Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5812

The Bode Family of Large Language Models: Investigating the Frontiers of LLMs in Brazilian Portuguese

Pedro Henrique Paiola [**São Paulo State University** | ***pedro.paiola@unesp.br***]
Gabriel Lino Garcia [**São Paulo State University** | ***gabriel.lino@unesp.br***]
João Vitor Mariano Correia [**São Paulo State University** | ***mariano.correia@unesp.br***]
João Renato Ribeiro Manesco [**São Paulo State University** | ***joao.r.manesco@unesp.br***]
Ana Lara Alves Garcia [**São Paulo State University** | ***ana-lara.garcia@unesp.br***]
João Paulo Papa [**São Paulo State University** | ***joao.papa@unesp.br***]

Department of Computing, School of Sciences, São Paulo State University, Av. Eng. Luiz Edmundo Carrijo Coube, 14-01, Vargem Limpa, Bauru, SP, 17033-360, Brazil

Received: 01 April 2025 • **Accepted:** 17 July 2025 • **Published:** 08 October 2025

Abstract The rapid advancement of Large Language Models (LLMs) has significantly impacted Natural Language Processing, yet their effectiveness remains uneven across languages. Most state-of-the-art models are trained predominantly in English, leading to performance disparities in lower-resource languages such as Brazilian Portuguese (BP). This paper explores fine-tuning strategies for adapting open-weight LLMs to BP, focusing on dataset translation techniques, linguistic adaptation challenges, and parameter-efficient fine-tuning methods, such as LoRA and Q-LoRA. We present a benchmark analysis evaluating multiple fine-tuning approaches across various open models, establishing a guiding framework for future BP-specific adaptations. Our results showcase the importance of specialized fine-tuning in improving cross-lingual transfer and NLP performance in BP, contributing to the broader goal of enhancing multilingual language model accessibility.

Keywords: Large Language Models, Brazilian Portuguese, Natural Language Processing, Small Language Models

1 Introduction

In a remarkably short time, large language models (LLMs) have considerably advanced the field of natural language processing (NLP), enabling machines to interpret context, infer meaning, and even mimic human reasoning. These models now serve as the foundation for several applications, including machine translation, sentiment analysis, and question-answering systems [Kumar, 2024].

Yet, despite their transformative potential, a fundamental challenge remains: Most state-of-the-art LLMs are developed primarily for high-resource languages, particularly English, leading to substantial performance disparities when applied to lower-resource languages.

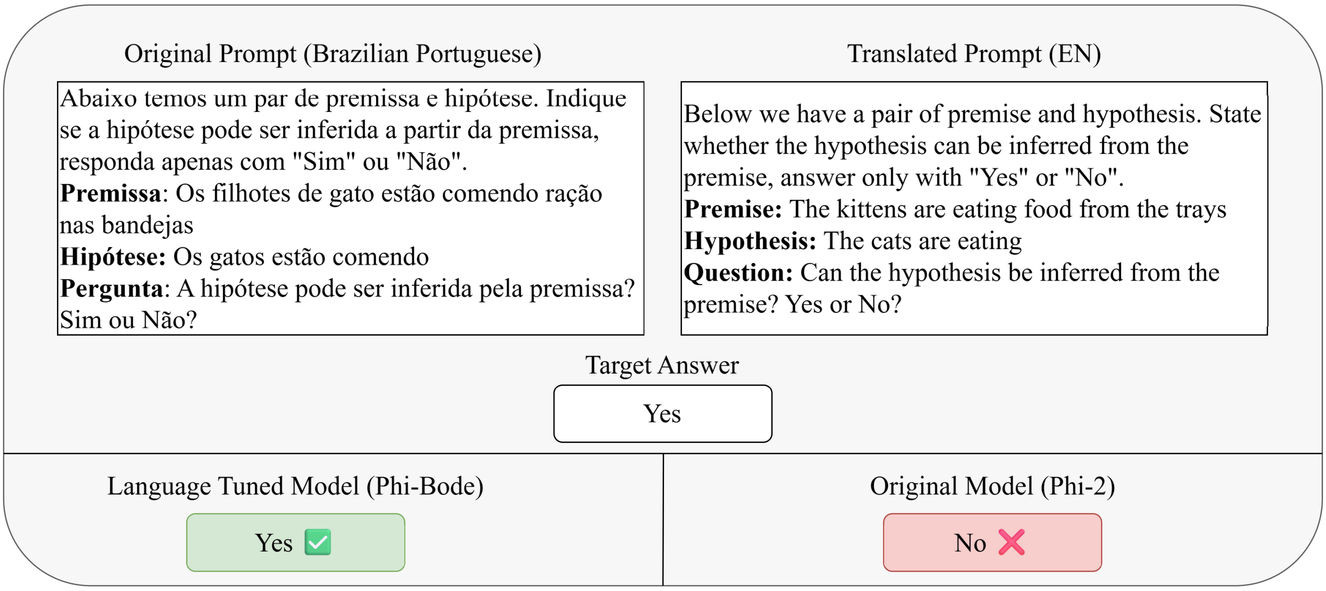
While multilingual LLMs like LLaMA [Touvron *et al.*, 2023a] and GPT [Radford *et al.*, 2018] aim to bridge this gap by approaching multiple languages, their effectiveness in non-English contexts is often constrained by insufficient linguistic specialization, cultural misalignment, and challenges in handling syntactic or pragmatic nuances [Li *et al.*, 2024]. As a result, these models exhibit performance bottlenecks in cross-lingual transfer, leading to errors such as mistranslations, hallucinations, and culturally insensitive outputs. In particular, Brazilian Portuguese (BP), a language spoken by over 200 million people, presents a significant case where existing models, trained predominantly on English-centric datasets, struggle to generalize effectively.

