GlórIA: A Generative and Open Large Language Model for Portuguese

Ricardo Lopes, João Magalhães, David Semedo
NOVA LINCS, NOVA School of Science and Technology, Portugal
rv.lopes@campus.fct.unl.pt
{jmag, df.semedo}@fct.unl.pt

Abstract

Significant strides have been made in natural language tasks, largely attributed to the emergence of powerful large language models (LLMs). These models, pre-trained on extensive and diverse corpora, have become increasingly capable of comprehending the intricacies of language. Despite the abundance of LLMs for many high-resource languages, the availability of such models remains limited for European Portuguese. We introduce GlórIA, a robust European Portuguese decoder LLM. To pre-train GlórIA, we assembled a comprehensive PT-PT text corpus comprising 35 billion tokens from various sources. We present our pre-training methodology, followed by an assessment of the model's effectiveness on multiple downstream tasks. Additionally, to evaluate our models' language modeling capabilities, we introduce CALAME-PT (Context-Aware LAnguage Modeling Evaluation for Portuguese), the first Portuguese zero-shot language-modeling benchmark. Evaluation shows that GlórIA significantly outperforms existing open PT decoder models in language modeling and that it can generate sound, knowledge-rich, and coherent PT-PT text. The model also exhibits strong potential for various downstream tasks.[1]

1 Introduction

The emergence of robust large language models (LLMs) has led to a significant step forward across the whole natural language processing (NLP) field spectrum, with remarkable advances in a myriad of tasks, all of this with minimal supervision. Among the key ingredients to obtain such LLMs and enable effective modeling of language intricacies, we have 1) rich, highly diverse, and broad pre-training corpora accompanied by task-specific benchmarks to assess model capabilities in multiple down-stream tasks (Wang et al., 2018; Paperno et al., 2016); 2) high-capacity deep Transformer decoder architectures (Vaswani et al., 2017; Workshop et al., 2023), and 3) state-of-the-art pre-training methodologies, to ensure stable convergence (Biderman et al., 2023; Workshop et al., 2023).

While such core language model learning ingredients have been thoroughly investigated and matured for English and other high-resource languages, the European Portuguese language is lagging behind. In fact, there is a shortage of PT resources for pre-training and downstream task benchmarking, which is further aggravated when dialing down to European Portuguese (PT-PT). Additionally, it is critical to understand how well-established LLM learning methodologies, from data preparation and selection to training methodologies, generalize and ensure convergence on PT-PT corpora. Despite these limitations and challenges, there have been promising advances, with recent PT encoder models (Rodrigues et al., 2023; Souza et al., 2020) addressing many discriminative tasks with great success. However, many challenges remain open in Portuguese LLMs, in particular, in tasks that require language generation capabilities, in zero and few-shot settings, on a wide range of domains.

With these models and resource gaps in mind, we propose a new European Portuguese large decoder model, **GlórIA**, trained on a diverse corpora comprising 35 billion tokens from a myriad of domains, including generic web content, news pieces, encyclopedic knowledge and dialog data. Furthermore, to evaluate the language modeling capabilities of GlórIA, we introduce CALAME-PT, a novel zero-shot PT benchmark for language modeling evaluation. In our experiments, we show that GlórIA consistently and significantly outperforms existing PT open language models in language modeling.

[1] For the source code, pre-trained models and data resources, refer to https://github.com/rvlopes/GlorIA.

2 Related Work

Generative LLMs have widely sparked the interest of the NLP community. All the way from GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), new *GPT-like* models have demonstrated impressive flexibility in addressing NLP tasks such as reading comprehension, question answering, among others, in zero and few-shot settings. All these models adopt billion-scale parameter decoder-only Transformer architectures (Vaswani et al., 2017; Radford et al., 2019), ranging from 1.3B up to 175B parameters. While each model uses its own pre-training corpora (some of which are not disclosed), the majority of the texts are in English. It is particularly interesting the case of the LLaMA (Touvron et al., 2023a,b) family of models that are open and were trained with publicly available datasets, thus contributing to reproducibility.

2.1 Moving towards PT LLMs

With the goal of generalizing language knowledge, some initial multilingual models such as mBERT (Devlin et al., 2019), mT5 (Xue et al., 2021) and mGPT (Shliazhko et al., 2022) were contributed. Nevertheless, it has been shown that single-language LLMs outperform multilingual ones (Martin et al., 2020; Virtanen et al., 2019). Souza et al. (2020) proposed BERTimbau, the first Brazilian Portuguese (PT-BR) encoder with that in mind. Moving towards larger encoder-decoder models, Carmo et al. (2020) proposed the PTT5 model, based on the T5 architecture. It was then fine-tuned for paraphrasing tasks (Schneider et al., 2021), and for Portuguese question-generation (Leite and Lopes Cardoso, 2022).

