Sabiá: Portuguese Large Language Models

Ramon Pires*, Hugo Abonizio*, Thales Sales Almeida*, and Rodrigo Nogueira*

Maritaca AI

Abstract

As the capabilities of language models continue to advance, it is conceivable that "one-size-fits-all" model will remain as the main paradigm. For instance, given the vast number of languages worldwide, many of which are low-resource, the prevalent practice is to pretrain a single model on multiple languages. In this paper, we add to the growing body of evidence that challenges this practice, demonstrating that monolingual pretraining on the target language significantly improves models already extensively trained on diverse corpora. More specifically, we further pretrain GPT-J and LLaMA models on Portuguese texts using 3% or less of their original pretraining budget. Few-shot evaluations on *Poeta*, a suite of 14 Portuguese datasets, reveal that our models outperform English-centric and multilingual counterparts by a significant margin. Our best model, Sabiá-65B, performs on par with GPT-3.5-turbo. By evaluating on datasets originally conceived in the target language as well as translated ones, we study the contributions of language-specific pretraining in terms of 1) capturing linguistic nuances and structures inherent to the target language, and 2) enriching the model's knowledge about a domain or culture. Our results indicate that the majority of the benefits stem from the domain-specific knowledge acquired through monolingual pretraining.

1 Introduction

Language Models have revolutionized the field of natural language processing with their exceptional ability to perform tasks with minimal supervision. Although these models have been primarily pretrained on English-centric corpora, they have shown impressive multilingual capabilities [10]. Given the abundance of languages worldwide, the majority of which are low-resource, it has become a common practice to pretrain single models on multiple languages simultaneously. Models like XLM-R [12], mBART [32], mT5 [74], and BLOOM [57] exemplify this approach.

Despite the success of these multilingual models, we argue that they may not be the optimal approach for capturing the cultural and knowledge richness inherent in individual languages. When a moderately-sized language-specific corpus is available, continued pretraining could integrate the missing knowledge into the model, thereby enhancing its performance on targeted tasks. To test this hypothesis, we extend the pretraining of English-centric models using Portuguese corpora and evaluate their performance on an extensive range of Portuguese datasets employing a few-shot learning approach. Our results indicate that, even for models trained beyond the recommendations by Hoffmann et al [19], this additional pretraining considerably improves the performance compared to multilingual models.

We evaluate our models on datasets comprising texts originally created by native Brazilian Portuguese speakers, as well as datasets translated from English to Portuguese. We observe improvements across all datasets due to the Portuguese pretraining, with the gains being particularly pronounced for datasets created by Brazilian speakers. One of the largest improvements was observed on the ENEM dataset, which is derived from entrance exams used by Brazilian universities and requires extensive knowledge of the country's history, geography, and literature. This result provides evidence that the major contribution of our language-specific pretraining is to inject domain-specific knowledge about a particular culture as opposed to solely enhancing language proficiency.

*Equal contribution.

2 Related Work

The success of multilingual pretraining has been well-documented in the literature, with models such as ByT5 [73], mT5 [74], XLM-R [12], XGLM [30] and mGPT [59] paving the way for more inclusive language understanding and generation by leveraging shared knowledge across multiple languages. However, there are limitations to this approach.

BLOOM, a 175B-parameter model pretrained on 46 languages, demonstrates inferior performance on English tasks when compared to OPT [78], a similarly sized model pretrained with comparable computational resources and data size but that was trained on English-centric corpora. We conjecture that BLOOM's underperformance may be attributed to its relatively limited exposure to English tokens during the pretraining phase. Consequently, this observation suggests that monolingual pretraining could offer supplementary advantages.

In support of this hypothesis, models with hundreds of millions of parameters pretrained on monolingual texts have demonstrated gains over multilingual counterparts [7, 62, 6, 55, 36, 27, 8, 2, 28, 39, 22]. Additionally, research has indicated that language adaptation is beneficial even for low-resource languages [41, 14, 4, 76]. However, there is a limited number of published research articles with comprehensive evaluations of the benefits of continued pretraining at the multi-billion-parameter scale [26, 77, 23]. Through this study, we contribute to the literature by demonstrating the effectiveness of continued language-specific pretraining for Portuguese language models up to the 65B-parameter scale.

The question concerning whether it is advantageous to train models for specific languages is closely associated with the question of whether it is beneficial to train models for particular domains of knowledge. Recent studies on models such as Minerva [29] and Galactica [66] have shown that domain-specific pretraining can lead to significant improvements, even when the pretraining corpus is smaller in comparison to web-scale, general-purpose pretraining corpora. Analogously, Fu et al. [16] demonstrated the feasibility of specializing smaller models to perform multi-step reasoning — a capability typically exclusive to models with at least 50B parameters — at the expense of diminished performance in other, more general tasks.

Pretraining with a combination of general and domain-specific corpora can potentially enhance performance in specialized tasks without compromising effectiveness in general-purpose tasks, albeit at the cost of increased computational demands. For example, BloombergGPT [72], a 50B-parameter model pretrained on heterogeneous corpus in which more than half of texts are from the financial domain, exhibits comparable performance to OPT-66B in general tasks. BloombergGPT, however, was pretrained on a dataset three times larger in size, and consequently used more computational resources.

Rather than pursuing a single model that performs well across multiple domains, Gururangan et al. [18] propose an alternative approach: using multiple expert models, each trained on a domain-specific subset within a broader, diverse dataset, to function as a single general-purpose model. Their models outperform dense ones across various domain-specific tasks, at the expense of an increased parameter count, consequently leading to larger memory requirements for efficient inference.[1]

3 Methodology

In this section, we outline the pretraining data and training details used to build our models, including data sources, preprocessing techniques, architectures, hyperparameters, and optimization methods.

