Sabiá: Portuguese Large Language Models

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

Maritaca AI
{ramon,hugo,thales,rodrigo}@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 impact 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 most benefits stem from the domain-specific knowledge acquired through monolingual pretraining. Finally, we show that our Portuguese model has a lower performance in English tasks, thereby substantiating the inherent compromise in refining models for specific domains. Sabiá-7B is avaliable at HuggingFace:
https://huggingface.co/maritaca-ai/sabia-7b

1 Introduction

Language Models have revolutionized the field of natural language processing with their exceptional ability to perform tasks with minimal supervision. Although primarily pretrained on English-centric corpora, the models 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 [28], mT5 [70], and BLOOM [54] 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, 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 [18], this additional pretraining considerably improves 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 [57], 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.

2   Related Work

The success of multilingual pretraining has been well-documented in the literature, with models such as ByT5 [69], mT5 [70], XLM-R [12], XGLM [27] and mGPT [56] 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, performs worse on English tasks compared to OPT [74], a similarly sized model pretrained on English-centric corpora using comparable computational resources and data size. 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 [7,59,6,52,32,24,8,2,25,36,21]. Additionally, research has indicated that language adaptation is beneficial even for low-resource languages [38,13,4,72]. 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 [50,73,22]. 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, such as Minerva [26] and Galactica [62], have shown that domain-specific pretraining can lead to significant improvements, even with a smaller pretraining corpus

compared to large-scale, general-purpose pretraining corpora. Analogously, Fu et al. [15] 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 [68], 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. However, BloombergGPT's pretraining dataset is three times larger, and consequently used more computational resources.

Rather than pursuing a single model that performs well across multiple domains, Gururangan et al. [17] 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 [40,41]. To increase the datasets's quality, we apply the quality filters from MassiveText [45], modifying them to accommodate the specific requirements of the Portuguese language. We normalize the text with $ftfy$[2], convert wikitexts into human-readable texts, and exclude documents containing less than 200 unique tokens.

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, this study did not explore such approaches 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 [44], produces 7.8 billion tokens, while

---

[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.

the LLaMA tokenizer produces 7.3 billion tokens. The discrepancy in the total number of tokens is primarily due to the different tokenization strategies each model employs, byte-level BPE and BPE based on sentencepiece [23], respectively along with the variation of the vocabularies used by each tokenizer.

We extended the training of three models — LLaMA 7B and 65B [63] as well as GPT-J [66] — 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 [64] 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. The 7B model was trained on 1 trillion tokens and the 65B model was trained on 1.4 trillion tokens. While the majority of the corpus is in English, it also includes an unspecified amount of Portuguese text.

Starting from the LLaMA weights, we train the Sabiá models on our Portuguese dataset (see Section 3.1) using the t5x and seqio frameworks [48]. Adhering closely to the hyperparameters used by PALM, we use the AdaFactor optimizer [55] 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. This equals to 10,000 training steps. We noticed improvements in few-shot tasks beyond one epoch, which corroborates results from Taylor et al. [62]. However, due to the high costs of pretraining, we did not continue training.[4]

---

[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 costs approximately 9,000 and 80,000 USD, respectively.

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) [61]. GPT-J was trained on 400B tokens from The Pile dataset [16], 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 [65] and AdamW optimizer [30] 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.[6]

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 [47], ENEM Challenge [57], ENEM 2022 [37], FaQuAD [53], TweetSentBr [5], AG News [75], IMDB [31], MASSIVE [14], MKQA [29], BoolQ [11], SST2 [58], WSC [33], 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.[7]

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 a 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

---

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

[6] Due to constraints in our hardware budget, this model was trained with fewer tokens compared to Sabiá.

[7] 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.

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 1. A summary of the datasets constituting the Poeta benchmark.

examples in alternating order. Each task on Poeta has a particular instruction that is placed at the beginning of the prompt.

Following Srivastava et al [60], we adopt the Normalized Preferred Metric (NPM) as our primary evaluation measure:

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

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, which is either 1 or 100. The preferred metric and random score for each dataset are presented in Table 1. 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 2. 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.

Table 2. Few-shot NPM results on the Poeta benchmark.

	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

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*" (BoolQ dataset), the model gains little from additional pretraining on a Portuguese corpus as it is unlikely that the corpus would provide new information regarding the formatting of US 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 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, OPT significantly outperforms Galactica in Portuguese tasks. This result underscores the trade-offs associated with domain-specific specialization, which often entails diminished performance in other tasks.

BLOOMZ [35], a multilingual instruction-tuned model, demonstrated superior performance compared to its baseline BLOOM model, rivaling LLaMA of equivalent size.[8] 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 3 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 [71,76,46,51]. Since the Poeta benchmark excludes unanswerable examples of the MKQA dataset, we decided not to include the SOTA result for this dataset.