However, most of these are not exclusive to a specific variant of Portuguese, or lack generative capabilities, or are just fine-tuned to a specific downstream task. Focusing only on the PT-PT variant, Rodrigues et al. (2023) proposed Albertina, a 900M parameter DeBERTa encoder with both PT-PT and PT-BR versions, trained on different corpora due to language differences. The authors demonstrated that the PT-PT model outperformed its PT-BR counterpart on PT-PT tasks. The same authors also released the Gervásio-PTPT LLM, a 1B decoder available in HuggingFace[2]. Recently, Sabiá (Pires et al., 2023), a 65B PT-BR LLM based on LLaMA was proposed, showing promising results on PT-BR few-show settings.

2.2 Large-Scale PT Text Data

Previously mentioned models leveraged large-scale unlabeled text data. Major efforts have been made to produce massive collections of text for heavily researched languages like English. For Portuguese, there have been some promising advances. BERTimbau used brWac (Wagner Filho et al., 2018), a 2.7B token dataset obtained from crawling PT-BR websites, while Albertina used a PT-PT filtered version of OSCAR, together with PT-PT transcripts datasets from the Portuguese and the EU Parliaments (Hajlaoui et al., 2014; Koehn, 2005). Sabiá uses the Portuguese subset of ClueWeb22 (Overwijk et al., 2022). Leveraging filtered massive web crawls such as ClueWeb22 (Overwijk et al., 2022) and OSCAR (Abadji et al., 2022), the Portuguese web-archive (Gomes et al., 2008) (Arquivo.pt), encyclopedic and dialog data, we assemble and contribute with a large and highly-diverse pre-training PT-PT corpus.

3 Preparing a new Large PT-PT Corpus

As evidenced by previous work, a large and diverse collection of texts, spanning over multiple domains, allows the model to better understand the language (and its intrincacies), thus improving the quality of the generated text (Radford et al., 2019; Touvron et al., 2023a; Brown et al., 2020). Given that the availability of European Portuguese texts at scale is not on par with English, our first objective is to further advance the diversification and availability of PT-PT resources, by gathering a large and rich collection of datasets.

3.1 PT Language Sources

To gather high-quality, large-scale, PT language resources, we resorted to multiple PT-PT text sources, summarized in Table 1. **OSCAR-2201 (Abadji et al., 2022) and ClueWeb-L 22 (Overwijk et al., 2022)** are web crawls – they both give us text from blogs, forums, among other websites. The **PTWiki**[3] provides our model with well-written and reviewed encyclopedic knowledge, in neutral and revised Portuguese text. **Europarl**[4] (Koehn, 2005) provides transcripts from diverse sessions that occurred in the European Parliament (such

[2] `https://huggingface.co/PORTULAN` - model name: `gervasio-ptpt-base`.

[3] `https://dumps.wikimedia.org/`

[4] `https://www.statmt.org/europarl/`

Table 1: Collected datasets and post-processing statistics.

Dataset	Domain	Documents	Tokens
ClueWeb22 PTPT Subset	Web Crawl	29M	31.6B
OSCAR PTPT	Web Crawl	1.5M	1.8B
ArquivoPT	News and periodicals	1.5M	0.8B
OpenSubtitles PTPT	Subtitles from movies	1.2M	1.0B
PTWiki	Encyclopedia	0.8M	0.2B
EuroParl PTPT	European Parliament Dialogs	1.3M	0.05B
	Total	35.3M	35.5B

as colloquial conversations between Eurodeputies). **OpenSubtitles** (Lison and Tiedemann, 2016) is comprised of essentially small and short movie conversations and narrations. Finally, our **Arquivo.pt subset** is a collection of scrapped text from periodicals and news websites archived by Arquivo.pt (Gomes et al., 2008), providing the model with high-quality reviewed news texts.

3.2 Data Processing

Once the individual datasets were gathered, they were filtered and processed. PT-PT documents were filtered using metadata when available (documents whose URL contains ".pt" in its domain), removing documents with low word count (<=15), fixing *mojibakes* and other encoding errors, removing remnant HTML tags, and removing exact duplicates through hashing. To avoid having the model learn "first-person" toxicity biases and insults, an extra processing step was applied to OpenSubtitles to discard samples based on the existence of profanity words, where a manually produced list of Portuguese bad words was used to explicitly filter out samples that contained them. We believed this had to be done specifically for OpenSubtitles due to its dialog nature - we wanted to avoid having the model learn "first-person" toxicity biases

After processing, our pre-training corpus reached a total of 35.3M documents and 35.5B tokens – Table 1 shows the detailed statistics.