3.1 Pretraining Data

The pretraining data is derived from the Portuguese subset of the ClueWeb 2022 dataset [43, 44]. To increase the quality of the dataset, we apply the quality filters from MassiveText [48], modifying them to accommodate the specific requirements of the Portuguese language, such as using appropriate stopwords. We normalize the text with *ftfy* [63][2], convert wikitexts into human-readable texts, and exclude documents made up with less than 200 unique tokens.

[1]To serve their ensemble with a low latency, the weights for each expert must be kept in GPU memory.

[2]*ftfy* normalization fixes *mojibakes* and remove remnant HTML tags.

These quality filters are primarily designed for web pages and may not seamlessly transfer to other domains. There is potential for improvement by employing more automated methods; however, such experimentation was not conducted in this work due to the resource-intensive nature of pretraining experiments.

Following the cleaning process, all documents are concatenated using an end-of-sequence token as a separator, and then tokenized. The GPT-J tokenizer, which is identical to the GPT-2 tokenizer [47], produces 7.8 billion tokens, while the LLaMA tokenizer produces 7.3 billion tokens.

We extended the training of three models — LLaMA 7B and 65B [67] as well as GPT-J [70] — originally trained on English-centric corpora, on Portuguese texts; these further pretrained models from LLaMA are denoted as Sabiá, while the one derived from GPT-J is referred to as Sabiá-J.[3]

3.2 Sabiá models

The LLaMA 7B and 65B models are decoder-only Transformer models [68] with a similar architecture to PALM's [10]. The models were trained using a causal language modeling objective on a massive dataset sourced from webpages, code, books, and scientific papers, with the 7B model trained on 1 trillion tokens and the 65B model trained on 1.4 trillion tokens. The majority of the corpus consists of English text; however, it also contains Portuguese text, with the precise quantity not specified by the authors.

Starting from the LLaMA weights, we train the Sabiá models on our Portuguese dataset (described in Section 3.1) using the `t5x` and `seqio` frameworks [52]. Adhering closely to the hyperparameters used by PALM, we use the AdaFactor optimizer [58] without factorization, a first-order momentum $\beta_1 = 0.9$, and a second-order momentum $\beta_2 = 1 - k^{-0.8}$, where k represents the step number. We apply global norm clipping at 1.0 and dynamic weight decay of lr^2, with lr denoting the current learning rate.

Besides the standard causal language modeling loss, we use an auxiliary loss of $10^{-4} \log^2(\sum_i e^{z_i})$, where z are the logits, to decrease the likelihood of loss spikes at the 65B-parameter scale. The learning rate is linearly increased from 0 to 1e-3 over the initial 1,000 steps, followed by a constant learning rate of 1e-3 for an additional 9,000 steps.

The models were trained on a TPU v2-512, using batches of 512 sequences, each containing 2048 tokens. We utilized gradient checkpointing, also known as rematerialization, to enable the use of larger batches, thereby increasing TPU utilization. For the 7B model, this configuration results in a throughput of 124,000 tokens/sec, corresponding to a Model FLOPs Utilization (MFU) [10] of 45.2%, excluding the self-attention operations. For the 65B model, we achieve a throughput of 14,000 tokens/sec, resulting in an MFU of 47.4%.

The resulting models were trained on a total of 10.4 billion tokens, or 1.52 epochs of the Portuguese dataset. We noticed improvements in few-shot tasks beyond one epoch, which corroborates results from Taylor et al. [66]. However, due to the high costs of pretraining, we did not continue training.[4]

3.3 Sabiá-J

The GPT-J model is a 6B-parameter decoder-only Transformer model whose architecture and training hyperparameters closely follow GPT-3 6.7B. The main differences reside on computing the MLP and self-attention in parallel, applying attention head with dimension 256 (twice larger than GPT-3 6.7B), and using Rotary Positional Embedding (RoPE) [65]. GPT-J was trained on 400B tokens from The Pile dataset [17], whose 97.4% tokens are in English.

We begin training Sabiá-J from the released GPT-J checkpoint,[5] using the `mesh-transformer-jax` framework [69] and AdamW optimizer [34] with a weight decay of 0.1. We start the pretraining by warming up the learning rate until 1.2e-5 over 13,500 steps, followed by a cosine annealing decay during 135,518 steps until the end learning rate of 2.4e-6, and kept it constant from there on. We train on a TPU v3-8 using an effective batch size of 32 sequences of 2048 tokens. This results in a throughput of 5,200 tokens/sec, corresponding to a MFU of 44.5% without self-attention. The model was trained for 18 days on 7.8B tokens, or one epoch of the Portuguese dataset.

[3] Sabiá is a tribute to the eponymous bird, renowned for its diverse and intricate vocalizations.

[4] Considering the on-demand pricing of 384 USD per hour for a TPU v2-512, pretraining Sabiá-7B and Sabiá-65B cost approximately 9,000 and 80,000 USD, respectively.

[5] `https://huggingface.co/EleutherAI/gpt-j-6b`

4 Evaluation on Poeta

We evaluate the Sabiá models on the Portuguese Evaluation Tasks (Poeta) benchmark, which comprises 14 downstream NLP datasets in Portuguese: ASSIN 2 RTE and STS [51], ENEM Challenge [60], ENEM 2022 [40], FaQuAD [56], TweetSentBr [5], AG News [79], IMDB [35], MASSIVE [15], MKQA [33], BoolQ [11], SST2 [61], WSC [13], and BLUEX [1]. Half of them (ASSIN 2 RTE and STS, BLUEX, ENEM Challenge, ENEM 2022, FaQuAD, and TweetSentBr) were originally written in Portuguese, and the remaining ones were either manually or automatically translated into Portuguese from their originals in English. We refer to the first group as "Native" datasets and the second group as "Translated" datasets.[6] More details about the datasets can be found in Appendix A.

