GemBode and PhiBode: Adapting Small Language Models to Brazilian Portuguese

Gabriel Lino Garcia[1](✉), Pedro Henrique Paiola[1], Eduardo Garcia[2], João Renato Ribeiro Manesco[1], and João Paulo Papa[1]

[1] School of Sciences, São Paulo State University (UNESP), Bauru, SP, Brazil
{gabriel.lino,pedro.paiola,joao.r.manesco,joao.papa}@unesp.br
[2] Institute of Informatics, Federal University of Goiás, Goiânia, GO, Brazil

Abstract. Recent advances in generative capabilities provided by large language models have reshaped technology research and human society's cognitive abilities, bringing new innovative capacities to artificial intelligence solutions. However, the size of such models has raised several concerns regarding their alignment with hardware-limited resources. This paper presents a comprehensive study on training Portuguese-focused Small Language Models (SLMs). We have developed a unique dataset for training our models and employed full fine-tuning, as well as PEFT approaches for comparative analysis. We used Microsoft's Phi and Google's Gemma as base models to create our own, named PhiBode and GemBode. These models range from approximately 1 billion to 7 billion parameters, with a total of ten models developed. Our findings provide valuable insights into the performance and applicability of these models, contributing significantly to the field of Portuguese language processing. This research is a step forward in understanding and improving the performance of SLMs in Portuguese. The comparative analysis of the models provides a clear benchmark for future research in this area. The results demonstrate the effectiveness of our training methods and the potential of our models for various applications. This paper significantly contributes to language model training, particularly for the Portuguese language.

Keywords: Small Language Models · Portuguese · Bode · Generative Artificial Intelligence · Natural Language Processing

1 Introduction

Recent Natural Language Processing (NLP) advances have been primarily driven by developing increasingly sophisticated and powerful language models [30]. Large Language Models (LLMs), in particular, are known for their ability to understand and generate text with great precision, opening ways for a wide range of applications, including automatic translation, text summarization, and question and answer, among others [21].

R. J. Barrientos and S. A. Velastin (Eds.): CIARP 2024, LNCS 15368, pp. 228–243, 2025.
https://doi.org/10.1007/978-3-031-76607-7_17

Despite their impressive effectiveness, these models usually come paired with a substantial cost in computing resources, often requiring huge amounts of memory and processing power, unattainable to several researchers and developers with limited infrastructure [18]. To tackle this issue, Small Language Models (SLMs) have emerged in the literature as a way to minimize the computational resource requirement. These models allow language models to be deployed in resource-constrained environments with limited memory and processing power.

The underlying concept behind SLMs is to offer a more affordable and resource-efficient alternative to various NLP applications. However, a fundamental issue faced when dealing with SLMs is their performance trade-off compared to large-scale models, as demonstrated by [30]. When trained with more data, these smaller models can match or even outperform their larger counterparts in certain benchmarks. Still, given their size and training capacity limitations, SLMs often fail to achieve the same accuracy and generalization capabilities as their larger counterparts in many other applications. Nonetheless, several prominent works have emerged in this direction, including Microsoft's Phi-1.5 [18], Phi-2[1] and Phi-3 [1], as well as Gemma models, based on Gemini, made available by Google [29].

As Portuguese is an important language spoken by more than 250 million people worldwide, it is paramount to address the necessity of developing language models for that scenario. As such, several works have already been proposed as a way to adapt LLMs to the Brazilian Portuguese (BP) language, such as Sabiá [22], a model developed by Maritaca AI, explicitly designed for BP, which has the first model based on the LLaMA-1 and Sabiá-J architecture, and a second version under a proprietary license, with specifics regarding parameter count and architectural design not publicly disclosed.

With the release of the LLaMa open-weight models, several models that enhance its capabilities for Brazilian Portuguese have been proposed, such as Canarim [10] and Bode [14]. The Canarim model was obtained by pretraining an LLaMa2-7B model using the Portuguese subset of Common Crawl solely on the language modeling task. Conversely, Bode employs low-rank adaptation (LoRA) to fine-tune and improve the LLaMa-2 model using the Alpaca dataset in prompt-based tasks for Portuguese instruction-following response.

Another notable model is the openCabrita [17], a 3B parameter model based on the OpenLLaMa, designed to provide a more cost-effective solution for Portuguese language processing. It employs a language-specific tokenizer adaptation approach to optimize token usage and reduce inference time, offering tangible benefits for real-world applications. Following a similar approach, the TeenyTinyLLaMa [6] model was proposed, presenting a compact model under the permissive Apache 2.0 license, with only 460 million parameters, following the Chinchilla scaling laws.

