com/togethercomputer/RedPajama-Data.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Vargas, F., Carvalho, I., de Góes, F. R., Pardo, T., & Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 7174–7183).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation*.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Weights&Biases (2017). Weights & biases: A tool for visualizing and tracking your machine learning experiments. https://github.com/wandb/wandb.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., et al. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4003–4012). Marseille, France: European Language Resources Association, ISBN: 979-10-95546-34-4.

Wikimedia Foundation (2024). Wikimedia Downloads. https://dumps.wikimedia.org.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., et al. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

Xue, F., Fu, Y., Zhou, W., Zheng, Z., & You, Y. (2023). To repeat or not to repeat: Insights from scaling LLM under token-crisis. arXiv preprint arXiv:2305.13230.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. (2022). Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems*, *32*.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385.

Zhang, X., Zhao, J. J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *NIPS*.

Zhao, J., Zhang, Z., Zhang, Q., Gui, T., & Huang, X. (2024). LLaMA beyond english: An empirical study on language capability transfer. arXiv preprint arXiv:2401.01055.

Nicholas Kluge Corrêa is a postdoc researcher at the Center for Science and Thought at the University of Bonn (Bonn, NRW, Germany) and a fellow from the RAIES (Rede de Inteligência Artificial Ética e Segura) network (Porto Alegre, RS, Brazil). His contact email is kluge@uni-bonn.de.

Sophia Falk is a Ph.D. researcher at the Bonn Sustainable AI Lab, Institute for Science and Ethics, University of Bonn (Bonn, NRW, Germany). Her contact email is falk@iwe.uni-bonn.de.

Shiza Fatimah is a master's student at the Institute for Computer Science, University of Bonn (Bonn, NRW, Germany). Her contact email is s39sfati@uni-bonn.de.

Aniket Sen is a Ph.D. researcher at the Helmholtz-Institut für Strahlen und Kernphysik, University of Bonn, and the Bethe Center for Theoretical Physics, University of Bonn (Bonn, NRW, Germany). His contact email is sen@hiskp.uni-bonn.de.

Prof. Dr. Nythamar De Oliveira is a full professor at the Pontifical Catholic University of Rio Grande do Sul and the coordinator of the RAIES network (Porto Alegre, RS, Brazil). His contact email is nythamar.oliveira@pucrs.br.