The models were evaluated in a few-shot manner using the maximum number of examples that fits into a 2048-token context for each task. We used the *GPT-2* tokenizer as reference because it results in more tokens. This allowed us to comfortably fit prompts tokenized with other tokenizers.

To evaluate the models, we manually select a set of few-shot examples for each dataset on Poeta. Depending on the dataset, these examples are balanced by class (except for FaQuAD, BLUEX, ENEM Challenge, ENEM 2022, MKQA, and WSC). For each test example, the prompts are built with the selected few-shot examples in alternating order. We also explore different alternatives to select few-shot examples in an ablation study (see Section 5.3). Each task on Poeta has a particular instruction that is placed at the beginning of the prompt.

Following Srivastava et al [64], we adopt the Normalized Preferred Metric (NPM) as our primary evaluation measure:

$$\texttt{NPM} = \frac{1}{N}\sum_{i=1}^{N} 100 \times \frac{\texttt{[raw preferred metric]}_i - \texttt{[random score]}_i}{\texttt{[high score]}_i - \texttt{[random score]}_i} \tag{1}$$

where N is the number of evaluation datasets, $\texttt{[raw preferred metric]}_i$ is the score obtained by the model on the i-th dataset, $\texttt{[random score]}_i$ is the score of a random model (e.g., 50% for a binary classification task) and $\texttt{[high score]}_i$ is the highest possible score on that dataset. The preferred metric and random score for each dataset is presented in Table 4. The rationale behind employing NPM rather than a straightforward average across all datasets is to mitigate the undue influence of datasets with inherently high scores, such as binary classification datasets, which could otherwise outweigh datasets characterized by lower scores.

5 Results

The main results can be found in Table 1. Models such as BLOOMZ, XGLM and Bertin-GPT struggled to generate answers in Portuguese. To address this issue, we adopted an approach akin to that used by the XGLM authors: by calculating the likelihood of each candidate answer string based on the input text and subsequently selecting the class with the highest probability. For FaQuAD, the only dataset in the benchmark without predetermined candidate answers, we allowed the models to generate answers in their original format.

We observe that the LLaMA baselines significantly outperform models of equivalent size trained with fewer tokens, such as Galactica and OPT. Furthermore, despite being trained on English-centric corpora, LLaMA-7B surpasses multilingual BLOOM and XGLM of similar sizes. The Sabiá models demonstrate considerable improvement in NPM compared to their respective baseline models. These NPM gains are more substantial for the smaller Sabiá-J and Sabiá-7B models. Notably, Sabiá-65B marginally outperforms OpenAI's GPT-3.5-turbo, which serves as the base model for ChatGPT.

Through our Portuguese pretraining, we observed that the improvement in NPM was higher in native datasets than that in translated datasets. For Sabiá-65B, improvements over LLaMA-65B were mostly from the native subset. We hypothesize that this is due to the "mechanistic" nature of translated datasets: since they were translated from English, the baseline model already possesses the knowledge needed to solve them and gains little from learning the linguistic, syntactic, and grammatical knowledge of the target language. For instance, to answer the question "*does p o box come before street address*" from the BoolQ dataset, the model gains little from additional pretraining on a Portuguese corpus as it

[6] The MASSIVE dataset underwent manual translation and localization; however, given that the original text was composed in English, it has been categorized as a translated dataset.

	Native	Translated	All
GALACTICA-6.7B	2.2	13.6	7.9
OPT-6.7B	5.3	39.7	22.5
OPT-66B	16.4	47.1	31.7
BERTIN-GPT	5.8	42.5	24.2
BLOOM-7.1B	10.6	44.2	27.4
BLOOMZ-7.1B	18.3	44.7	31.5
XGLM-7.5B	14.0	46.9	30.4
GPT-3.5-turbo	67.9	66.0	67.0
GPT-4	78.8	82.5	80.6
GPT-J	10.2	33.9	22.0
Sabiá-J	25.0	43.1	34.0
LLaMA-7B	20.2	45.8	33.0
Sabiá-7B	43.4	53.6	48.5
LLaMA-65B	59.1	68.4	63.7
Sabiá-65B	69.2	69.6	69.4

Table 1: Few-shot NPM results on the Poeta benchmark.

is unlikely that the corpus would provide new information regarding the formatting of American mailing addresses that the model has not already encountered during its initial English-centric pretraining. Conversely, language-specific pretraining introduces the specific knowledge required to solve tasks in the native subset.

Although GPT-J exhibited lower few-shot performance in English tasks relative to LLaMA, we use it in this study to illustrate that not only highly optimized models like LLaMA, but also less refined models such as GPT-J, can benefit from extended pretraining. We chose not to use BLOOM-7.1B as our initial checkpoint for pretraining due to its inferior performance compared to GPT-J in preliminary few-shot experiments on three Portuguese datasets. However, we later discovered that its performance on Poeta surpassed GPT-J's. Nonetheless, BLOOM still exhibits lower performance compared to LLaMA.

Analogous to Sabiá-J, BERTIN-GPT is a model pretrained on Spanish text starting from the GPT-J weights. Since Spanish and Portuguese are similar languages, it is reasonable to expect that BERTIN-GPT would perform better than its baseline model. Nevertheless, the observed NPM for BERTIN-GPT is only slightly higher than GPT-J's.