Despite the progress in language models for Brazilian Portuguese, there is still a need for more inclusive and versatile language approaches that prioritize efficiency and are based on modern SLM architectures. This work aims to

[1] Available at: https://huggingface.co/microsoft/phi-2 Accessed on: April 18th, 2024.

enhance the adaptation of linguistic models, specifically focusing on the Phi and Gemma models provided by Microsoft and Google, respectively. We introduce the adapted versions of PhiBode (1.5B, 2B, and 3B) and GemBode (2B, 2B-it, 7B, and 7B-it). These models are designed to bridge the gap in developing linguistic resources for Brazilian Portuguese, with the aim of improving performance across various Natural Language Processing (NLP) tasks. Our ultimate goal is to make these models more resource-efficient and enable a broader range of researchers and developers to benefit from these technologies. As such, the main contributions of this work are described as follows:

- **Comparison of Tuning Techniques:** This work compares the effectiveness of fully fine-tuned models with others fine-tuned through parameter-efficient tuning techniques (PEFT), aiming to evaluate not only models with smaller parameter sizes but also the impact of using efficient tuning techniques.
- **Introduction of UltraAlpaca Dataset:** A new multi-task dataset named UltraAlpaca, has been created by aggregating multiple English datasets translated into Portuguese.
- **In-depth Evaluation of Small Language Models:** We conduct a thorough evaluation of small language models specifically within the context of the Portuguese language.

This paper is organized as follows: Sect. 2 introduces the methodology employed by the proposed method, technical details of their architectures, the proposed dataset, and information on the tuning process. Section 3 discusses the experimental setup configurations used for evaluating the models on the benchmarks. Section 4 outlines the results, and Sect. 5 states conclusions.

2 Methodology

This section describes the proposed method and the foundational models used as the basis for our approach, along with details on our proposed dataset and the conditions under which it was created. Moreover, we describe the tuning techniques employed for evaluation over different scenarios.

2.1 Proposed Method

In this study, we have chosen the linguistic models Phi and Gemma, provided by Microsoft and Google, respectively, as the foundation for our investigation. Despite their sophistication, they were not specifically trained in Portuguese. Nevertheless, it is reasonable to assume that these models may have been exposed to Portuguese data during their training. To optimize the performance of these models for Portuguese, we have decided to use the UltraAlpaca dataset for further training.

The UltraAlpaca dataset, a unique collection of exclusively Portuguese samples, is a rare resource exceptionally suited to enhancing language models for this language. Leveraging this dataset, we conducted both PEFT and full fine-tuning of the Phi and Gemma models.

By performing PEFT and full fine-tuning based on UltraAlpaca, we aimed to adapt the Phi and Gemma models to yield more accurate and relevant results in Portuguese. This approach will allow us to maximize the potential of these pre-trained models, adapting them to the nuances and peculiarities of the Portuguese language. We hope that the results obtained with this methodology will significantly contribute to the advancement of natural language processing in Portuguese and for specific applications of this language.

2.2 Foundation Models

Two families of SLMs, Phi and Gemma, developed by Microsoft and Google, respectively, were chosen as the baseline for developing our Portuguese language models. The decision to use these models was motivated by their effectiveness in Portuguese and English, especially their performance per number of parameters. Both are capable of achieving performance comparable to that of large-scale models. This section will introduce these models, with details of their architectures discussed in more depth.

Introduced by Microsoft, the Phi [18] family of language models represents a turning point in the research of SLMs. The first iteration of the Phi model family focused on training a compact model using high-quality textbook-like texts and synthetic data from GPT-3.5, achieving significant results with only 1.3B parameters. Its successor expanded to 2.7 billion parameters, trained on an augmented dataset from the first model, delivering comparable performance to much larger models. The latest Phi-3-mini iteration of the family [1] is made of a 3.8 billion parameter SLM trained on3.3 trillion tokens, rivaling larger models while being mobile-friendly. This generation also introduced configurations with 7B and 14B parameters, capable of achieving even more significant results.

Gemma, another staple in the SLM state-of-the-art, comprises a family of open models based on Google's Gemini [29]. These models are trained on up to 6 trillion tokens, featuring variants with parameter counts ranging from 2 billion to 7 billion. Like the Phi model, Gemma adopts a transformer decoder architecture with a context length of $8,192$ tokens. However, Gemma incorporates several enhancements built upon previous works, such as multi-query attention, RoPE embeddings, GeGLU Activations, and RMSNorm. Although inspired by the Gemini architecture, the Gemma model was trained primarily on English data and did not demonstrate multimodal capabilities.

2.3 UltraAlpaca

In our research, we used models provided by Microsoft and Google and trained them using a dataset created by our team called UltraAlpaca. This dataset is based on several English databases translated into Portuguese. The datasets used to compose UltraAlpaca were:

- **Alpaca** [28]: A dataset of 52, 000 instruction-following samples generated by OpenAI's text-davinci-003 engine. We used the translated version of Alpaca [17] for training the first version of Cabrita.
- **UltraChat** [9]: Self-refinement dataset that comprises 1.47 million multi-turn dialogues generated by GPT-3.5-TURBO, spanning 30 topics and 20 distinct types of text material. From this dataset, 70, 000 samples were selected and translated to compose UltraAlpaca.
- **Aya** [27]: A multilingual instruction fine-tuning dataset curated by an open-science community via the Aya Annotation Platform from Cohere For AI. It contains 204, 000 human-annotated prompt-completion pairs, along with annotator demographics data. For this study, we filtered the Portuguese samples from this dataset.
- **OpenAssistant Conversations (OASST1)** [16]: Multilingual corpus of 161, 443 assistant-style messages in 35 languages, annotated with 461, 292 quality ratings, forming over 10, 000 fully annotated conversation trees. The dataset was created through a global crowd-sourcing effort with over 13, 500 volunteers. For UltraAlpaca, the Portuguese samples from this dataset were filtered.
- **Code Alpaca** [5]: A dataset of 20, 000 samples built similarly to Alpaca, focusing on code generation. The dataset was fully translated to be integrated into UltraAlpaca.
- **MetaMathQA-40K-PTBR** [32]: Specialized dataset designed to enhance mathematical reasoning capabilities in LLMs comprising 395, 000 samples. We used 40, 000 previously translated samples from this dataset.