4 The GlórIA Model

GlórIA is a decoder-based LLM with an architecture similar to GPT-3's (Brown et al., 2020), competing with it in linguistic, physical, and scientific reasoning tasks. Specifically, it adopts the GPT-Neo (Black et al., 2021)'s 1.3B and 2.7B architectures, following the HuggingFace's implementation of the model. Being a decoder, GlórIA uses a Causal LM pre-training objective, using cross-entropy as its loss. Table 2 shows the architecture configuration for GlórIA's both versions. GPT-Neo also employs local attention (Beltagy et al., 2020), which replaces standard self-attention and combines a dilating sliding window strategy with pre-selected global attention on some input locations, making the self-attention scale linearly, and linear attention (Zhuoran et al., 2021), which optimizes the dot-products by providing linear memory and processing complexities while maintaining representational capability.

Table 2: GlórIA architecture configurations. l denotes the number of layers, #AH the number of attention heads, and h denotes the model hidden layer size.

Model	#Params.	l	#AH	h
GlórIA 1.3B	1.3B	24	16	2048
GlórIA 2.7B	2.7B	32	20	2560

4.1 Pre-training details

To pre-train GlórIA, a total batch size of 512 was used (128 p/ GPU), with 16 gradient accumulation steps. We prepared a GPT-2-like BPE tokenizer, with a vocabulary size of 50257 tokens. Training was performed with BF-16 mixed-precision and a weight decay of 0.01. For the 1.3B version, GlórIA was trained for a total of 3M steps, on 4x NVIDIA A100s 40GB, for a total of 21 days (7 days p/ 1M steps), while, for the 2.7B, due to hardware resource constraints, we trained it only for 1M steps on 7x NVIDIA A100s (10 days). A cosine annealing scheduler was used for both models, with hard restarts every 500k steps and 10k warmup steps. Periodic evaluations and data shuffling were conducted every 1 million steps.

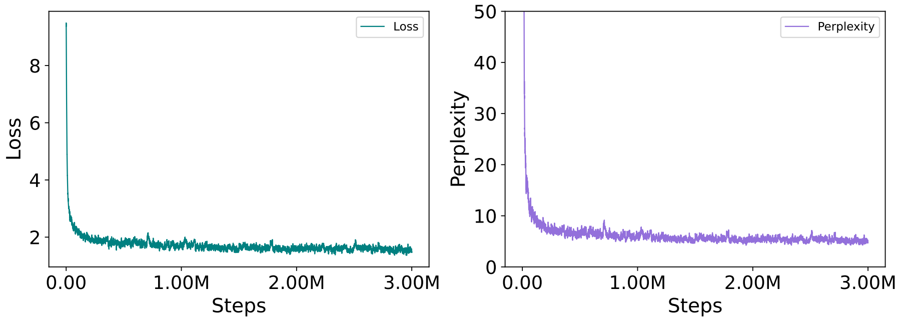

Figure 1: GlórIA 1.3B pre-training loss and perplexity.

Table 3: Documents seen per each dataset during GlórIA 1.3B's pre-training - a total of 96M documents. **Seen Docs.** denotes the number of documents seen in training, **#E** denotes the number of epochs of the corresponding subset, and $P(i)$ denotes the probability of sampling a document i from that subset.

Dataset	Seen Docs.	#E	$P(i)$
ClueWeb PTPT	59.870M	2.06	0.62
PTWiki	9.516M	11.60	0.10
OSCAR PTPT	7.610M	4.88	0.08
ArquivoPT	7.598M	5.07	0.08
OpenSub. PTPT	5.707M	4.41	0.06
EuroParl PTPT	5.700M	4.08	0.06

4.2 Data Sampling Strategy

In order to take advantage of the diversity of our data, we implemented a sampling strategy similar to LLAMA's (Touvron et al., 2023a) where we attribute specific probabilities to each dataset, so that we can control which and how much data the model sees. In sum, a batch is prepared by sampling documents from every dataset according to pre-assigned sampling probabilities. Table 3 presents the total data seen during the 1.3B model pre-training as well as the sampling probabilities $p(i)$. Higher probability was given to ClueWeb since it constitutes the bulk of our data. Thus, we decided to spread the remaining datasets with balanced percentages, akin to LLAMA's distribution. The same weights were used for the 2.7B version.

4.3 Training Convergence

Figure 1 depicts the loss (at the left) and perplexity (at the right) evolution during pre-training, for the GlórIA1.3B variant. We start by observing a rapid loss decrease in the first 1 million steps. Then, a slower but steady decrease can be observed, until the end of the training.