A noteworthy comparison involves Galactica, a model pretrained on scientific text, predominantly in English, and a similarly-sized OPT model, which utilized comparable pretraining compute but was pretrained on a larger and more diverse English-centric corpus. In their study, the authors demonstrate that Galactica performs on par with OPT on English tasks and largely outperforms OPT on scientific-related tasks. Conversely, in Portuguese tasks, OPT significantly outperforms Galactica. This result underscores the trade-offs associated with domain-specific specialization, which often entails diminished performance in other tasks.

BLOOMZ [38], a multilingual instruction-tuned model, demonstrated superior performance compared to its baseline BLOOM model, rivaling LLaMA of equivalent size.[7] Nevertheless, our approach of pretraining in Portuguese appears to yield superior results, as Sabiá-J surpasses BLOOMZ despite originating from a lower-performing baseline model. We envision continued pretraining and instruction tuning as complementary techniques to be combined in future research.

5.1 Results per Dataset

Table 2 presents the results per Poeta dataset for Sabiá models, their baselines, and for the supervised state-of-the-art. The SOTA results reported for the translated datasets were obtained using their original English versions [75, 80, 49, 54]. Since the Poeta benchmark excludes unanswerable examples of the MKQA dataset, we decided not to include the SOTA result for this dataset.

[7]This model was used in the experiments: https://huggingface.co/bigscience/bloomz-7b1-mt

In more challenging datasets, such as ENEM Challenge, ENEM 2022, and BLUEX, which are derived from admission exams to Brazilian universities, we see the most significant gains due to language-specific pretraining. Substantial improvements are also observed in TweetSentBr, a dataset containing tweets with an abundance of slang and references to Brazilian popular culture. We hypothesize that this pretraining imparts specific knowledge about the country's culture, literature, and geography that is less frequently encountered and learned during the original pretraining, which incorporates more diverse texts.

		Native							Translated						
	Avg	ASSIN 2 RTE (F1)	ASSIN 2 STS (Pearson)	BLUEX (Acc)	ENEM (Acc)	ENEM 2022 (Acc)	FaQuAD (F1)	TweetSentBr (F1-macro)	AG News (Acc)	BoolQ (Acc)	IMDB (Acc)	MASSIVE (F1-macro)	MKQA (F1)	SST2 (Acc)	WSC (Acc)
SOTA supervised	-	90.00[1]	86.00[2]	-	-	-	82.40[3]	77.27[4]	95.55[5]	92.40[6]	96.21[5]	-	-	97.50[7]	90.10[8]
GPT-4	84.99	90.96	77.58	76.40	92.00	79.66	84.74	82.40	93.50	86.50	97.00	83.30	55.67	97.50	92.63
GPT-3.5-turbo	76.08	88.28	66.41	60.11	80.57	75.42	78.28	74.39	87.71	71.43	84.86	84.19	44.92	91.71	76.84
Galactica-6.7B	34.11	34.92	11.63	28.65	20.74	22.88	40.16	21.98	38.33	57.13	51.08	35.62	3.01	62.27	49.12
Bertin-GPT-6B	45.18	33.24	6.23	22.47	20.52	22.03	64.00	35.52	82.44	44.25	87.66	55.46	15.56	81.08	62.11
OPT-6.7B	43.35	43.33	21.35	24.16	19.87	20.34	56.45	14.37	55.67	61.31	90.42	51.84	13.64	86.47	47.72
OPT-66B	49.68	65.66	7.88	29.78	20.41	17.80	71.12	32.54	81.87	58.75	92.66	61.64	21.17	87.50	46.67
BLOOM-7.1B	47.01	50.32	12.16	25.84	20.85	17.08	72.67	25.12	79.48	60.43	89.60	56.23	15.72	83.83	48.77
BLOOMZ-7.1B	50.94	33.57	24.50	34.27	28.38	27.12	79.90	50.36	83.82	38.23	93.80	55.31	12.36	86.93	64.56
XGLM-7.5B	48.79	53.75	15.07	24.16	19.10	19.49	44.84	63.23	77.47	49.76	91.46	59.74	13.72	89.11	62.11
GPT-J	43.51	54.88	17.86	24.72	20.85	20.34	59.52	20.98	64.15	48.75	72.68	55.67	10.69	83.94	54.04
Sabiá-J	52.84	35.49	22.97	39.89	39.41	36.44	69.28	64.16	84.30	51.53	90.86	58.82	13.84	87.16	45.61
LLAMA-7B	51.30	56.82	7.39	32.02	29.04	23.73	77.38	44.19	76.94	57.37	86.92	59.90	30.08	88.76	47.72
Sabiá-7B	62.43	64.87	13.63	47.75	60.59	60.17	77.43	67.17	83.28	64.07	92.70	68.95	31.98	90.60	50.88
LLAMA-65B	73.84	74.98	62.85	53.93	75.00	62.71	87.25	68.05	88.01	73.12	94.98	78.71	48.34	94.27	71.58
Sabiá-65B	77.65	88.07	63.29	57.87	90.39	72.03	88.47	72.91	88.34	75.96	92.76	79.41	49.47	93.43	74.74

Table 2: Results per dataset. [1][62]; [2][9]; [3][37]; [4][3]; [5][75]; [6][80]; [7][49]; [8][54].

Certain capabilities only emerge at scale, as evidenced by [71]. For example, 6-7B models perform close to the random baseline in datasets such as ASSIN 2 RTE and STS, and WSC. However, at the 65B scale, we observe substantial improvements, approaching or surpassing state-of-the-art supervised models on the ASSIN 2 RTE and FaQuAD datasets.