Thus, prior efforts to address BP-specific NLP challenges have arisen in the literature in the form of models such as BERTimbau [Souza *et al.*, 2020], Sabiá [Pires *et al.*, 2023], TeenyTinyLLaMa [Corrêa *et al.*, 2024], among several others that have improved performance in tasks such as named entity recognition, text classification and text generation, posing finetuning for specific languages as a crucial step towards mitigating language disparities in NLP. This process enables models to capture language-specific structures, idiomatic expressions, and domain-specific variations that are often overlooked in broadly trained multilingual systems.

However, existing studies lack systematic evaluations across diverse LLM architectures and fail to establish standardized benchmarks for effective multilingual transfer. Furthermore, many models prioritize performance in high-resource computing environments, which are not often available, limiting practical deployment in resource-constrained scenarios.

When dealing with real-world applications, where models are required to adapt for proper linguistic nuances, but also to learn proper domain-specific knowledge, fine-tuning and finding out the effectiveness of proper architectures are paramount. In the legal sector, for example, precise language understanding is essential for contract analysis and regulatory compliance, where misinterpretations can have significant consequences [Lai *et al.*, 2024]. The same happens in healthcare, where models trained on English-centric datasets often fail to accurately interpret medical records or patient interactions in BP, leading to potential risks in decision support systems [Zhou *et al.*, 2023].

As Brazilian Portuguese presents unique linguistic and cultural challenges that require careful consideration in model fine-tuning, such as complex verb conjugations, flexible word



Original Prompt (Brazilian Portuguese)
Translated Prompt (EN)
Target Answer
Yes
Language Tuned Model (Phi-Bode)
Yes ✓
Original Model (Phi-2)
No ✗

Abaixo temos um par de premissa e hipótese. Indique se a hipótese pode ser inferida a partir da premissa, responda apenas com "Sim" ou "Não".
Premissa: Os filhotes de gato estão comendo ração nas bandejas
Hipótese: Os gatos estão comendo
Pergunta: A hipótese pode ser inferida pela premissa? Sim ou Não?

Below we have a pair of premise and hypothesis. State whether the hypothesis can be inferred from the premise, answer only with "Yes" or "No".
Premise: The kittens are eating food from the trays
Hypothesis: The cats are eating
Question: Can the hypothesis be inferred from the premise? Yes or No?

Figure 1. Example from the ASSIN2 RTE dataset, we put an English translation of the prompt for discussion purposes. In this case, the SLM model trained in the English language fails to answer the question correctly. Meanwhile, our proposed model, Phi-Bode, which tunes the same model to the Brazilian Portuguese language domain, can infer the answer correctly.

order, and regional dialectal variations that influence model predictions, on top of, idiomatic expressions and cultural references embedded in BP. Models trained primarily on English corpora can often misinterpret information. For example, phrases with figurative meanings, such as ”ficar na mão” (literally “to stay in the hand”, meaning ”to be left stranded”), may be misunderstood if not explicitly learned from BP texts. Cultural misalignment also manifests in bias-related issues, where models might generate responses that are insensitive to socio-linguistic nuances in Brazil. Addressing these challenges through targeted fine-tuning enhances the reliability and inclusivity of NLP applications.

A tangible example of the limitations in existing models is illustrated in Figure 1, which presents a comparison between the Phi-2 model and our fine-tuned variant, Phi-Bode. In this example, the original Phi-2 model fails to correctly answer a question from the ASSIN2 RTE dataset [Real *et al.*, 2020], whereas our BP-optimized model successfully interprets and responds accurately, highlighting the impact of language-specific adaptation.

To tackle these challenges, we introduce Bode, a family of foundation language models fine-tuned for Brazilian Portuguese. Our work makes three key contributions:

- We establish the first comprehensive evaluation framework for BP-optimized LLMs, spanning multiple architectures (e.g., LLaMA, Phi-2, Mistral) and tasks.
- We demonstrate how targeted fine-tuning addresses BP’s unique linguistic features, reducing errors in language interpretation.
- While previous models for BP exist, our systematic analysis and benchmarking effort serve as a reference point for future research, ensuring more effective adaptation strategies and broader accessibility of BP-specific NLP models.

The remainder of this paper is organized as follows: Section 2 discusses prior work on multilingual and language-specific model adaptation. Section 3 provides the theoretical foundations of fine-tuning approaches Sections 4 and 5 detail the methodology regarding the fine-tuning approach for Bode as well as architectural details in each model within the family. Section 6 presents experimental evaluations and performance benchmarks. Finally, Section 8 summarizes our contributions and outlines future research directions.

2 Related Works

Recent advances in Brazilian Portuguese language models have prioritized task-specific performance over architectural exploration, resulting in a fragmented landscape of models tied to singular frameworks. Below, we review key contributions and their limitations in the area.

Early efforts focused on training models from scratch using Portuguese corpora. GlórIA, a GPT-3-style model for European Portuguese, exemplifies this approach. Trained on 35B tokens from web texts and news articles, it targeted tasks like summarization and dialogue generation. However, its performance suffered from an inadequate pretraining scale (limited to 350B tokens vs. GPT-3’s 300B+) and aggressive data filtering, which stripped crucial linguistic diversity. This led to underfitting, particularly for morphologically complex tasks like verb conjugation [1]. The development of LLMs for the Portuguese language has progressed significantly in recent years, with models that aim to improve language understanding, fluency, and cultural relevance while addressing the limitations of English-centric LLMs [Lopes *et al.*, 2024].