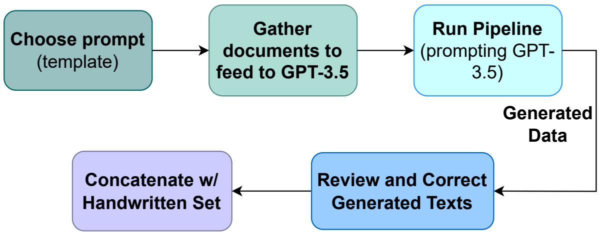

Figure 2: Overview of the CALAME-PT's generated set creation process.

5 Evaluation of PT Language Generation

We introduce the first zero-shot Portuguese language modeling benchmark, CALAME-PT (Context-Aware LAnguage Modeling Evaluation for Portuguese). Inspired by the widely used LAMBADA (Paperno et al., 2016) benchmark, the task consists of guessing the final word given the context that comes before it. It comprises **a total of 2076 texts and respective last words**, covering a wide variety of domains and contexts, whose context should be enough to guess the word. The topic diversity and the zero-shot setting directly requires models to leverage their inner knowledge to correctly solve the task. The target word can either be present or not in the context, which should be enough to predict it.

5.1 Building CALAME-PT

When creating the CALAME-PT benchmark, an hybrid approach is used to strike a balance between scale and diversity (w.r.t. to different domains and difficulty). As such, we produced two sets of samples: one with fully handwritten samples (**H**) and one with automatic generation+human review samples (**A**). For the handwritten set, a total of 406 samples were handwritten by 4 annotators, where it was sought to cover a broad set of domains.

For the automatic generation+human review, a pipeline was built to generate new texts grounded

Table 4: Examples of CALAME-PT's samples. We present the *prompts* and the target words the models should predict given the context, and if they're generated or handwritten.

Handwritten (H): *Um gato andava atrás do rato mas não o conseguia apanhar. Para todo o lado o rato fugia e fugia e o gato não o conseguia apanhar. Até que o gato se conseguiu adiantar e finalmente comeu o* rato	**Handwritten (H):** *A tragédia atingiu a família quando ele caiu no chão e não havia ninguém no local com formação em primeiros* socorros
Generated+Reviewed (A): N*o contexto apresentado, várias organizações do trabalho, como sindicatos e associações sindicais, estão envolvidas em negociações e revisões contratuais com várias empresas. Essas interações destacam a importância das negociações coletivas para garantir condições justas de trabalho. As organizações do trabalho trabalham em conjunto para representar os interesses dos* trabalhadores	**Generated+Reviewed (A):** *Depois de um período de controvérsia, uma empresa decidiu suspender a partilha de dados de utilizadores para fins publicitários. A decisão foi tomada após protestos em diferentes países. A suspensão é temporária e a empresa está a trabalhar com as autoridades para retomar a partilha de dados. Esta situação levanta questões sobre a segurança e privacidade dos* utilizadores

Table 5: The chosen prompt that was fed to GPT3.5 to generate a new, smaller text based on our documents.

Dado o seguinte contexto:
< DOC HERE >
Escreve um pequeno texto inspirado pelo contexto com poucas frases. Não deves mencionar nomes de pessoas ou países, eventos, marcas, e datas (dias, anos e horas).

on a set of randomly sampled documents from a small subset of documents from ArquivoPT, PTWiki, and OSCAR were chosen, purposely left out from the training set. This was accomplished by prompting GPT-3.5 - the prompt is shown in Table 5 - and generating a total of 2.5k samples. This process cost $\approx$ 7 euros. Then, these samples were human-reviewed to remove low-quality samples, anonymize samples, fix minor mistakes, and address ambiguity by performing small rewrites. In the end, we were left with 1670 generated samples. The handwritten and automatically generated+human reviewed sets were combined (**ALL**) to create the final version, resulting in a total of 2076 samples.

5.2 Caveats of LLM-based Sample Creation

When preparing CALAME-PT, the first difficulty was ensuring anonymization and removal of encyclopedic knowledge-dependent contexts, in order to make each samples' context self-contained. We asked GPT-3.5 to perform these steps but sometimes it would fail. The second issue was GPT-3.5's struggle to generate accurate European Portuguese text. While the generated text was generally correct, it had a tendency to shift to PT-BR, which led to the presence of sporadic PT-BR linguistic traits in some of the samples. Another note is the on ambiguous contexts, which can needlessly harm a model's performance (models may generate a word that makes sense but does not exactly correspond to the target word). Thus, we aimed toward a sensible balance between the ambiguity and predictability present in the samples.

5.3 Evaluation Protocol

We compared our 1.3B (1M to 3M steps' checkpoints) variants of GlórIA to two decoder-based models: Gervásio-PTPT and mGPT (Shliazhko et al., 2022). Gervásio-PTPT is based on the Pythia 1B model (Biderman et al., 2023), and mGPT is a 1.3B multilingual variant resembling the GPT-3 architecture. We chose greedy and beam search + top-k decoding strategies for evaluation, with 4 beams, $k = 50$, with a temperature of 1.0 and a token repetition penalty of 2. Due to its non-deterministic nature, we report the average of 3 runs.