GPT-4 [42] results indicate that there is still room for improvement for Sabiá-65B in the majority of the datasets evaluated in this work. Nevertheless, Sabiá-65B performs on par with GPT-4 in datasets such as ASSIN 2 RTE, ENEM Challenge, and FaQuAD.

5.2 Data Contamination

The pretraining data for Sabiá models were collected up until February 2022. Since ENEM 2022 was publicly released in November 2022, the model could not have access to the answers for the questions present within its pretraining data. Consequently, the improvements observed at least for ENEM 2022, which were higher than the average of the datasets, cannot be attributed to data contamination.

5.3 Ablation: Alternative Prompt Types

In this ablation study, we explored alternative prompt types beyond the use of prompts with manually-selected few-shot examples. We studied another three types, described below.

Static prompt

This prompt configuration restricts the few-shot examples to the first k training examples of the dataset. Each test set example has the few-shot context composed of the same examples but in a different order. For datasets requiring the leave-one-out validation protocol: if the current test example is among the selected few-shot ones, it is replaced by the $k+1$-th example. For classification tasks, the prompt was class-balanced.

Dynamic-sampled prompt

For each test example, k examples are randomly sampled from the training set. For classification tasks, these few-shot examples are not necessarily balanced by class as they follow the training data distribution.

Dynamic-similar prompt

As proposed by Liu et al. [31], we treat the training examples as documents and employ a search engine to retrieve and rank them using the text of the current test example as a query, selecting the k most relevant ones. The BM25 algorithm [53] is used as a ranking function. If documents consist of multiple fields (e.g., premise and hypothesis), both the indexed contents and the query are formed by concatenating these fields.

Similar to the dynamic-sampled prompts, each test example will have a different set of few-shot examples. For classification tasks, to achieve a class-balanced set of few-shot examples, we organize N individual lists of top-M relevant documents for each class, where N is the number of classes and $M = k/N$, and then merge the lists. The merged list is reversed to place the most relevant documents at the end of the prompt.

5.3.1 Results

Table 3 reports the NPM results obtained for the Sabiá-7B model with the four different strategies to select few-shot examples explored in this ablation study. The manual choice of few-shot examples (manual) outperforms the use of the first examples of the training set to build the context (static) by more than two percentage points. The advantage of the manual approach is even greater when compared to randomly sampling the few-shot examples (dynamic-sampled).

	Manual	**Static**	**Dynamic-sampled**	**Dynamic-similar**
Sabiá-7B	48.5	46.4	45.3	52.3

Table 3: Few-shot NPM results on the Poeta benchmark on different prompt types.

Moreover, we observe that using dynamic-similar prompts improves the model's performance. However, the use of dynamic-similar prompts is discouraged for datasets that were intentionally biased in the creation of their training and test examples. For instance, the WSC dataset has pairs of passages that are identical except for one or two words:

- The city councilmen refused the demonstrators a permit because they *feared* violence.
- The city councilmen refused the demonstrators a permit because they *advocated* violence.

When one of these sentences is used as test, the other will be used in the prompt as the most similar case, which can give a clue to the model. Therefore, the use of dynamic-similar prompts should be avoided in such cases.

6 Limitations

Owing to the financial constraints associated with pretraining and, more significantly, the manual labor involved in collecting and curating evaluation datasets, experiments were conducted exclusively in Portuguese. Given that our models started pretraining from English-pretrained models and that Portuguese and English exhibit relatively close linguistic proximity, we anticipate that other researchers conducting further pretraining on languages closely related to English will observe comparable improvements in their target tasks. However, determining whether the benefits of our this method persist for languages more distant from English remains an open research question.

Portuguese is a language with an abundance of high-quality web-based texts. Thus, the gains observed with the proposed method may not necessarily extend to low-resource languages with limited availability of quality texts. In such cases, parameter-efficient methods [20, 46, 45] could be advantageous, as evidenced by Yong et al. [76]. We did not experiment with these techniques in this study due to the associated training costs, which are approximately equivalent to training the entire model.[8]

[8]Although parameter-efficient methods adjust only a fraction of the weights, they use only marginally fewer training FLOPs, as activations and gradients are computed for the entire model. For instance, LoRA [21], a parameter-efficient method, improves training throughput of a GPT-3 175B model by only approximately 32%.

7 Conclusion

In this study, we contributed to the expanding body of scientific evidence that specializing models for individual languages leads to improvements, even when the baseline model is large and extensively trained. We achieved this for the Portuguese language utilizing a near state-of-the-art model with 65 billion parameters. Given the relatively low pretraining cost and significant performance gains observed, we foresee a future landscape consisting of a diverse array of models, each tailored to a specific domain, rather than a single, all-encompassing model.

8 Acknowledgments

We thank Google Cloud for the generous TPU grant.

References

[1] Thales Sales Almeida, Thiago Laitz, Giovana K. Bonás, and Rodrigo Nogueira. Bluex: Brazillian leading universities entrance exams. *To appear*, 2023.

[2] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May 2020. European Language Resource Association.

[3] Tiago Martinho de Barros et al. Employing transformers and emoji to perform sentiment classification of social media texts: Utilizando transformers e emoji na classificação de sentimento de textos oriundos de redes sociais. 2021.

[4] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] Henrico Brum and Maria das Graças Volpe Nunes. Building a sentiment corpus of tweets in Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[6] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

[7] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[8] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[9] Ruan Chaves Rodrigues, Marc Tanti, and Rodrigo Agerri. Evaluation of Portuguese Language Models, 3 2023.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[11] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[13] Gabriela de Melo, Vinicius Imaizumi, and Fábio Cozman. Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798. SBC, 2019.