The choice of these datasets to compose UltraAlpaca resulted from an attempt to write a comprehensive database comprising several tasks. Some works in the literature [19,33] point to the benefits of using multi-task datasets for training LLMs, increasing their chances of gaining emerging capabilities.

After translating these databases, we combined them to create UltraAlpaca. It is important to note that syntactic and semantic problems may occur even with translations, as translation does not guarantee high-quality data. However, in our experiments, these translated datasets improved the performance of the original models, as later discussed. This work represents a significant leap forward in developing LLMs, particularly for languages other than English. While our approach has challenges, it offers a promising direction for future research.

2.4 Efficient Tuning Techniques

As a route to adapt already existing language models to Portuguese, it is necessary to fine-tune the model in a language-specific dataset. One way to do that is called *full fine-tuning*, in which a pre-trained model is loaded, and all model weights are adjusted based on the calculated error. Although heavy in computer resource usage, this approach usually has the potential to provide the best results since it is capable, in theory, of working on the finer details of the model.

On the other hand, to use fewer computer resources while still maintaining acceptable model accuracy, fine-tuning can be performed on only the most relevant weights of the model. Certain techniques such as Low-rank Adaptation (LoRA) [15] perform *Parameter Efficient Fine-Tuning (PEFT)* by introducing two low-rank decomposition matrices within the dense layers, which are then fine-tuned to represent an update to the original model's weights.

QLoRA [8] refers to a more efficient alternative of LoRA, in which, by integrating quantization techniques, fewer computer resources are used for fine-tuning. QLoRA applies quantization to 4-bit precision on a pre-trained model to reduce space usage. After that, non-quantized decomposition matrices are calculated using LoRA to fine-tune the model properly. This work compares the fully fine-tuned techniques with the PEFT-based fine-tuning, in which QLoRA was employed.

3 Experimental Setup

This section discusses the experimental setups used on the developed models for each experiment and provides detailed information on the benchmark and techniques.

3.1 Open Portuguese LLM Leaderboard

Resources like the Hugging Face Open LLM Leaderboard [3] offer valuable insights for English language models. Still, they need more coverage for Portuguese, creating a significant gap in understanding the capabilities of LLMs for this language. Existing Portuguese-centric benchmarks, such as the Poeta benchmark [22], do not have a fully open reproducible evaluation script and incorporate machine-translated datasets, potentially introducing biases and limiting their reliability for assessing proper language understanding.

The Open Portuguese LLM Leaderboard [13] is a fully open and reproducible benchmark designed to evaluate the performance of Large Language Models. It was created using the EleutherAI-Language Model Evaluation Harness [12], a unified framework for testing generative language models on various evaluation tasks. The leaderboard ensures consistent and standardized evaluation across diverse tasks, enabling meaningful comparisons between different models and fostering transparency in performance assessment. The leaderboard currently includes nine benchmarks:

- **ENEM** [20,23,26]: The *Exame Nacional do Ensino Médio* (ENEM) is a standardized Brazilian national exam for high school students, covering various subjects such as natural sciences, human sciences, languages, and mathematics. This benchmark comprises $1,430$ questions from exams between 2010 and 2018, as well as 2022 and 2023, excluding questions that require image understanding. For utilizing this dataset in the Open Portuguese LLM Leaderboard, 3 examples were used for each prompt, following a few-shot approach. The performance was evaluated based on accuracy.