The Sabiá family [Pires *et al.*, 2023] addressed these issues by combining continued pretraining of multilingual architectures (GPT-J, LLaMA) with 10.4B high-quality Portuguese tokens. By retaining cross-lingual knowledge while specializing for Portuguese, Sabiá achieved state-of-the-art results in syntactic tasks (e.g., dependency parsing) but showed limited cultural adaptation, as its training data lacked region-specific idioms and dialectal variations.

In order to enhance multimodal capabilities, Cabrita [Larcher *et al.*, 2023] adopted parameter-efficient fine-tuning (PEFT) on OpenLLaMA-3B using Portuguese translations of the Alpaca dataset. While this enabled competitive few-shot performance, its reliance on translated synthetic data introduced semantic distortions, particularly for figurative language (e.g., metaphors like "água mole em pedra dura"). Cabrita's 3B parameter size also constrained its ability to capture long-range dependencies in legal and medical texts.

Aiming to solve the computational constraints in Large Language Models, TeenyTinyLLaMa [Corrêa *et al.*, 2024] was proposed as a minimalist 460M-parameter model under Apache 2.0, and it demonstrated that scaled-down architectures can retain 65% of LLaMA-7B's downstream performance. Yet, its generic pretraining and lack of BP-specific tuning led to poor cultural alignment.

We can see from the literature that existing BP models evaluated in academic settings exhibit two systemic limitations: (i) most models derive from decoder-only LLaMA variants, and (ii) prior work focused more on generic BP adaptation and neglected domain-specific tuning (e.g., medical).

Thus, we propose the Bode family of models in order to systematically dissect how architectural choices themselves shape performance of finetuning in Brazilian Portuguese. While the original LLaMA-2-based variant [Garcia *et al.*, 2024] demonstrated the viability of Portuguese instruction tuning, subsequent improved the analysis of the LLM. Studies quantized its layers [Jodas *et al.*, 2024], shrunk it into SLMs [Garcia *et al.*, 2025], and specialized it for medicine [Paiola *et al.*, 2024], but always in isolation. This work aims to offer a compilation of models, probing how different architectures impact in BP understanding.

3 Theoretical Foundation

Large Language Models have revolutionized natural language processing through their ability to learn intricate linguistic patterns from vast text corpora. These models are typically based on transformer architectures [Vaswani *et al.*, 2017], which employ self-attention mechanisms to capture contextual relationships between words. Despite their effectiveness, the computational demands of training and fine-tuning LLMs present significant challenges, particularly for resource-limited languages like Brazilian Portuguese.

LLMs can be broadly categorized into different families based on their architecture and training paradigm. Autoregressive models, such as GPT [Radford *et al.*, 2018], predict tokens sequentially and excel in text generation tasks. Encoder-decoder models, like T5 [Raffel *et al.*, 2020], process input sequences into a latent representation before generating outputs, making them well-suited for translation and summarization. Multilingual models like LLaMA [Touvron *et al.*, 2023a] are trained on data from multiple languages but often underperform in low-resource settings due to uneven corpus distributions.

One of the fundamental challenges in adapting LLMs to new languages is the availability of high-quality datasets. Dataset translation serves as a means of augmenting training data by leveraging parallel corpora or synthetic translation techniques. However, direct translation often introduces noise and fails to preserve linguistic and cultural nuances [Li *et al.*, 2024]. A more effective approach involves fine-tuning models using domain-specific and monolingual corpora, ensuring that syntactic and semantic structures are accurately captured.

In the case of Brazilian Portuguese, a critical issue in dataset translation arises from the morphological richness and syntactic flexibility of the language. Simple word-to-word translation often leads to unnatural phrasing, loss of idiomatic expressions, and incorrect verb conjugations. To mitigate these issues, hybrid approaches combining human-verified parallel corpora with neural machine translation systems can improve data quality. Additionally, unsupervised and semi-supervised learning methods, such as self-training, where a model iteratively improves by training on its own predictionsDu *et al.* [2021], and back-translation, where monolingual data is used to create synthetic training pairsSennrich *et al.* [2016] help generate more contextually accurate datasets. These methods enhance model adaptation by refining lexical choices and preserving domain-specific terminologies.

3.1 Parameter Efficient Fine-Tuning

Fine-tuning Large Language Models on new languages can be computationally prohibitive. Parameter-efficient tuning techniques such as Low-Rank Adaptation (LoRA) [Hu *et al.*, 2021] and adapters mitigate this issue by modifying a smaller subset of parameters while keeping the majority of the pre-trained model unchanged.

LoRA reduces the number of trainable parameters by decomposing weight updates into low-rank matrices. For a given pre-trained weight matrix $W \in \mathbb{R}^{d \times k}$, where d and k represent the output and input dimensions of the layer, LoRA freezes W and introduces two smaller, trainable matrices: $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ Here, the hyperparameter r is the rank of the adaptation, which is much smaller than d and k. This rank determines the expressiveness and the number of trainable parameters in the LoRA layers. The adapted weights are then expressed as:

$$W' = W + AB$$

where $r \ll \min(d, k)$, significantly reducing the number of trainable parameters while maintaining performance.

Q-LoRA [Dettmers *et al.*, 2023] extends LoRA by integrating quantization techniques, enabling fine-tuning of LLMs under extreme memory constraints. By quantizing model weights into low-bit representations, Q-LoRA minimizes memory footprint while preserving model adaptability. This method is particularly advantageous for deploying LLMs in low-resource environments with limited computational power. These methods are capable of democratizing different types of adaptation.