[14] Abteen Ebrahimi and Katharina Kann. How to adapt your pretrained multilingual model to 1600 languages. 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 4555–4567, Online, August 2021. Association for Computational Linguistics.

[15] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

[16] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*, 2023.

[17] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[18] Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Scaling expert language models with unsupervised domain discovery. *arXiv preprint arXiv:2303.14177*, 2023.

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[21] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[22] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*, 2021.

[23] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, 2021.

[24] Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*, 2020.

[25] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[26] Javier De la Rosa and Andres Fernández. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In Manuel Montes y Gómez, Julio Gonzalo, Francisco Rangel, Marco Casavantes, Miguel Ángel Álvarez Carmona, Gemma Bel-Enguix, Hugo Jair Escalante, Larissa Freitas, Antonio Miranda-Escalada, Francisco Rodríguez-Sánchez, Aiala Rosá, Marco Antonio Sobrevilla-Cabezudo, Mariona Taulé, and Rafael Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, 2022.

[27] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association.

[28] Hyunjae Lee, Jaewoong Yoon, Bonggyu Hwang, Seongho Joe, Seungjai Min, and Youngjune Gwon. Korealbert: Pretraining a lite bert model for korean language understanding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5551–5557. IEEE, 2021.

[29] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

[30] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, 2022.

[31] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[32] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[33] Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[35] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

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

[37] Guilherme Moraes, Luiz Henrique Bonifácio, Leandro Rodrigues de Souza, Rodrigo Nogueira, and Roberto Lotufo. A cost-benefit analysis of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*, 2021.

[38] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022.

[39] Dat Quoc Nguyen and Anh Tuan Nguyen. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online, November 2020. Association for Computational Linguistics.

[40] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams, 2023.

[41] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[42] OpenAI. Gpt-4 technical report, 2023.

[43] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362, 2022.

[44] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. Clueweb22: 10 billion web documents with visual and semantic information, 2022.

[45] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online, April 2021. Association for Computational Linguistics.

[46] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020.

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

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

[50] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[51] Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. The ASSIN 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 406–412. Springer, 2020.

[52] Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*, 13, 2022.

[53] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at TREC-3. *Nist Special Publication Sp*, 109:109, 1995.

[54] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[55] Gabriele Sarti and Malvina Nissim. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*, 2022.

[56] Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448, 2019.

[57] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[58] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[59] Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*, 2022.

[60] I. Cataneo Silveira and D. Deratani Maua. Advances in automatically solving the enem. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48, Los Alamitos, CA, USA, oct 2018. IEEE Computer Society.

[61] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[62] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 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*, pages 403–417. Springer, 2020.

[63] Robyn Speer. ftfy. Zenodo, 2019. Version 5.5.

[64] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[65] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[66] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[69] Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

[70] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

[71] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

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

[73] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[74] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[75] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* Curran Associates Inc., Red Hook, NY, USA, 2019.

[76] Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. Bloom+1: Adding language support to bloom for zero-shot prompting. *arXiv preprint arXiv:2212.09535*, 2022.

[77] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[78] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[79] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

[80] Barret Zoph. Designing effective sparse expert models. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1044–1044. IEEE, 2022.

A Details of the Evaluation Datasets

In this section, we detail the datasets present in the Poeta benchmark. Table 4 shows statistics of the datasets. We translated into Portuguese the datasets BoolQ, AG News, IMDB, and SST2 using Google Translate.

For the datasets that do not have training examples, we used the leave-one-out validation protocol. Particularly for ENEM Challenge and ENEM 2022, questions from other exams (except the one the current question came from) were selected to compose the set of few-shot examples.

In addition, we excluded large training examples that exceeded a certain number of characters for some datasets. To do this, we seek to find the maximum amount of characters that minimizes the cutoff (excluding a maximum of 25% of training) while maximizing the number of few-shot examples that would fit in a context of 2048 tokens. For instance, the AG News dataset had a cutoff of 298 characters and retained 110,953 examples, while the BoolQ dataset had a cutoff of 749 characters and kept 7,015 examples, and IMDB dataset had a cutoff of 1,544 characters and retained 18,613 examples. In some cases, large testing examples hindered the use of few-shot learning, so we truncated the documents in the BoolQ dataset at 1,169 characters (affecting 5% of the test set) and in the IMDB dataset at 2,095 characters (affecting 13% of the test set). A different strategy was applied for the two largest documents of the FaQuAD dataset. Instead of truncating the documents, we chose to reduce the number of few-shot examples by 1.

Dataset	Type	Preferred Metric	Rand. Score	Transl.	Avg Len (chars)	Num Train	Num Test	Num Few-shot
AG News	Multiclass classification (4)	Accuracy	25	Yes	282.34	120,000 (110,953)	7,600	12
ASSIN 2 RTE	Binary classification	F1	50	No	139.99	6,500	2448	18
ASSIN 2 STS	Regression	Pearson	0	No	139.99	6,500	2448	15
BLUEX	Multiple choice (4)	Accuracy	25	No	1,228.08	-	178	1
BoolQ	Binary classification	Accuracy	50	Yes	562.30	9,427 (7,015)	3,270	4
ENEM Challenge	Multiple choice (5)	Accuracy	20	No	1,286.68	-	916	1
ENEM 2022	Multiple choice (5)	Accuracy	20	No	1,170.24	-	118	1
FaQuAD	Extractive QA	F1	0	No	1,056.47	-	63	4
IMDB	Binary classification	Accuracy	50	Yes	1,114.56	25,000 (18,613)	25,000	2
MASSIVE	Multiclass classification (18)	F1-macro	0.58	Yes	68.35	11,514	2,974	36
MKQA	Extractive QA	F1	0	Yes	80.32	-	10,000 (6,758)	40
SST2	Binary classification	Accuracy	50	Yes	84.19	67,349	872	34
TweetSentBR	Multiclass classification (3)	F1-macro	32.4	No	93.32	12,990	2010	30
WSC	Binary classification	Accuracy	50	Yes	102.15	-	285	18

Table 4: A summary of the datasets constituting the Poeta benchmark.