- **BLUEX** [2]: The Brazilian Leading Universities Entrance eXams (BLUEX) dataset comprises entrance exams from two top Brazilian universities, spanning from 2018 to 2024. This benchmark includes 724 questions that do not necessitate image understanding. For this dataset's application in the Open Portuguese LLM Leaderboard, 3 examples were employed for each prompt, and this performance was assessed based on accuracy.
- **OAB Exams** [7]: The Order of Attorneys of Brazil (OAB) Exams are professional certification exams for lawyers in Brazil, evaluating their legal knowledge and reasoning skills. This benchmark includes over $2,000$ questions from exams between 2010 and 2018. This dataset was used with 3 examples for each prompt. The performance metric used was accuracy.
- **ASSIN2 RTE** [24]: The ASSIN2 Recognizing Textual Entailment (RTE) task assesses a model's ability to determine whether a given text entails another text. This benchmark uses a dataset of sentence pairs annotated with human judgments for RTE. For this dataset, 15 examples were used for each prompt and were evaluated using the macro F1 score as the performance metric.
- **ASSIN2 STS** [24]: The ASSIN2 Semantic Textual Similarity (STS) task measures a model's capability to determine the degree of semantic similarity between two sentences. This benchmark utilizes the same dataset as ASSIN2 RTE but with annotations for STS. This dataset used 15 examples for each prompt, and the performance metric used was the Pearson Correlation Coefficient.
- **FAQUAD NLI** [25]: The FAQUAD Natural Language Inference (NLI) task is derived from the FAQUAD question-answering dataset and evaluates a model's ability to perform textual entailment between a question and its possible answers. This benchmark contains 900 questions about 249 reading passages. In this dataset, we used 15 examples for each prompt. The performance metric used was the macro F1 score.
- **HateBR** [31]: The HateBR dataset is a collection of $7,000$ Brazilian Instagram comments annotated for hate speech and offensive language detection. For this dataset's application in the Open Portuguese LLM Leaderboard, 25 examples were employed for each prompt. Performance was assessed based on a macro F1 score.
- **PT Hate Speech** [11]: The Portuguese Hate Speech dataset consists of $5,668$ tweets labeled for hate speech detection in Portuguese. This dataset used 25 examples for each prompt, and the performance metric used was macro F1 score.
- **TweetSentBR** [4]: The tweetSentBR dataset is a corpus of $15,000$ Brazilian Portuguese tweets annotated for sentiment analysis. For this dataset, 25 examples were used for each prompt and were evaluated using the macro F1 score as the performance metric.

4 Results and Discussion

In this section, we present and discuss the results obtained by our models on each dataset. The models were categorized based on the number of parameters they possess, resulting in three distinct subsections: Models around 1B parameters, Models around 3B parameters, and Models around 7B parameters. For this analysis, the best result will be displayed in gold, the second best in silver, and the third best in bronze.

4.1 Models Around 1B Parameters

In this subsection, we present an analysis of the evolution from the base model to the specific model we have developed, with an emphasis on optimization for the Portuguese language. This analysis aims to assess how our model compares with other models of similar dimensions (approximately 1 billion parameters) that have also been trained for Portuguese.

PhiBode-1.5 vs Phi-1.5 Comparison: Table 1 presents a detailed comparison between PhiBode-1.5 and its base model, Phi-1.5. Each dataset's results and final average are provided, allowing for a comprehensive evaluation of the models' performance.

Table 1. Comparison results between the baseline model and our Portuguese model of around 1 billion parameters.

Model	ENEM	BLUEX	OAB	Assin2RTE	Assin2STS	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
PhiBode-1.5	**23.58%**	20.72%	**24.87%**	**69.07%**	04.94%	43.97%	**34.94%**	41.23%	24.19%	**31.95%**
Phi-1.5	21.62%	**23.50%**	23.92%	33.33%	**13.02%**	43.97%	22.20%	41.23%	**32.88%**	28.41%

The results indicate that the PhiBode-1.5 model outperforms the Phi-1.5 model across various metrics. For instance, the PhiBode-1.5 model achieved a score of 34.94% on the dataset involving the HateBR, while the Phi-1.5 model scored 22.20%. Similarly, on the Assin2RTE dataset, the PhiBode-1.5 model reached 69.07%, whereas the Phi-1.5 model scored 33.33%. However, when considering the Assin2STS dataset, the PhiBode-1.5 model did not perform as well as the Phi-1.5 model. This discrepancy may be attributed to the training data, as some samples were artificially created. Additionally, the translation process into Portuguese may have resulted in the loss of semantic information.

Despite this, the overall findings demonstrate that the PhiBode-1.5 model is capable of surpassing the original model in several tasks, achieving an average percentage increase of more than 3% compared to the base model. This suggests that the PhiBode-1.5 model holds promise for future applications, even though there are areas where further optimization could enhance performance.

Benchmarking PhiBode-1.5 vs Other Portuguese Models: Table 2 provides a comprehensive comparison between the PhiBode-1.5 model and other Portuguese models that have approximately 1B parameters. The objective is to benchmark the performance of these models against each other.

Table 2. Comparison between Portuguese models with approximately 1B parameters.

Model	ENEM	BLUEX	OAB	Assin2RTE	Assin2STS	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
PhiBode-1.5	23.58%	20.72%	24.87%	69.07%	04.94%	43.97%	34.94%	41.23%	24.19%	31.95%
TeenyTinyLlama	20.15%	25.73%	27.02%	53.61%	13.00%	46.41%	33.59%	22.99%	17.28%	28.86%
TinyLLama1.1B	17.07%	21.56%	23.23%	43.90%	00.52%	43.97%	33.33%	42.33%	28.04%	28.22%
gp2-small-pt	19.31%	21.42%	03.14%	33.59%	03.44%	43.97%	33.33%	22.99%	13.62%	21.65%
Samba1.1b	10.22%	08.07%	15.03%	33.33%	01.30%	17.78%	35.79%	27.26%	03.27%	16.89%
Glria1.3B	01.89%	03.20%	05.19%	00.00%	02.32%	00.26%	00.28%	23.52%	00.19%	04.09%