For Brazilian Portuguese, parameter-efficient fine-tuning strategies are particularly relevant given the scarcity of large-scale, high-quality datasets. By employing LoRA, Q-LoRA, and adapters, it becomes feasible to efficiently adapt LLMs without requiring extensive computational resources.

4 Methodology and Datasets

A significant challenge in developing instruction-tuned models for Brazilian Portuguese is the scarcity of large-scale, high-quality, native instruction datasets. To overcome this, we adopt a hybrid approach that combines translated versions of established English datasets with existing native Portuguese, primarily using OpenAI's GPT-3.5 model. We acknowledge that potential biases might be inherited from the translation model, and that direct translation can introduce artifacts and fail to capture subtle linguistic and cultural nuances, the very issues this work aims to mitigate. However, this strategy is a pragmatic necessity to obtain the volume and diversity of instruction-following examples required for robust model training. By integrating these translated resources with authentic Portuguese data, we aim to balance the breadth of task coverage from the former with the linguistic and cultural fidelity of the latter. This section details the datasets used, including the translation procedures employed.

Our work leverages two primary datasets: Alpaca [1] and UltraAlpaca [2]. The former is a well-established dataset in the field of instruction-following models that was translated to Portuguese, while the latter is a novel dataset proposed by [Garcia *et al.*, 2025], designed to enhance the performance of language models in Portuguese. This section provides a comprehensive review of the datasets, their composition, and their relevance to the broader research landscape.

4.1 Alpaca

The Alpaca dataset, introduced by Stanford's Center for Research on Foundation Models (CRFM), is a widely recognized resource in the field of instruction-following models. It consists of 52,000 instruction-following samples generated by OpenAI's text-davinci-003 engine, a model from the GPT-3 family. Alpaca has been instrumental in advancing research on fine-tuning LLMs for task-specific applications. In our study, we utilize a translated version of Alpaca, which was previously employed in training the first version of Cabrita, a Portuguese-focused language model. The dataset's structured and diverse instruction set provides a robust foundation for training models to generalize across a wide range of tasks.

4.2 UltraAlpaca

UltraAlpaca is a dataset specifically designed to address the limitations of existing datasets for Portuguese language models. It is composed of multiple high-quality datasets, each contributing unique strengths to the final corpus. Below, we detail the datasets integrated into UltraAlpaca:

- **Alpaca:** As mentioned earlier, we incorporated the translated version of Alpaca, which includes 52,000 instruction-following samples. This dataset serves as the backbone of UltraAlpaca, providing a strong foundation of task-specific instructions.
- **UltraChat:** UltraChat [Ding *et al.*, 2023] is a self-refinement dataset comprising 1.47 million multi-turn dialogues generated by GPT-3.5-TURBO. These dialogues span 30 topics and 20 distinct types of text material. From this dataset, we selected and translated 70,000 samples to enrich UltraAlpaca with diverse conversational contexts.
- **Aya:** The Aya dataset [Singh *et al.*, 2024] is a multilingual instruction fine-tuning dataset curated by the 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 UltraAlpaca, we filtered and included only the Portuguese samples, ensuring linguistic relevance.
- **OpenAssistant Conversations (OASST1):** OASST1 [Köpf *et al.*, 2023] multilingual corpus includes 161,443 assistant-style messages in 35 languages, annotated with 461,292 quality ratings. The dataset was created through a global crowd-sourcing effort involving over 13,500 volunteers. We filtered the Portuguese samples from this dataset to enhance UltraAlpaca's conversational diversity.
- **Code Alpaca:** Code Alpaca [Chaudhary, 2023] is a dataset of 20,000 samples focused on code generation, built similarly to Alpaca. We fully translated this dataset to integrate it into UltraAlpaca, thereby incorporating programming-related tasks into the training corpus.
- **MetaMathQA-40K-PTBR:** This specialized dataset [Yu *et al.*, 2024] is designed to enhance mathematical reasoning capabilities in LLMs. It comprises 395,000 samples, of which we utilized 40,000 previously translated samples to bolster UltraAlpaca's mathematical reasoning components.

5 Bode Family

The Bode family of models represents a collection of open-source language models specifically fine-tuned for the Portuguese language. Built on state-of-the-art base models such as LLaMa, Gemma, Phi, Mistral, InternLM, Qwen, and Zephyr, the Bode models leverage advanced fine-tuning techniques like LoRA (Low-Rank Adaptation) and QLoRA (Quantized LoRA) to optimize performance for Portuguese-language tasks. This section details the diverse, state-of-the-art base models (LLaMa, Gemma, etc.) that serve as the foundation for the Bode family, followed by a comprehensive overview of the resulting fine-tuned Portuguese models.

5.1 LLaMa

LLaMa is a series of open-source large language models developed by Meta AI, available in various versions and sizes. These models are designed to achieve state-of-the-art performance while maintaining computational efficiency, democratizing access to NLP capabilities through publicly available datasets. The LLaMa family has evolved through several iterations, each introducing architectural improvements, enhanced training methodologies, and superior performance compared to its predecessors.

The original LLaMa 1 [Touvron *et al.*, 2023a] model is

[1] https://huggingface.co/datasets/tatsu-lab/alpaca

[2] https://huggingface.co/datasets/recogna-nlp/ultra-alpaca-ptbr

based on the transformer architecture [Vaswani *et al.*, 2017], incorporating optimizations such as RMSNorm [Zhang and Sennrich, 2019] for pre-normalization, the SwiGLU [Shazeer, 2020] activation function, and rotary positional embeddings (RoPE) [Su *et al.*, 2023]. These enhancements improve training stability and model performance. LLaMa 1 was trained on 1.4 trillion tokens of publicly available data, including sources like CommonCrawl, C4, GitHub, and Wikipedia, ensuring broad language understanding and knowledge representation.