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 with more diverse texts.

Certain capabilities only emerge at scale, as evidenced by [67]. 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 [39] 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.

---

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

Table 3. Results per dataset. [1][49]; [2][9]; [3][34]; [4][3]; [5][71]; [6][76]; [7][46]; [8][51].

	Avg	Native							Translated						
		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	-	92.07[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

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. However, for the other datasets, the possibility of data contamination cannot be ruled out.

5.3 Ablation: English datasets

In this ablation study, we investigate the potential impact of Portuguese pretraining on the performance of the model in English datasets. We evaluated the LLaMA-7B and the Sabiá-7B models in English multiple-choice tasks. For simplicity, we employed a few-shot evaluation setup with 10 randomly selected examples (dynamic-sampled prompt). Importantly, we did not incorporate any descriptions or include Portuguese keywords to delimit the few-shot examples. We also restricted all the datasets to 350 test examples.

Following LLaMA's [63] approach, given the provided context, we select the answer with the highest likelihood normalized by the number of characters. The results in Table 4 indicate that the Sabiá-7B model exhibits a slightly reduced performance in English tasks compared to the baseline. This result corroborates our premise that model specialization invariably entails a balancing act, where improvements in one domain frequently coincide with degradation in another.

Table 4. Results in English datasets.

	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	*NPM*
LLaMA-7B	83.43	77.43	74.29	69.43	48.86	43.14	50.10
Sabiá-7B	80.86	75.71	72.29	72.86	50.00	42.29	49.02

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 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 [19,43,42] could be advantageous, as evidenced by Yong et al. [72]. We did not use these techniques in this study due to the training costs, which are approximately equivalent to training the entire model.[9]

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

References

1. Almeida, T.S., Laitz, T., Bonás, G.K., Nogueira, R.: Bluex: A benchmark based on brazilian leading universities entrance exams. To appear (2023)
2. Antoun, W., Baly, F., Hajj, H.: 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. pp. 9–15. European Language Resource Association, Marseille, France (May 2020)

---

[9] 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 [20], a parameter-efficient method, improves training throughput of a GPT-3 175B model by only nearly 32%.

3. Barros, T.M.d., 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. Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K.S., Islam, M.S., Iqbal, A., Rahman, M.S., Shahriyar, R.: BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1318–1327. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.98

5. Brum, H., Volpe Nunes, M.d.G.: Building a sentiment corpus of tweets in Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)

6. Carmo, D., Piau, M., Campiotti, I., Nogueira, R., Lotufo, R.: Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144 (2020)

7. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pretrained BERT model and evaluation data. In: PML4DC at ICLR 2020 (2020)

8. Chan, B., Schweter, S., Möller, T.: German's next language model. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6788–6796. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.598

9. Chaves Rodrigues, R., Tanti, M., Agerri, R.: Evaluation of Portuguese Language Models (3 2023). https://doi.org/10.5281/zenodo.7781848, https://github.com/ruanchaves/eplm

10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)

11. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: 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). pp. 2924–2936. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1300

12. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)

13. Ebrahimi, A., Kann, K.: 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). pp. 4555–4567. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.351

14. FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., Natarajan, P.: MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages (2022)

15. Fu, Y., Peng, H., Ou, L., Sabharwal, A., Khot, T.: Specializing smaller language models towards multi-step reasoning. arXiv preprint arXiv:2301.12726 (2023)

16. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)

17. Gururangan, S., Li, M., Lewis, M., Shi, W., Althoff, T., Smith, N.A., Zettlemoyer, L.: Scaling expert language models with unsupervised domain discovery. arXiv preprint arXiv:2303.14177 (2023)

18. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)

19. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)

20. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9

21. Kalyan, K.S., Rajasekharan, A., Sangeetha, S.: Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542 (2021)

22. Kim, B., Kim, H., Lee, S.W., Lee, G., Kwak, D., Hyeon, J.D., Park, S., Kim, S., Kim, S., Seo, D., 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. pp. 3405–3424 (2021)

23. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012

24. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: FlauBERT: Unsupervised language model pre-training for French. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2479–2490. European Language Resources Association, Marseille, France (May 2020)

25. Lee, H., Yoon, J., Hwang, B., Joe, S., Min, S., Gwon, Y.: Korealbert: Pretraining a lite bert model for korean language understanding. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5551–5557. IEEE (2021)

26. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al.: Solving quantitative reasoning problems with language models. arXiv preprint arXiv:2206.14858 (2022)

27. Lin, X.V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al.: Few-shot learning with multilingual generative language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9019–9052 (2022)

28. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8, 726–742 (2020)

29. Longpre, S., Lu, Y., Daiber, J.: MKQA: A linguistically diverse benchmark for multilingual open domain question answering. Transactions of the Association for Computational Linguistics 9, 1389–1406 (2021)

30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
31. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011)
32. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.645
33. de Melo, G., Imaizumi, V., Cozman, F.: Winograd schemas in portuguese. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. pp. 787–798. SBC (2019)
34. Moraes, G., Bonifácio, L.H., Rodrigues de Souza, L., Nogueira, R., Lotufo, R.: A cost-benefit analysis of cross-lingual transfer methods. arXiv preprint arXiv:2105.06813 (2021), https://arxiv.org/abs/2105.06813
35. Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C.: Crosslingual generalization through multitask finetuning (2022)
36. Nguyen, D.Q., Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1037–1042. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.92
37. Nunes, D., Primi, R., Pires, R., Lotufo, R., Nogueira, R.: Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams (2023)
38. Ogueji, K., Zhu, Y., Lin, J.: 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. pp. 116–126. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021)
39. OpenAI: Gpt-4 technical report (2023)
40. Overwijk, A., Xiong, C., Callan, J.: Clueweb22: 10 billion web documents with rich information. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3360–3362 (2022)
41. Overwijk, A., Xiong, C., Liu, X., VandenBerg, C., Callan, J.: Clueweb22: 10 billion web documents with visual and semantic information (2022)
42. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: AdapterFusion: Nondestructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 487–503. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.39
43. Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. arXiv preprint arXiv:2005.00052 (2020)
44. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
45. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)

46. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)

47. Real, L., Fonseca, E., Gonçalo Oliveira, H.: 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. pp. 406–412. Springer (2020)

48. Roberts, A., Chung, H.W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., et al.: Scaling up models and data with t5x and seqio. arXiv preprint arXiv:2203.17189 **13** (2022)

49. Rosa, G.M., Bonifacio, L.H., de Souza, L.R., Lotufo, R., Nogueira, R.: A cost-benefit analysis of cross-lingual transfer methods. arXiv preprint arXiv:2105.06813 (2021)

50. la Rosa, J.D., Fernández, A.: Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In: y Gómez, M.M., Gonzalo, J., Rangel, F., Casavantes, M., Ángel Álvarez Carmona, M., Bel-Enguix, G., Escalante, H.J., Freitas, L., Miranda-Escalada, A., Rodríguez-Sánchez, F., Rosá, A., Sobrevilla-Cabezudo, M.A., Taulé, M., Valencia-García, R. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings (2022)

51. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM **64**(9), 99–106 (2021)

52. Sarti, G., Nissim, M.: It5: Large-scale text-to-text pretraining for italian language understanding and generation. arXiv preprint arXiv:2203.03759 (2022)

53. Sayama, H.F., Araujo, A.V., Fernandes, E.R.: FaQuAD: Reading comprehension dataset in the domain of brazilian higher education. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 443–448 (2019). https://doi.org/10.1109/BRACIS.2019.00084

54. Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)

55. Shazeer, N., Stern, M.: Adafactor: Adaptive learning rates with sublinear memory cost. In: International Conference on Machine Learning. pp. 4596–4604. PMLR (2018)

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

57. Silveira, I.C., Maua, D.D.: Advances in automatically solving the enem. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). pp. 43–48. IEEE Computer Society, Los Alamitos, CA, USA (oct 2018). https://doi.org/10.1109/BRACIS.2018.00016

58. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013)

59. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for brazilian portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9. pp. 403–417. Springer (2020)

60. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
61. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864 (2021)
62. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R.: Galactica: A large language model for science. arXiv preprint arXiv:2211.09085 (2022)
63. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
65. Wang, B.: Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax (May 2021)
66. Wang, B., Komatsuzaki, A.: GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model (May 2021)
67. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. Transactions on Machine Learning Research (2022), survey Certification
68. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: BloombergGPT: A large language model for finance (2023)
69. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: Byt5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics **10**, 291–306 (2022)
70. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)
71. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. Curran Associates Inc., Red Hook, NY, USA (2019)
72. Yong, Z.X., Schoelkopf, H., Muennighoff, N., Aji, A.F., Adelani, D.I., Almubarak, K., Bari, M.S., Sutawika, L., Kasai, J., Baruwa, A., et al.: Bloom+ 1: Adding language support to bloom for zero-shot prompting. arXiv preprint arXiv:2212.09535 (2022)
73. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
74. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
75. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. In: NIPS (2015)
76. Zoph, B.: Designing effective sparse expert models. In: 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). pp. 1044–1044. IEEE (2022)