The models were evaluated on the entire CALAME-PT dataset, in a zero-shot setting, followed by a separate evaluation of the handwritten (**H**) and automatically generated + human reviewed (**A**) sets. In practice, we have the models generate up to 5 new tokens, and we only consider the first full generated word. We then compare it against the ground-truth target last word, by ignoring casing

Table 6: CALAME-PT benchmark results (**Exact-Match**) comparison using the **greedy decoding strategy**.

Models	ALL	A	H
Gervásio-PTPT	19.03	19.88	15.52
mGPT	29.47	31.55	20.93
GlórIA 1.3B (1M Chk)	35.07	37.36	25.62
GlórIA 1.3B (2M Chk)	35.93	38.14	26.84
GlórIA 1.3B (3M Chk)	**36.61**	**38.86**	**27.34**

Table 7: CALAME-PT benchmark results (**Exact-Match**) comparison using the **beam search with top-k sampling strategy**. Each score is the average of 3 runs.

Models	ALL	A	H
Gervásio-PTPT	44.01	45.97	34.90
mGPT	47.14	50.03	35.87
GlórIA 1.3B (1M Chk)	50.99	53.21	38.75
GlórIA 1.3B (2M Chk)	51.80	53.69	41.95
GlórIA 1.3B (3M Chk)	**52.79**	**55.39**	**42.61**

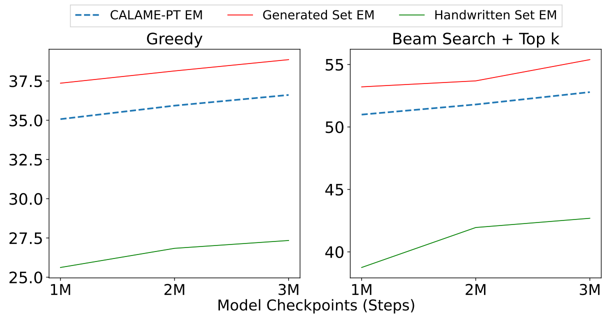

Figure 3: Evolution of GlórIA 1.3B performance on CALAME-PT. Evaluated at 3 distinct checkpoints (1M, 2M, and 3M steps) for both decoding strategies. **EM** denotes Exact-Match.

Table 8: CALAME-PT's generated set results (exact-match as percentage) discriminated by the source dataset used to create the samples (using beam search). *PW* - PTWiki. *Arq* - ArquivoPT. *Osc* - OscarPTPT.

Models	*PW*	*Arq*	*Osc*
Gervásio-PTPT	46.15	45.42	45.80
mGPT	49.69	50.08	50.57
GlórIA 1.3B (3M Chk)	53.84	55.76	56.20
GlórIA 2.7B (1M Chk)	54.61	54.06	55.89

and accents.

5.4 CALAME-PT Results Discussion

Overall Results. We present the results for CALAME-PT for the greedy and beam search+top-k strategies in, respectively, Tables 6 and 7. The first conclusion is that the beam-search + top-k sampling is significantly better for text generation, matching our initial qualitative observations. The second is that both versions of GlórIA outperform Gervásio-PTPT and mGPT by a relevant margin, in all settings. It can also be seen that training longer leads to a consistent performance improvement, with an observed $\approx 4\%$ relative improvement between the 1M and 3M checkpoints. This is also supported by Figure 3, which evidences the consistent performance evolution throughout training checkpoints.

Results per Subset (H vs. A). Regarding the results on each subset - handwritten (H) vs. automatically generated+human reviewed (A), it is interesting to see that samples from the H set are more challenging. In particular, we observe a 10% performance drop in GlórIA1.3B (3M Chk), with both decoding strategies, compared to the A set. We posit that there is an inherent bias to GPT-3.5 generated samples, that leads to more predictable target words.

Results Per Source on the Automatically Generated set (A). We recall that the automatically generated + human reviewed subset (A) was created by sampling documents from three different sources (ArquivoPT, PTWiki, OSCAR PT). To understand the models' performance per source, we present in Table 8 the results, by discriminating by the samples' dataset source. The main observation is that performance is quite balanced over the three distinct sources, over all the compared models. We observe that for samples grounded in OSCAR PT, performance is consistently (but marginally) higher. For GlórIA1.3B and mGPT, samples grounded on PTWiki are the most challenging.

5.5 Comparing 1.3B and 2.7B Models

To understand the model scaling possibilities of GlórIA, in this section we compare GlórIA1.3B with its 2.7B variant, both trained on 1M steps. Table 9 shows the results, where it can be observed that the 2.7B is able to outperform the 1M steps 1.3B variant. This leads us to strongly believe that GlórIA performance has the potential to increase by scaling the model and by conducting further pre-training.