LLaMa 2 [Touvron *et al.*, 2023b] builds on LLaMa 1, introducing architectural improvements such as grouped-query attention (GQA) [Ainslie *et al.*, 2023], which reduces memory usage and speeds up inference by grouping queries during attention computation. The training dataset was expanded to 2 trillion tokens (40% larger than LLaMa 1), and the context length was doubled, enabling the model to handle longer sequences more effectively. LLaMa 2 also emphasizes safety and alignment through Reinforcement Learning with Human Feedback (RLHF), with chat variants fine-tuned specifically for dialogue use cases. According to its authors, these improvements are intended to enhance the model's safety and utility in dialogue-based applications.

The latest release, LLaMa 3 [Grattafiori *et al.*, 2024], retains the transformer-based architecture of its predecessors but introduces significant optimizations in model scale and training efficiency. The largest LLaMa 3 model scales up to 405 billion parameters, a substantial increase from the 70 billion parameters in LLaMa 2. To manage this scale, LLaMa 3 is trained on a dataset of 15.6 trillion tokens and employs advanced training techniques such as supervised fine-tuning (SFT), rejection sampling (RS), and direct preference optimization (DPO) [Rafailov *et al.*, 2023]. These methods improve the model's alignment with human preferences and enhance its ability to generate high-quality, contextually appropriate responses.

5.2 Gemma

Gemma is a family of lightweight, state-of-the-art open models developed by Google DeepMind. These models are designed to achieve strong performance across a wide range of language understanding, reasoning, and safety benchmarks while maintaining computational efficiency. Gemma models are available in two sizes: 2 billion and 7 billion parameters, with pre-trained and instruction-tuned checkpoints released.

The models are trained on 3 trillion tokens (2B parameters) and 6 trillion tokens (7B parameters) of primarily English data, including web documents, mathematics, and code. They are not multimodal and are not expected to perform well on multilingual tasks. The model architecture is based on the transformer decoder, with several improvements such as multi-query attention [Raffel *et al.*, 2020], RoPE, GeGLU activation functions [Shazeer, 2020], and RMSNorm.

It is noteworthy that these models were trained on datasets comprising 3 trillion (2B model) and 6 trillion (7B model) tokens of primarily English data, including web documents, mathematics, and code. Consequently, while powerful, their inherent exposure to Portuguese during pre-training is limited, making fine-tuning crucial for optimal performance in the language.

5.3 Phi

The Phi family of models, developed by Microsoft Research, is a series of small language models (SLMs) designed to challenge conventional scaling laws of LLMs by focusing on high-quality, curated training data. These models achieve competitive performance across a variety of tasks despite their compact size. The key distinction of the Phi models lies in their reliance on synthetic and filtered datasets, which enable them to match or surpass the capabilities of much larger models while reducing computational and environmental costs [Bender *et al.*, 2021].

The first model in the series, Phi-1 [Gunasekar *et al.*, 2023], introduced the concept of using textbook-quality data to train a 1.3 billion-parameter model for code generation. Phi-1's architecture is a conventional decoder-only transformer, with the major innovation lying in its training data, which was carefully curated to include clear, self-contained, and instructive examples, both synthetically generated and filtered from web sources. By focusing on data quality, Phi-1 achieves remarkable performance on Python coding benchmarks despite being orders of magnitude smaller than competing models.

Building on Phi-1, Phi-1.5 expanded the scope to commonsense reasoning, a more challenging task for NLP [Li *et al.*, 2023]. With the same 1.3 billion-parameter architecture, Phi-1.5 was trained on a combination of Phi-1's data and a new synthetic dataset designed to teach general knowledge and reasoning. This model demonstrated performance comparable to models ten times its size, highlighting the effectiveness of synthetic data in addressing complex reasoning tasks.

Phi-2 marked a significant leap in the series, scaling up to 2.7 billion parameters while maintaining the focus on training data quality. Trained on 1.4 trillion tokens from a mix of synthetic and web datasets, Phi-2 showcased state-of-the-art reasoning and language understanding capabilities, outperforming models up to 25 times larger [Hughes, 2023]. Unlike its predecessors, Phi-2 was designed as a base model without alignment or fine-tuning, yet it exhibited better behavior regarding toxicity and bias, thanks to its tailored data curation techniques.

The latest iteration, Phi-3, introduced two models: Phi-3-mini (3.8 billion parameters) and Phi-3-small (7 billion parameters). These models further advanced the data quality approach, training on larger and more refined datasets to achieve performance comparable to models like GPT-3.5 and Mixtral 8x7B [Jiang *et al.*, 2024]. Phi-3-mini featured innovations like LongRope [Ding *et al.*, 2024] for extended context lengths and a LLaMa-2-compatible architecture, making it accessible for local inference on devices like smartphones. Phi-3-small introduced novel techniques such as blocksparse attention and GeGLU activation, optimizing both training and inference efficiency.

5.4 Mistral

Mistral 7B is a 7-billion-parameter language model designed to balance performance and efficiency, leveraging architectural innovations to achieve competitive results across a range of benchmarks. The model is based on a transformer architecture but introduces key modifications, such as grouped-query attention (GQA) and sliding window attention (SWA) [Child *et al.*, 2019; Beltagy *et al.*, 2020], which optimize inference speed and memory usage. GQA reduces memory requirements during decoding, enabling higher batch sizes and throughput, while SWA allows the model to handle longer sequences more efficiently by limiting the attention span to a fixed window size. These mechanisms collectively reduce the quadratic complexity of traditional attention, improving both computational efficiency and scalability [Jiang *et al.*, 2023].