When comparing the PhiBode-1.5 model with other models of similar size, such as the TeenyTinyLlama and TinyLLama1.1B, the PhiBode-1.5 model continues to stand out. For instance, in the HateBR dataset, the PhiBode-1.5, TeenyTinyLlama, and TinyLLama1.1B models achieved scores of 34.94%, 33.59%, and 33.33%, respectively, demonstrating that the PhiBode-1.5 model is competitive even against other models. However, it is important to note that there is still room for improvement. For example, in the BLUEX dataset, the PhiBode-1.5 model scored 20.72%, while the TeenyTinyLlama model scored 25.73%. This suggests that future iterations of the PhiBode-1.5 model could benefit from further adjustments and optimizations.

4.2 Models Around 3B Parameters

In this subsection, we embark on a comprehensive analysis of the progression from the base model to the specific model we have developed, with a distinct focus on optimization for the Portuguese language. The purpose of this analysis is to evaluate how our model stands in comparison with other models of similar dimensions (approximately 3 billion parameters) that have also been trained for Portuguese. This analysis includes models such as PhiBode-3B, PhiBode-2B, Gembode-2B-it, and Gembode-2b. This will allow us to understand where our model stands regarding performance and effectiveness relative to these models.

PhiBode and GemBode vs Base Models: Table 3 presents a detailed comparison between the PhiBode-3B and PhiBode-2B models and their base models, Phi-3B-mini and Phi-2B, respectively. These models are part of Microsoft's Phi family and have approximately 3 billion parameters. In addition, we also compare the Gembode-2B and Gembode-2B-it models and their base models, Gemma-2B and Gemma-2B-it, which are based on the Gemma family from Google. The results for each dataset are provided, along with the final average, allowing for a comprehensive evaluation of the performance of these models.

Table 3. Comparison of PhiBode and GemBode models against baseline models.

Model	ENEM	BLUEX	OAB	Assin2RTE	Assin2STS	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
PhiBode-3-3B	56.12%	40.75%	38.50%	88.56%	69.63%	50.65%	82.19%	68.10%	51.67%	60.69%
Phi-3-mini-4k	**65.22%**	**53.96%**	**45.69%**	**90.64%**	**73.60%**	**56.06%**	**84.34%**	**70.82%**	**57.40%**	**66.41%**
PhiBode-2B	**38.85%**	25.17%	**29.61%**	**45.39%**	**24.43%**	**43.97%**	54.15%	**54.59%**	**43.34%**	**39.89%**
Phi-2	34.99%	**26.98%**	28.29%	38.38%	8.87%	43.92%	**59.63%**	51.23%	36.37%	36.52%
Gembode-2B	**34.71%**	25.87%	**31.71%**	**71.37%**	34.08%	**60.09%**	47.01%	**57.04%**	49.37%	**45.69%**
Gemma-2B	26.45%	**28.37%**	28.34%	63.53%	**36.35%**	44.75%	**77.82%**	36.81%	**54.25%**	44.07%
GemBode-2B-it	21.69%	25.31%	26.83%	52.71%	**16.28%**	52.95%	**67.52%**	**24.22%**	37.54%	**36.12%**
Gemma-2B-it	**28.76%**	**25.87%**	**28.38%**	**57.17%**	5.51%	**55.20%**	44.55%	23.59%	**39.20%**	34.25%

The PhiBode-3B model performed less effectively than its base model, Phi-3B-mini, indicating that the recently released base model from Microsoft already has a strong proficiency in Portuguese, showing a significant improvement when compared to the Phi-2B, moving from an average of 36.52% to 66.41%. This also suggests that our training dataset, UltraAlpaca, may contain grammatical and syntactical errors, and its translations may not have been satisfactory. In contrast, the PhiBode-2B showed an improvement of approximately 3% compared to the Phi-2B, achieving a general average of 39.89%, demonstrating that its training was effective.

Regarding the Gemma family from Google, the overall results improved compared to the base model but did not reach a 3% improvement in the general average. This could also be related to the training data despite slightly improving overall performance.

PhiBode and GemBode vs Other Portuguese Models: Table 4 presents a comprehensive comparison between the PhiBode and GemBode models and other Portuguese models. It provides a detailed view of these models' performance across various tasks, allowing for a complete evaluation of their effectiveness and accuracy in comparison to other models.

Table 4. Comparison between Portuguese models with approximately 3B parameters.