Figure 1 presents prompts for some of Poeta datasets, translated into English. For simplicity, we omit the few-shot examples. The text in blue represents the target output. In the following subsections, we provide a more comprehensive description of the datasets.

A.1 FaQuAD

FaQuAD [56] is an extractive question-answering dataset about Brazilian higher education institutions, covering documents from a federal university, complemented with Wikipedia articles. The task parallels that of SQuAD [50], requiring the prediction of an answer span based on a given question and context as input. The primary evaluation metric is the token-level F1, wherein the predicted and ground truth answers are considered as sets of tokens, followed by the calculation of their respective F1 scores. Figure 1(a) presents an example of prompt for FaQuAD dataset.

A.2 ASSIN 2

ASSIN 2 is the second edition of the ASSIN (Avaliação de Similaridade Semântica e Inferência Textual - Evaluating Semantic Similarity and Textual Entailment) dataset and includes two tasks in the Portuguese language: Recognizing Textual Entailment (RTE), recently known as Natural Language Inference (NLI), and Semantic Textual Similarity (STS) [51]. The former predicts if a given text (premise) implies another text (hypothesis); while the latter measures the degree of semantic equivalence between two sentences. ASSIN 2 is composed of about 10,000 sentence pairs: 6,500 used for training, 500 for validation, and 2,448 for testing. The sentence pairs were annotated by native speakers of Brazilian Portuguese with linguistic training. Figure 1(b) shows an example of prompt for ASSIN 2 STS dataset.

A.3 ENEM Challenge

The ENEM Challenge is a dataset with questions from ten editions (2009-2017) of a multidisciplinary admission test widely used in Brazilian universities, the *Exame Nacional do Ensino Médio* (ENEM). ENEM is an annual exam[9] that has an essay and 180 multiple-choice questions with 5 alternatives divided into four knowledge areas. The dataset poses a challenge for LM because it requires a deep understanding of various fields of knowledge and the ability to integrate information from diverse domains. For comparison purposes, this work follows the authors [60] criteria by eliminating questions that require image comprehension and mathematical reasoning, and have chemical elements.

A.4 ENEM 2022

ENEM 2022 is a dataset composed of questions from the most recent edition of the ENEM exam [40]. The dataset is structured and annotated similarly to the ENEM Challenge dataset. Unlike the ENEM Challenge dataset, we kept questions that require mathematical reasoning and have chemical elements, excluding only those that depend on understanding images. However, we also exclude the five questions

[9] Although two exams were applied in 2016

(a) FaQuAD

Answer the question based on the context.

... Few-shot examples go here, separated with ###

Example N:

Context: Complementary rules for registration renewal. In the online course selection stage, the student must fill out an electronic form, through the Academic System (Siscad), indicating the subjects he intends to take in the subsequent academic period, respecting the prerequisites and the compatibility of class schedules of the selected subjects. Place priority is not guaranteed to students who have attended any prerequisites in the immediately preceding summer or winter term or who, for any reason, do not select the respective subject(s) at this stage of the registration renewal.

Question: Which subjects must the academic fill in the electronic form?

Answer: *subjects he intends to take in the subsequent academic period*

(b) ASSIN 2 STS

Rate the sentence pair as 1, 2, 3, 4, or 5.
1. completely different sentences, on different topics.
2. unrelated sentences, but on the same topic.
3. somewhat related sentences: they can describe different facts, but they share some details.
4. closely related sentences, but differ in some detail.
5. sentences mean essentially the same thing.

... Few-shot examples go here, separated with ###

Example N:

Sentence 1: The green field for horse racing is completely full of jockeys

Sentence 2: Jockeys are racing on horseback in the field, which is completely green

Answer: *4*

(c) ENEM 2022

Select the correct alternative.

... Few-shot examples go here, separated with ###

Example N:

Header: Emotional urgency. If everything is for yesterday, if life engages a first gear and takes off, if there is no more time for strategic stops, we fatally fall into the addiction of wanting love to be equally resolved in a split second. We are in a hurry to hear "I love you". We can't wait for the rules of coexistence to be established: are we boyfriends, hookups, married, lovers? Emotional urgency. A trap. We associate several words with LOVE: passion, romance, sex, adrenaline, palpitation. We forget, however, the word that makes this feeling possible: "patience". Love without patience does not avenge. Love cannot be chewed and swallowed with emergency, desperate hunger. It's a meal that can last a lifetime. MEDEIROS, M. Available at: http://porumavidasimples.blogspot.com.br. Accessed on: 20 Aug. 2017 (adapted).

Statement: In this opinion text, the linguistic marks reveal a relaxed situation with little formality, which is evidenced by the

Alternatives:

A. impersonalization throughout the text, as in: "if there is no more time".

B. construction of an atmosphere of urgency, in words like: "hurry".

C. repetition of a certain syntactic structure, as in: "If everything is for yesterday".

D. emphasis on the use of hyperbole, as in: "a meal that can last a lifetime".

E. use of metaphors, as in: "life engages a first gear and takes off".

Answer: *E. use of metaphors, as in: "life engages a first gear and takes off".*

(d) TweetSentBR

Rate the message as "positive", "neutral" or "negative".