The model was trained on a diverse dataset, though the paper does not provide extensive details on its composition. The training process leveraged distributed computing, and the model was fine-tuned for specific tasks, such as instruction following, using publicly available datasets. The fine-tuned version, Mistral 7B - Instruct, demonstrates strong performance on instruction-based benchmarks, outperforming comparable models like LLaMa 2 7B and 13B in human evaluations and automated metrics.

Mistral 7B also incorporates system prompts for enforcing guardrails in front-facing applications, ensuring that generated content adheres to ethical and safety guidelines. The model can perform content moderation through self-reflection, accurately classifying prompts or generated responses as acceptable or falling into categories such as illegal activities, hateful content, or unqualified advice. This feature is particularly useful for applications requiring strict content control, such as social media moderation or brand monitoring.

5.5 InternLM

The InternLM series represents a significant step forward in the development of open-source large language models. Despite the open-source community's enthusiasm, bridging the gap between proprietary LLMs like ChatGPT or GPT-4 and their open-source counterparts remains challenging [InternLM Team, 2023]. InternLM addresses this challenge through a combination of architectural innovations, a multi-stage data preparation pipeline, and novel alignment techniques, making it a competitive model in the LLM landscape.

InternLM builds on the transformer architecture, incorporating several optimizations to enhance efficiency and performance. Following the design principles of LLaMA, InternLM2 replaces LayerNorm with RMSNorm and uses the SwiGLU activation function, which improves training stability and model performance. To ensure compatibility with the well-established LLaMA ecosystem, InternLM adheres to LLaMA's structural design while introducing several enhancements.

The quality of data is considered the most crucial factor during pre-training. While technical reports on LLMs often overlook data processing details, InternLM provides extensive documentation on its data preparation pipeline. The pre-training dataset includes text, code, and long-context data, curated and filtered to ensure high quality, totaling 2.0T to 2.6T tokens, depending on the model size (1.8B, 7B, or 20B parameters). The model also employs positional encoding extrapolation to further extend its context window beyond 32k tokens.

Following pre-training, the model undergoes SFT and reinforcement learning from human feedback to ensure alignment with human instructions and values. Notably, the RLHF process includes the construction of 32k context data to further enhance the model's long-context processing capabilities. A key feature is the Conditional Online Reinforcement Learning from Human Feedback (COOL RLHF). Traditional RLHF methods often struggle with preference conflicts and reward hacking, where the model exploits the reward system rather than genuinely aligning with human preferences.

Since InternLM2 was pre-trained on a diverse corpus that included medical data, it provided a strong foundational knowledge base for biomedical and healthcare-related tasks. This made it an ideal candidate for further specialization into a medical domain-specific model, named doutor-bode [Paiola *et al.*, 2024].

5.6 Qwen

The Qwen model, developed by Alibaba Group, focuses on model accessibility and reproducibility while bridging the gap between proprietary models and open-source alternatives. The architecture is based on the LLaMa model but proposes modifications such as embeddings and output projection, RoPE for positional encoding, removal of bias for most layers to enhance extrapolation ability, SwiGLU activation functions, and RMSNorm for layer normalization.

The models are pretrained on a massive dataset of up to 3 trillion tokens, encompassing diverse text and code from public web documents, encyclopedias, books, and multilingual sources. The dataset is carefully curated and filtered to ensure high quality. Qwen uses Byte Pair Encoding (BPE) [Jain, 2022] with a vocabulary size of approximately 152K, optimized for multilingual tasks, particularly in Chinese. The tokenizer achieves high compression efficiency, reducing serving costs while maintaining performance.

Qwen-Chat models are fine-tuned on curated datasets that include human-style conversations, tool use, and safety-related annotations. The ChatML format is used to distinguish between system, user, and assistant inputs, improving the model's ability to handle complex conversational data. The models also undergo RLHF to align them with human preferences. A reward model is trained on high-quality comparison data, and Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] is used to fine-tune the models further. This process enhances the model's ability to generate responses that are more aligned with human expectations.

5.7 Zephyr

The Zephyr [Tunstall *et al.*, 2023] model focuses on aligning smaller, open-source LLMs with user intent, leveraging distillation techniques to achieve competitive performance without extensive human feedback. The model is based on Mistral-7B and proposes a distilled supervised fine-tuning (dSFT) step, where the model is trained on high-quality instruction-response pairs generated by a more capable teacher model. This step ensures that the model learns to respond to user prompts effectively. The dataset used for dSFT is UltraChat [Ding *et al.*, 2023], which consists of multi-turn dialogues generated by GPT-3.5-TURBO. The dataset is refined to remove unhelpful responses and grammatical errors, resulting in a high-quality training set.

The model employs AI feedback from an ensemble of teacher models to collect preference data. This feedback is used to rank model outputs based on criteria like helpfulness, honesty, and instruction-following. The model then undergoes distilled direct preference optimization (dDPO) [Rafailov *et al.*, 2023], a method that directly optimizes the model to prefer higher-ranked responses over lower-ranked ones. This approach avoids the need for sampling during fine-tuning, making the process more efficient.

5.8 Overview of Bode Variants

Building upon the diverse base architectures previously described, the Bode family currently encompasses 35 distinct models specifically fine-tuned for Brazilian Portuguese (as detailed in Table 1). These variants utilize different fine-tuning datasets (like Alpaca or UltraAlpaca variations) and techniques. The table below summarizes key specifications for each Bode model, including its base architecture, parameter count, fine-tuning dataset, adapter state, and format. Regarding the 'Adapter State' column: models listed with 'LoRA' or 'QLoRA' retain separate adapter weights, while those marked 'Merged' have had the adapter weights fused into the base model weights.