Table 9: Comparison between GlórIA 1.3B and GlórIA 2.7B (EM), after 1M training steps, using beam search with top-k sampling. Each score is the average of 3 evaluations.

Models	ALL	A	H
GlórIA 1.3B (1M Chk)	50.99	53.21	38.75
GlórIA 2.7B (1M Chk)	52.20	54.57	40.40

6 Comparison to PT Encoder Models

We now compare GlórIA with state-of-the-art PT encoder models on PT discriminative/non-generative tasks. In these tasks, classification/regression heads are added to the pre-trained model and fine-tuned in a fully supervised setting. Previous research has shown that mostly due to their bidirectional nature, encoder models are particularly well-suited for many discriminative tasks, generally outperforming decoder-only models. For example, the GLUE leaderboard[5] is dominated by BERT-based models. In this section we compare GlórIA to other PT-encoder models. While we know priori that this is not the setting in which decoders excel, it will allow us to understand how GlórIA positions itself against encoder approaches.

6.1 Methodology Overview

In the following evaluations, we considered the 1.3B version of GlórIA and evaluated its 1M, 1.5M, 2M, and 3M step checkpoints.

For each task/subtask, we defined sets of hyperparameters to be evaluated (comprising learning rate, number of epochs, scheduler, etc.). Each model (including baselines) was fine-tuned in all hyperparameter sets, using the same protocol. In tasks with multiple target metrics, for each experiment, we kept the best checkpoint for each metric, based on the validation set. We then report the results obtained with the best set of hyperparameters. Furthermore, to increase robustness, each metric result was obtained by averaging the individual checkpoints' metric results.

6.2 ASSIN2

ASSIN-2 (Real et al., 2020) is a PT-BR multitask benchmark whose goal is to train and evaluate models for assessing both entailment (RTE) and similarity (STS) relations between sentences. Its training, validation, and test sets comprise 6.5k, 500, and 3k sentence pairs with annotations for both tasks, respectively. Due to ASSIN-2 being PT-BR, we compared GlórIA to BERTimbau-Large.

[5]GLUE Benchmark leaderboard

Table 10: Best results achieved for each baseline, on the ASSIN-2 task, across all experiments.

Model	F1	Accuracy	Pearson
GlórIA 1.3B	0.8960	0.8967	**0.8510**
BERTimbau-Large	**0.9020**	**0.9020**	0.8460

ASSIN-2 Protocols. For the ASSIN-2 benchmark, we follow (Souza et al., 2020) and perform a multi-task fine-tuning, by attaching two extra heads, each taking as input the embedding of the last token of the sequence. The final loss is the sum of the two losses from each task. RTE is treated as a classification task, thus we adopt the cross-entropy loss. STS is treated as a regression task, thus, we adopt the mean-squared error loss. To prepare the input, we tokenize the pair of sentences and pass the corresponding RTE and STS labels to the model, with a max sequence length of 128.

For this task's experimental space, we evaluated learning rates *1e-5* and *1e-6*, for 5 to 10 epochs, and for both linear and constant schedulers. A batch size of 32 was used with 2 GA steps. From these variations, we prepared 8 hyperparameter sets, and found that the most optimal combination for both our model and BERTimbau used a LR of *1e-5*, 10 epochs, and a constant scheduler.

ASSIN-2 Results. Table 10 shows the best results from each model on the ASSIN-2 task. A key observation is that GlórIA achieves equivalent results to the encoder-based baseline, BERTimbau-large. In fact, our model achieves top-performance in terms of Pearson score, and comes very close to BERTimbau's F1 and Accuracy scores.

6.3 Glue-PTPT

Given our focus on PT-PT, we evaluate GlórIA on GLUE-PTPT (Rodrigues et al., 2023), a PT-PT machine-translated version of GLUE (Wang et al., 2018). GLUE-PTPT comprises 4 subtasks of the original GLUE benchmark, from which we chose: RTE, MRPC, and STS-B. We compare GlórIA against Albertina-PTPT (encoder) (Rodrigues et al., 2023) and Gervásio-PTPT (decoder)[6].

[6]`https://huggingface.co/PORTULAN` - model name: `gervasio-ptpt-base`.

Table 11: Evaluation results on the GLUE-PTPT tasks across all experiments (all fine-tunes). *Enc.* stands for Encoders, and *Dec.* stands for Decoders.