Model	**ENEM**	**BLUEX**	**OAB**	**Assin2RTE**	**Assin2STS**	**FAQUADNLI**	**HateBR**	**PTHateSpeech**	**tweetSentBR**	**Average**
PhiBode-3-3B	56.12%	40.75%	38.50%	88.56%	69.63%	50.65%	82.19%	68.10%	51.67%	60.69%
Zephyr-3B	41.64%	36.02%	33.03%	87.57%	61.89%	67.10%	67.08%	54.25%	40.75%	54.37%
RecurrentGemma-2B-it	32.68%	31.43%	30.34%	79.42%	35.66%	28.02%	75.47%	60.20%	57.86%	47.92%
Gembode-2B	34.71%	25.87%	31.71%	71.37%	34.08%	60.09%	47.01%	57.04%	49.37%	45.69%
PhiBode-2B	38.85%	25.17%	29.61%	45.39%	24.43%	43.97%	54.15%	54.59%	43.34%	39.89%
GemBode-2B-it	21.69%	25.31%	26.83%	52.71%	16.28%	52.95%	67.52%	24.22%	37.54%	36.12%
Luana-2B	24.42%	24.34%	27.11%	70.86%	01.51%	43.97%	40.05%	51.83%	30.42%	34.95%
Periquito-3B	17.98%	21.14%	22.69%	43.01%	08.92%	43.97%	50.46%	41.19%	47.96%	33.04%
OpenCabrita-3B	17.98%	21.14%	22.69%	43.01%	08.92%	43.97%	50.46%	41.19%	47.96%	33.04%

Compared to other 3B models explicitly trained for the Portuguese language, the PhiBode-3-3B shows a significant performance improvement, beating the second-best model by a margin of 6.32%. It also greatly dominates almost all the other benchmarks, except for FAQUADNLI and tweetSentBR, in which the model was bested by Zephyr and RecurrentGemma, respectively. The PhiBode-3 model also excels in specific tasks, such as HateBR and PTHateSpeech, with a significant margin compared to other models, indicating the model's effectiveness when dealing with problems related to hate speech. Regarding the Phibode-2 model, although it can maintain a relevant result on average compared to other Portuguese-based models, it still needs to improve in terms of efficacy by a large amount to the more robust models, including the Gembode-2B.

The other developed models, Gembode-2B and GemBode-2B-it, although displaying a gap in effectiveness in comparison to the PhiBode-3 model and falling behind other models such as Zephyr and RecurrentGemma, are still able to outperform all the other Portuguese language models, including Luana-2B, which also belongs to the Gemma Family and Periquito and OpenCabrita, from the OpenLLaMA family. These models also display excellent results in the FAQUADNLI task, with Gembode-2B even surpassing the PhiBode-3 model, indicating that the proposed dataset may contribute to enhancing the model's capabilities to deal with textual entailment and question-answering scenarios.

4.3 Models Around 7B Parameters

In this subsection, we analyze the efficacy of our models of around 7 billion parameters by comparing them to their base models and other models of similar size from the literature. This comparison enables us to properly understand the capabilities and limitations of our models in various Portuguese language tasks.

GemBode X Base Model: Table 5 displays a comparison between the GemBode-7B and GemBode-7B-it models and their respective base models, Gemma-7B and Gemma-7B-it, as listed on the Open Portuguese LLM Leaderboard. It is important to note that it was impossible to perform full tuning on the 7B models, meaning the models presented are trained using PEFT.

Table 5. GemBode: comparison against baseline models.

Model	ENEM	BLUEX	OAB	Assin2RTE	Assin2STS	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
GemBode-7B	66.90%	**57.16%**	**45.47%**	**86.61%**	**71.39%**	67.40%	79.81%	**63.75%**	**65.49%**	**67.11%**
Gemma-7b	**67.04%**	56.47%	42.87%	81.34%	64.28%	**69.23%**	**85.69%**	42.51%	68.19%	64.18%
GemBode-7B-it	**49.34%**	**36.58%**	**34.76%**	79.09%	**64.95%**	**64.67%**	**86.27%**	**63.61%**	**66.17%**	**60.60%**
Gemma-7b-it	36.60%	30.32%	27.70%	**81.44%**	60.84%	57.36%	72.73%	55.99%	23.53%	49.61%

One can observe that the GemBode-7B model, when compared to its base model Gemma-7B, achieved an overall improvement of nearly 3% in the benchmark, demonstrating satisfactory performance across various datasets. The GemBode-7B-it model, on the other hand, exhibited a substantial performance increase compared to its base model Gemma-7B-it, achieving an increase of more than 10% in the general average of the benchmark. Consequently, its performance improved significantly in almost all the analyzed datasets, indicating its successful adaptation and acquisition of information in Portuguese.

Comparative Analysis of GemBode-7B with Portuguese SLMs: Table 6 provides a comparative analysis of the GemBode-7B model with other Portuguese models of equivalent size.

Table 6. Comparison between Portuguese models with approximately 7B parameters.