6 Experimental Setup

6.1 Fine-tuning Configuration

To enhance reproducibility, this section details the hardware and software configurations for the fine-tuning experiments.

- **Hardware:** The fine-tuning processes were primarily conducted on a high-performance computing (HPC) infrastructure with a processing capacity of 5.1 PetaFLOPS. The specific compute nodes used were of the Bull Sequana X1120 compute blade, each equipped with 4 NVIDIA Volta V100 GPUs, 384GB of RAM, and 2 Intel Xeon Skylake 6252 CPUs. Inference tasks for evaluation were performed on the Kaggle platform, which provided an environment with an NVIDIA K80 GPU.
- **Hyperparameters:** The fine-tuning scripts were largely based on the xtuner [XTuner Contributors, 2023] library developed by the InternLM team. As precise logs for all models were not retained, we present representative hyperparameters. For many models, the default parameters of the xtuner library were used, with primary adjustments made to the learning rate depending on whether a full or parameter-efficient (LoRA/Q-LoRA) tuning method was employed.
  - Training Schedule: Models were typically trained for a single epoch. The effective batch size was 16, achieved using a per-device batch size of 1 with 16 gradient accumulation steps.
  - Optimizer: The AdamW optimizer was used with beta values of (0.9, 0.999) and no weight decay.
  - Learning Rate: This was the most significant variable, adjusted based on the tuning method. A learning rate of $2e^{-4}$ was generally used for parameter-efficient techniques like LoRA/Q-LoRA. For full fine-tuning, a lower learning rate of $2e^{-5}$ was often applied. Both were scenarios were paired with a warmup ratio of 0.03.
  - Regularization: Gradient clipping was employed with a max norm of 1.
  - Input Handling: The maximum sequence length ranged from 512 to a max of 2048 on more recent models, and inputs were packed to this maximum length to improve training efficiency.

6.2 Evaluation Benchmarks

Model evaluation was conducted using the specific configurations implemented on the **Open PT-LLM Leaderboard** Garcia [2024]. This leaderboard utilizes the **Eleuther AI Language Model Evaluation Harness** [3] to standardize testing across various benchmarks. The datasets and their specific settings for the leaderboard evaluation are detailed below:

- **ENEM [Nunes *et al.*, 2023]:** This datasets comprises 1,430 questions from exams between 2019 and 2018 as well as 2022 and 2023, excluding questions that require Image Comprehension (IC), Mathematical Reasoning (MR), or interpretation of Chemical Elements (CE). Each question comprises a header (context), a statement (question), and five multiple-choice alternatives. For the leaderboard, models are evaluated using a **3-shot** setting on the `enem_challenge` task, with **accuracy (acc)** as the metric.
- **BLUEX [Almeida *et al.*, 2023]:** This dataset consists of over 1,000 multiple-choice questions sourced from the 2018-2023 entrance examinations of two prominent Brazilian universities, Unicamp and USP. Questions, alternatives, and images were automatically extracted and subsequently manually verified and annotated with metadata. Key metadata tags include indications of required Prior Knowledge (PRK), Text Understanding (TU), Image Understanding (IU), Mathematical Reasoning (MR), Multilingualism (ML), and Brazilian-specific Knowledge (BK), alongside subject classification and image positioning information. For the evaluation of the models were select 724 questions that do not necessitate image understanding in a **3-shot** setting via the `bluex` task, measured by **accuracy (acc)**.
- **OAB Exams [Delfino Pedro *et al.*, 2017]:** This benchmark utilizes questions from the first phase of the unified Order of Attorneys of Brazil (OAB) exams, administered nationally since 2010. The dataset comprises 1,820 multiple-choice questions (each with four options) extracted from 22 exams obtained in PDF format and converted to text using Apache Tika. The questions cover various areas of law, with a notable emphasis on Ethics, often based on specific legal statutes, the OAB General Regulation, and the OAB Ethics Code. Evalua-

[3] `https://github.com/EleutherAI/lm-evaluation-harness`

Table 1. Specifications of Fine-tuned Bode Models for Portuguese.