	Models	RTE	MRPC		STS-B
		Acc	F1	Acc	Pearson
Dec.	GlórIA	0.6679	0.8775	0.8162	0.8500
	Gervásio-PTPT	0.6534	0.8599	0.7941	0.8360
Enc.	Albertina-PTPT	**0.8628**	**0.9261**	**0.8971**	**0.898**
	BERTimbau-Large	0.6968	0.9030	0.8652	0.8700

Glue-PTPT Protocols. Following the methodology, 4 hyperparameter sets were prepared for each subtask. The RTE and MRPC tasks share the same 4 sets - varying LR (*1e-4* and *1e-5*), linear and constant schedulers - while STS-B uses different ones - adding *1e-6* as an extra LR value. For all subtasks, models were fine-tuned for 5 epochs, with a batch size of 32, and 2 gradient accumulation steps. For the input, each pair of sentences is tokenized with their corresponding label, with a max sequence length of 128, due to the sentences being relatively short.

At the time of writing, GLUE's official evaluation service was not available, so we followed Albertina's protocol (Rodrigues et al., 2023) and used the original validation set as a test set, and took 10% from the original train split to create a new validation split. All models and baselines were fine-tuned using the created splits, to ensure comparability.

Glue-PTPT Results. The results, presented in Table 11, show that encoder-base models achieve better performance than decoder-based ones, with Albertina-PTPT achieving top performance followed by BERTimbau-large. Nevertheless, among decoder-base models, GlórIA significantly outperforms Gervásio-PTPT. This entails that among PT-PT decoder models, GlórIA is a robust choice.

7 Qualitative Results

To complement quantitative evaluation, we conduct a qualitative evaluation of GlórIA, by prompting the model to generate text for a set of topic-diverse prompts, using beam search w/ top-k sampling. The generated examples are illustrated in Table 12. The different generations showcase the model acquired knowledge across the different topics, ranging from *Culinary*, *Sports*, *Health*, *History*, etc. Namely, we observe that GlórIA **can output coherent and contextually correct PT-PT text**. In particular, the diversity of topics that we highlight in Table 12 hints that the model was able to capture the full range of topics that were present in the training data.

8 Discussion and Conclusions

8.1 Generative and Open Portuguese LLM

In this paper we proposed GlórIA, a generative and open large language model for Portuguese. In addition, we assemble a large-scale corpora for European Portuguese and contribute with CALAME-PT, a new benchmark for Portuguese generation tasks.

GlórIA achieves state-of-the-art results in Portuguese generative tasks and is a competitive model on many discriminative tasks. We believe that this success is attributed to its larger size, training duration, and especially to its large and rich 35+ billion tokens corpora, comprising multiple high-quality PT-PT sources.

8.2 Foundational Portuguese LLM and Broader Impact

GlórIA establishes a strong foundation to pursue new advances in language modeling for European Portuguese. Results demonstrated that GlórIA generates syntactically correct Portuguese language for a wide range of domains (Table 12). The generated language is also semantically correct, with sentence structures demonstrating a sound knowledge about multiple topics with limited hallucinations. Despite lacking a structured knowledge training task, the model was able to produce semantically coherent generations, by inferring entities, their relations, and context. For these reasons, we believe that GlórIA model lays out a strong foundation to tackle complex NLP tasks requiring chain-of-thought, zero/few-shot reasoning, human alignment, among other challenging scenarios.

Table 12: GlórIA 1.3B's text generation samples, with prompts and *generations*, using beam search w/ top-k (T=1.0) decoding. *NL* represents a *new line*.