Model	ENEM	BLUEX	OAB	Assin2RTE	Assin2STS	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
Llama-3-8B-Dolphin	67.39%	53.82%	45.38%	92.97%	75.99%	80.17%	88.26%	58.45%	69.56%	70.22%
Internlm-ChatBode-7B	63.05%	51.46%	42.32%	91.33%	80.69%	79.80%	87.99%	68.09%	61.11%	69.54%
GemBode-7B	66.90%	57.16%	45.47%	86.61%	71.39%	67.40%	79.81%	63.75%	65.49%	67.11%
Zephyr-7B-alpha	56.26%	51.04%	40.27%	90.11%	72.33%	69.63%	85.26%	65.29%	65.49%	66.19%
Llama-3-8B	67.46%	57.30%	48.11%	89.51%	76.65%	43.97%	75.63%	62.43%	70.50%	65.73%
InternLM2-7B	60.18%	51.88%	39.86%	88.20%	81.15%	60.07%	67.98%	68.24%	66.23%	64.87%
Mistral-7B-Instruct	58.92%	52.16%	39.23%	90.52%	74.32%	66.10%	81.21%	70.26%	50.57%	64.81%
GemBode-7B-it	49.34%	36.58%	34.76%	79.09%	64.95%	64.67%	86.27%	63.61%	66.17%	60.60%
Vicuna-7B	50.59%	41.31%	36.08%	79.57%	50.76%	59.38%	69.81%	65.87%	59.89%	57.03%
Bode-7B	34.36%	28.93%	30.84%	79.83%	43.47%	67.45%	85.06%	65.73%	43.25%	53.21%
Llama-2-7b	31.91%	31.29%	35.44%	67.02%	31.10%	53.87%	75.16%	55.26%	59.06%	48.90%
Canarim-7B	25.96%	29.76%	31.48%	75.74%	12.08%	43.92%	79.57%	64.01%	66.00%	47.36%
Sabi-7B	55.07%	47.71%	41.41%	46.68%	01.89%	58.34%	61.93%	64.13%	46.64%	47.09%

The Gembode-7B, despite being only third in the overall ranking of models, still displays impressive performance, achieving a 67.11% accuracy on average, being the best-scorer in certain benchmarks such as BLUEX and OAB test scores. It is still relevant to notice that despite the top-scorer being tuned for the Portuguese language on the much more recent LLaMA-3 model (70.22%), the difference between them and our Gembode-7B model regarding the average score is only 3.11%.

An interesting analysis is that Gembode-7B can beat the average score of the base versions of the top scorers. For example, it surpasses the base Llama-3-8B's average of 65.73%, despite Llama-3-8B undergoing significant finetuning to reach its higher performance. Gembode-7B also performs better than the InternLM2-7B model, which scores an average of 64.87%.

The Gembode-7B-it model, however, despite improving the performance of the Gemma-7B-it, as shown in Table 5, and being capable of competing with certain 7B models, still lags behind the top-scorers, especially the Gembode-7B model of the same family, indicating some limitation inherited from its base model. Still, the Gembode-7B-it can outperform several models in certain tasks, such as in the case of the HateBR dataset, in which it achieves the third-best score. In the case of the tweetSentBR, on top of achieving the third-best score, it also beats the top two average scorers, indicating that this model may be more adequate for certain types of domain-specific tasks.

4.4 Full Tuning X PEFT

In addition to the training on models with approximately 3B parameters involving full tuning, we have incorporated an analysis of the training of these models using the same dataset but with a PEFT approach. This allows us to explore the extent to which these training approaches can influence the final outcome of the model. Table 7 presents the results and comparisons of outcomes between the types of training, focusing on the use of the PhiBode-2B models and the GemBode-2B and GemBode-2B-it models.

Table 7. Comparison between fine-tuned and full-tuned models. In this case, $*$ indicates full-tuning and † PEFT.

Model	ENEM	BLUEX	OAB	Assin2RTE	FAQUADNLI	HateBR	PTHateSpeech	tweetSentBR	Average
PhiBode-2B*	**38.35%**	25.17%	**29.61%**	45.39%	**43.97%**	54.15%	54.59%	43.34%	39.89%
PhiBode-2B-PEFT†	33.94%	**25.31%**	28.56%	**68.10%**	**43.97%**	**60.51**%	**54.60**%	**46.78%**	**43.59%**
GemBode-2B*	**31.77%**	**24.20%**	**27.84%**	69.51%	**55.55%**	**53.18%**	**64.74%**	**50.18%**	**45.25%**
GemBode-2B-PEFT†	24.14%	20.31%	25.56%	**69.75%**	52.63%	33.33%	41.65%	19.15%	32.30%
GemBode-2b-it*	**34.71%**	**25.87%**	**31.71%**	**71.31%**	**60.09%**	**47.01%**	**57.04%**	49.37%	**45.69%**
GemBode-2b-PEFT†	32.05%	21.56%	27.47%	33.33%	43.00%	36.41%	34.22%	**51.79%**	31.19%

The results indicate that in the case of PhiBode, its full-tuning training did not yield the expected impact, falling short compared to its PEFT training alternative. This raises the question of whether the data may not have had the desired impact and resulted in optimal performance during the model's training. In contrast, for GemBode, the results were quite significant, achieving more than a 10% increase in overall performance.

5 Conclusion and Future Works

This paper thoroughly investigates the training of Portuguese-centric Small Language Models (SLMs). A unique dataset was curated for training the models, and both full-tuning and PEFT methodologies were employed for comparative evaluation. Base models from Microsoft's Phi and Google's Gemma were used to construct our models, namely PhiBode and GemBode, which span from approximately 1 billion to 7 billion parameters. Ten models were developed in total.