... Few-shot examples go here, separated with ###

Example N:

Message: Father Fabio has been my spiritual director for a long time and I am grateful for the many times his teachings have guided me.

Answer: *positive*

(e) WSC

Select the correct alternative.

... Few-shot examples go here, separated with ###

Example N:

Sentence: The city councilmen refused the demonstrators a permit because

A. the city councilmen feared violence

B. the demonstrators feared violence

Answer: *A. the city councilmen feared violence*

Figure 1: Examples of prompt for selected datasets, translated into English for the convenience of non-Portuguese speaking readers. The text in blue is the expected output.

related to the foreign language (English or Spanish). Figure 1(c) illustrates a prompt for ENEM 2022 dataset.

A.5 BLUEX

The BLUEX datasets consists of high school level multiple-choice entrance exam questions from two distinguished Brazilian universities, namely Universidade Estadual de Campinas (UNICAMP) and Universidade de São Paulo (USP), covering various subjects such as geography, history, literature, chemistry, physics and mathematics. In this research, we have exclusively used questions from UNI-

CAMP exams administered in 2021 and onwards, which resulted in a total of four exams. We have eliminated questions that require image comprehension to be answered, resulting in a final set of 178 textual multiple-choice questions, each with 4 alternatives.

A.6 TweetSentBR

TweetSentBR is a sentiment classification dataset composed of 15.000 tweets on the TV show domain, collected between January and July in 2017 [5]. The sentences were manually labeled by seven annotators in three classes (positive, neutral and negative). The test set has 2010 tweets with a distribution similar to the training set: positive (45%), neutral (25%), and negative (29%). Figure 1(d) presents an example of prompt for TweetSentBR dataset.

A.7 AG News

AG News is a document categorization dataset obtained from the AG corpus of news articles on the web. The documents were selected from the four largest categories, i.e., world, sports, business, and science/technology, with equal distribution among them. The dataset is composed of 120,000 training examples and 7,600 test examples [79].

A.8 IMDB

The Internet Movie Database (IMDb) dataset is a vast collection of informal movie reviews gathered from the website with the same name. The dataset consists of 25,000 training examples and 25,000 test examples labeled as either positive or negative [35]. The documents in this dataset tend to be longer than those in other text classification datasets. Due to the considerable computational costs arising from its large test set, we opted to subsample it by utilizing only 5,000 examples while preserving a balanced class distribution.

A.9 MASSIVE

The Multilingual Amazon Slu resource package (SLURP) for Slot-filling, Intent classification, and Virtual assistant Evaluation (MASSIVE) is a 1M-example dataset [15] with realistic virtual utterances spanning 51 languages translated from English by professional translators. The dataset is annotated for slot (55 classes) and intents (60 classes) prediction tasks. As the number of classes is prohibitive for evaluating language models with in-context learning, instead of using the dataset for the originally proposed intent prediction, we simplify it to predict the 18 domains/scenarios. Specifically, we simplify intents like *alarm_query*, *alarm_remove* and *alarm_set* to the domain *alarm*. In this study, we use only its Portuguese subset.

A.10 MKQA

The Multilingual Knowledge Questions and Answers (MKQA) dataset is an open-domain question-answering dataset comprising 10k documents aligned across 26 typologically diverse languages [33]. The dataset was created by sampling 10,000 queries from the Natural Questions dataset and collecting new passage-independent answers for each query. These queries and answers were then translated into 25 non-English languages by human translators. MKQA was created to enable a fair comparison of question-answering performance across languages, without imposing assumptions on the underlying QA approach. Even in English, MKQA is a challenging evaluation set, and it is particularly difficult in low-resource languages. In this study, we use only the answerable examples (ignoring *unanswerable* and *long-answers* types) of the Portuguese language.

A.11 BoolQ

Boolean Questions (BoolQ) is a question-answering dataset for yes/no questions [11]. The authors gathered naturally occurring yes/no questions using the same data collection method as used by Natural Questions [25]. Each example in BoolQ is a triplet consisting of a question, a paragraph from Wikipedia containing the answer, and the yes/no answer.

A.12 WSC

The Winograd Schema Challenge (WSC) dataset consists of pairs of sentences that differ in one or two words, with a highly ambiguous pronoun that is resolved in opposite ways in the two sentences [24]. Although these examples are easy for humans to solve, they pose a challenge for machines, making WSC a benchmark for both commonsense reasoning and natural language understanding.

To conduct our study, we used a Portuguese-based collection of schemas that was based on the original set of 285 English-based schemas [13]. Each schema was manually translated by three native Portuguese speakers. The Portuguese collection also includes adaptations of some sentences, made to preserve the original meaning of the sentence and comply with the rules of Winograd schemas. For example, unlike in English, objects in Portuguese are not gender-neutral, so some objects were changed to other plausible ones. In addition, the authors replaced the original names with popular names in Brazil, with gender being the only restriction. Names of famous personalities were kept. Figure 1(e) shows an example of prompt for WSC dataset.

A.13 SST2

The Stanford Sentiment Treebank (SST) [61] is a corpus that contains fully labeled parse trees. The original corpus was parsed using the Stanford parser, resulting in a total of 215,154 unique phrases from those parse trees, each annotated by three human judges. In this study, we used the most common version of this dataset, referred to as SST2 or SST binary, designed for a sentiment analysis task. As neutral sentences are excluded, this is a binary classification task. Another commonly used version of this dataset, referred to as SST5 or SST fine-grained, considers five levels of sentiment polarity but was not used in this study.