Culinary: A cozinha portuguesa é rica em aromas e sabores. Existem muitos pratos *tradicionais, como o bacalhau com natas, o polvo à lagareiro ou a feijoada de búzios, mas também existem novos pratos que estão a ganhar cada vez mais adeptos. É o caso dos rojões à moda do* (...)	**Tech-Auto:** Os carros são considerados uma das maiores criações da engenharia. Os avanços no transporte de seres humanos *e máquinas, bem como as novas tecnologias, permitiram que o automóvel se tornasse um meio de transporte muito mais rápido e eficiente do que qualquer outro tipo de veículo.[NL] No entanto, a maior parte dos automóveis atuais ainda é baseada em motores de combustão interna* (...)
Sports: Surfing é um dos desportos mais praticados nas zonas costeiras de Portugal. Os adeptos do surf *em Portugal são cada vez mais e o número de praticantes tem vindo a aumentar nos últimos anos.[NL] O surf é um desporto que se pratica com uma prancha, geralmente de madeira, para deslizar sobre as ondas. É praticado em todo o mundo,* (...)	**Finance:** A literacia financeira permite às pessoas gerirem o seu dinheiro de forma mais responsável. Para aumentar a literacia financeira*, é importante que as crianças e os jovens aprendam a gerir o seu dinheiro de forma adequada.[NL] Aprender a gerir o seu dinheiro vai ajudá-lo a:[NL] Compreender como funciona o dinheiro;[NL] Identificar diferentes fontes de rendimento* (...)
Environment: Reciclar é muito importante para ajudar o ambiente. Reciclar *garrafas de plástico, latas e frascos de vidro pode ser uma boa forma de ajudar o meio ambiente.[NL] Porquê reciclar?[NL] A reciclagem é um processo que tem como objetivo reaproveitar os materiais em fim de vida, transformando-os* (...)	**History:** Antes do início do milénio, Portugal foi marcado por vários eventos marcantes, como por exemplo, em 1998, *a Expo'98, que se realizou no Parque das Nações, em Lisboa. Este evento serviu para mostrar ao mundo o que de melhor se fazia na área da engenharia e construção civil em Portugal.[NL] A Expo'98 é um dos maiores eventos* (...)
Music: A nível artístico, o legado musical português é rico. Alguns do maiores artistas portugueses são *reconhecidos internacionalmente, como Amália Rodrigues, Carlos Paredes, José Afonso, Adriano Correia de Oliveira, José Mário Branco, Fausto, Sérgio Godinho, entre muitos outros. A nível internacional, destacam-se nomes como a Orquestra Sinfónica Portuguesa, a Orquestra Gulbenkian, a* (...)	**Health:** É importante proteger a nossa saúde. Um dos cuidados a ter é a toma de vacinas anuais. As vacinas *são essenciais para a prevenção de doenças e, por isso, devemos tomá-las todos os anos.[NL] A vacina contra o tétano é uma das vacinas mais importantes para a proteção da nossa saúde. O tétano é uma doença que* (...)

8.3 Limitations

Our contributed CALAME-PT enables the evaluation of one particular facet of language modeling. However, the flexibility of such LLMs goes far beyond text completion, being capable of addressing tasks like abstractive summarization and dialog, either in zero or few-shot settings. Albeit such benchmarks are lacking for the Portuguese language, performing such evaluations would strengthen PT LLMs' research.

While the GlórIA generated text is syntactically, grammatically, and contextually correct, similarly to LLMs in other languages, *artifacts* may still be generated, including wrongly contextualized and non-factual generations. While some of these issues can be overcome with improved data selection (Ji et al., 2023), carefully designed prompts (Jin et al., 2022), or constrained decoding strategies (Rashkin et al., 2021), further research is still required to mitigate this behavior, as these are challenges that go beyond PT LLMs. Finally, while GlórIA is focused on European Portuguese, the ideal Portuguese LLM would cover other Portuguese variants as well (e.g. Mozambique, Guinea-Bissau, and others). Such promising research directions are left for future work.

8.4 Open Challenges

The framework proposed in this paper enables tackling open LLM challenges. This includes scaling the model to a larger number of parameters, including more training corpus, and expanding the model towards a multimodal LLM (Liu et al., 2023). In addition, GlórIA enables bringing new learning paradigms to Portuguese language modeling, such as LLM human-aligned generation: instruction tuning (Ouyang et al., 2022; Rafailov et al., 2023), factuality (Lee et al., 2022), and dialog (Silva et al., 2024; Ferreira et al., 2023).

Acknowledgements

We would like to thank Arquivo.pt's team for their content preservation efforts, and for all the help and guidance in accessing the archived web pages at scale. This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by CMU|Portugal project iFetch, Ref. CMUP LISBOA-01-0247-FEDER-045920, and by the FCT project Ref. Nº CPCA-IAC/AV/594875/2023.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2020. PTT5: pretraining and validating the T5 model on brazilian portuguese data. *CoRR*, abs/2008.09144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.

Rafael Ferreira, Diogo Tavares, Diogo Silva, Rodrigo Valério, João Bordalo, Inês Simões, Vasco Ramos, David Semedo, and Joao Magalhaes. 2023. Twiz: The wizard of multimodal conversational-stimulus. In *Alexa Prize TaskBot Challenge 2 Proceedings*.

Daniel Gomes, André Nogueira, João Miranda, and Miguel Costa. 2008. Introducing the portuguese web archive initiative.

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP -digital corpus of the European parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

Bernardo Leite and Henrique Lopes Cardoso. 2022. Neural question generation for the portuguese language: A preliminary study. In *Progress in Artificial Intelligence*, pages 780–793, Cham. Springer International Publishing.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John

Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362, New York, NY, USA. Association for Computing Machinery.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Claudia Moro, and Emerson Cabrera Paraiso. 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.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Diogo Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and João Magalhães. 2024. Plan-grounded large language models for dual goal conversational settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, St. Julian's, Malta. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *ArXiv*, abs/1912.07076.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Shen Zhuoran, Zhang Mingyuan, Zhao Haiyu, Yi Shuai, and Li Hongsheng. 2021. Efficient attention: Attention with linear complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538.