Most models exhibited enhancements over their base models, except PhiBode-3B. This could be attributed to the significant advancements in the Portuguese language by its base model, Phi-3k, compared to Phi-2B. Notably, the dataset used for our training, termed UltraAlpaca, could have adversely influenced this training. The training data for Phi-3k could be more refined and devoid of translation errors, which are potential issues in our dataset.

Moreover, our dataset's size is relatively small compared to other Portuguese training datasets, which could also impact the training. However, it is crucial to highlight that we achieved superior results in all other models, with improvements ranging from 3% to over 10% of the general average of the Open Portuguese LLM Leaderboard benchmark.

Our research is a significant stride in the development of open-source models in Portuguese. We contribute a new dataset for training large models, novel training methodologies, and a comprehensive discussion of the results obtained through this training. This not only enhances the performance of full fine-tuning training compared to PEFT but also enriches the understanding and application of language model training, particularly for the Portuguese language.

Looking ahead, our research sets a promising trajectory for the field. We aspire to train a 7B parameter model with full fine-tuning, refine our dataset for even more impressive results, and release new models to the Brazilian scientific community. This will not only promote and boost Brazilian models for the scientific community but also for companies and individuals to use without additional costs. Our research significantly contributes to the field of language model training, particularly for the Portuguese language, and paves the way for exciting future advancements.

Acknowledgments. This study was funded by the São Paulo Research Foundation (FAPESP) grants 2013/07375-0, 2019/07665-4, 2023/14427-8, 2024/00789-8, and 2024/01336-7 as well as the National Council for Scientific and Technological Development (CNPq) grants 308529/2021-9 and 400756/2024-2.

References

1. Abdin, M., et al.: Phi-3 technical report: a highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Almeida, T.S., Laitz, T., Bonás, G.K., Nogueira, R.: Bluex: a benchmark based on Brazilian leading universities entrance exams (2023)
3. Beeching, E., et al.: Open LLM leaderboard (2023). https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
4. Brum, H., das Graças Volpe Nunes, M.: 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 (2018)
5. Chaudhary, S.: Code alpaca: an instruction-following llama model for code generation (2023). https://github.com/sahil280114/codealpaca
6. Corrêa, N.K., Falk, S., Fatimah, S., Sen, A., de Oliveira, N.: Teenytinyllama: open-source tiny language models trained in Brazilian Portuguese. arXiv preprint arXiv:2401.16640 (2024)
7. Delfino, P., Cuconato, B., Haeusler, E.H., Rademaker, A.: Passing the Brazilian OAB exam: data preparation and some experiments. In: Legal Knowledge and Information Systems, pp. 89–94. IOS Press (2017)

8. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
9. Ding, N., et al.: Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:2305.14233 (2023)
10. Domingues, M.: canarim-7b (2023). https://doi.org/10.57967/hf/1356
11. Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the 3rd Workshop on Abusive Language Online (ALW3), pp. 94–104. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-3510
12. Gao, L., et al.: A framework for few-shot language model evaluation (2023). https://doi.org/10.5281/zenodo.10256836
13. Garcia, E.A.S.: Open Portuguese LLM leaderboard (2024). https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard
14. Garcia, G.L., et al.: Introducing bode: a fine-tuned large language model for Portuguese prompt-based task (2024)
15. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Köpf, A., et al.: Openassistant conversations – democratizing large language model alignment (2023)
17. Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., Caridá, V.: Cabrita: closing the gap for foreign languages (2023)
18. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are all you need ii: phi-1.5 technical report (2023)
19. Minaee, S., et al.: Large language models: a survey (2024)
20. Nunes, D., Primi, R., Pires, R., Lotufo, R., Nogueira, R.: Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams (2023)
21. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al: GPT-4 technical report (2024)
22. Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese large language models. In: Naldi, M.C., Bianchi, R.A.C. (eds.) BRACIS 2023. LNCS, vol. 14197, pp. 226–240. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-45392-2_15
23. Pires, R., Almeida, T.S., Abonizio, H., Nogueira, R.: Evaluating GPT-4's vision capabilities on Brazilian university admission exams (2023)
24. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 shared task: a quick overview. In: Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., Gonçalves, T. (eds.) PROPOR 2020. LNCS (LNAI), vol. 12037, pp. 406–412. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41505-1_39
25. 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)
26. Silveira, I.C., Mauá, D.D.: University entrance exam as a guiding test for artificial intelligence. In: Proceedings of the 6th Brazilian Conference on Intelligent Systems, pp. 426–431. BRACIS (2017)
27. Singh, S., et al.: Aya dataset: an open-access collection for multilingual instruction tuning (2024)
28. Taori, R., et al.: Alpaca: a strong, replicable instruction-following model. Stanford Center Res. Found. Models **3**(6), 7 (2023)
29. Team, G., et al: Gemma: open models based on Gemini research and technology (2024)

30. Touvron, H., et al: Llama: open and efficient foundation language models (2023)
31. Vargas, F., Carvalho, I., de Góes, F.R., Pardo, T., Benevenuto, F.: HateBR: a large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 7174–7183 (2022)
32. Yu, L., et al.: Metamath: bootstrap your own mathematical questions for large language models (2024)
33. Zhao, W.X., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)