Model Name	**Base Model**	**Parameters**	**Dataset**	**Adapter State**	**Format**
bode-7b-alpaca-pt-br	LLaMa 2	7B	Alpaca	LoRA	Standard
bode-7b-alpaca-pt-br-no-peft	LLaMa 2	7B	Alpaca	LoRA + Merged	Standard
bode-7b-alpaca-pt-br-gguf	LLaMa 2	7B	Alpaca	LoRA + Merged	GGUF
bode-13b-alpaca-pt-br	LLaMa 2	13B	Alpaca	LoRA	Standard
bode-13b-alpaca-pt-br-no-peft	LLaMa 2	13B	Alpaca	LoRA + Merged	Standard
Bode-3.1-8B-Instruct-lora	LLaMa 3	8B	Alpaca	LoRA	Standard
Bode-3.1-8B-Instruct-full	LLaMa 3	8B	Alpaca	Full	Standard
GemBode-2b-it	Gemma-2b-it	2B	Alpaca	QLoRA	Standard
gembode-2b-ultraalpaca	Gemma-2b	2B	UltraAlpaca	Full	Standard
gembode-2b-ultraalpaca-qlora	Gemma-2b	2B	UltraAlpaca	QLoRA	Standard
gembode-2b-it-ultraalpaca	Gemma-2b-it	2B	UltraAlpaca	Full	Standard
gembode-2b-it-ultraalpaca-qlora	Gemma-2b-it	2B	UltraAlpaca	QLoRA	Standard
gembode-7b	Gemma-7b	7B	UltraAlpaca	QLoRA	Standard
gembode-7b-it	Gemma-7b-it	7B	UltraAlpaca	QLoRA	Standard
phibode_1_5_ultraalpaca	Phi-1.5B	1.3B	UltraAlpaca	Full	Standard
phibode_1_5_ultraalpaca_qlora	Phi-1.5B	1.3B	UltraAlpaca	QLoRA	Standard
Phi-Bode	Phi-2B	2.7B	Alpaca	QLoRA	Standard
phi-bode-2-ultraalpaca	Phi-2B	2.7B	UltraAlpaca	Full	Standard
phibode-3-mini-4k-ultraalpaca	Phi-3-mini-4k-instruct	3.8B	UltraAlpaca	LoRA	Standard
mistral-bode	Mistral-7B	7B	Alpaca	LoRA	Standard
mistralbode_7b_lora_ultraalpaca	Mistral-7B	7B	UltraAlpaca	LoRA	Standard
mistralbode_7b_qlora_ultraalpaca	Mistral-7B	7B	UltraAlpaca	QLoRA	Standard
zephyr_7b_beta_ultraalpaca	Zephyr	7B	UltraAlpaca	QLoRA	Standard
internlmbode-7b	InternLM2	7B	UltraAlpaca	QLoRA	Standard
internlm-chatbode-7b	InternLM2 Chat	7B	UltraAlpaca	QLoRA + Merged	Standard
internlm-chatbode-20b	InternLM2 Chat	20B	UltraAlpaca	QLoRA + Merged	Standard
internlm-chatbode-20b-gguf	InternLM2 Chat	20B	UltraAlpaca	QLoRA + Merged	GGUF
internlm-chatbode2-7b-alpha	InternLM2 Chat	7B	UltraAlpaca	QLoRA	Standard
internlm-chatbode2-7b-beta	InternLM2 Chat	7B	UltraAlpaca	QLoRA	Standard
internlm2-chat-7b-ultraalpaca	InternLM2 Chat	7B	UltraAlpaca	QLoRA	Standard
internlm2-chat-1_8b-ultracabrita	InternLM2 Chat 1_8	1.8B	UltraCabrita	QLoRA	Standard
doutor-bode-7b-240k	InternLM2 Chat	7B	MedQuAD	QLoRA	Standard
doutor-bode-7b-360k	InternLM2 Chat	7B	MedQuAD	QLoRA	Standard
qwenbode_1_8b_chat_ultraalpaca	Qwen-1.8B Chat	1.8B	UltraAlpaca	Full	Standard
qwenbode_1_8b_chat_ultraalpaca_qlora	Qwen-1.8B Chat	1.8B	UltraAlpaca	QLoRA	Standard

Note: Adapter State 'Merged' indicates fine-tuning weights are integrated into the base model; 'LoRA'/'QLoRA' indicates separate adapter weights. All models are available on Hugging Face[4].

tion on the leaderboard employs a **3-shot** configuration for the `oab_exams` task, using **accuracy (acc)**.

- **ASSIN 2 (RTE & STS) [Real *et al.*, 2020]**: This benchmark, derived from the second Avaliação de Similaridade Semântica e Inferência Textual (ASSIN 2) shared task, provides data for two tasks: Recognizing Textual Entailment (RTE) and Semantic Textual Similarity (STS). The dataset was specifically created for these tasks, focusing on simple factual sentences to avoid complex linguistic phenomena present in earlier benchmarks like ASSIN 1. The RTE task uses a binary classification (entailment/non-entailment). The STS task involves predicting a similarity score. The corpus contains approximately 10,000 sentence pairs, generated through semi-automated methods and manual creation/revision, annotated by native Portuguese speakers. For RTE the leaderboard evaluates this using a **15-shot** setup and reporting the **macro F1-score (f1_macro)** while for STS it is evaluated in a **15-shot** setting using the **Pearson correlation coefficient (pearson)** as the metric.
- **FAQUAD NLI [Sayama *et al.*, 2019]:** This datasets is designed for reading comprehension in Portuguese, similar in structure to SQuAD. It contains 900 questions based on 249 contexts extracted from Wikipedia articles and official documents concerning the Brazilian higher education system. The task involves identifying the answer to a question as a continuous span of text within the provided context. Multiple correct answer spans were annotated for many questions to allow for more robust evaluation, following the SQuAD methodology. The leaderboard uses a **15-shot** configuration for the `faquad_nli` task, evaluated with the **macro F1-score (f1_macro)**.
- **HateBR [Vargas *et al.*, 2022]:** This corpus focuses on offensive language and hate speech detection in Brazilian Portuguese. It comprises 7,000 Instagram comments

[4] https://huggingface.co/recogna-nlp

collected from public posts by Brazilian politicians (balanced by gender and political alignment) during the second half of 2019. The data underwent cleaning to remove noise. Annotation was performed by specialists following a three-layer schema: (1) Binary classification (offensive/non-offensive), (2) Offensiveness level (highly/moderately/slightly offensive) for offensive comments, and (3) Hate speech classification (identifying specific hate categories like xenophobia, racism, sexism, etc., or classifying as offensive but not hate speech). The dataset is balanced for the binary task (3,500 offensive, 3,500 non-offensive). Evaluation employs a **25-shot** setting on the `hatebr_offensive` task, measured by the **macro F1-score (f1_macro)**.

- **PT Hate Speech [Fortuna *et al.*, 2019]:** This dataset contains 5,668 Portuguese tweets collected in 2017 using Twitter's API via keyword and profile searches related to hate speech. After filtering and sampling, the tweets were annotated in two ways. First, a binary annotation (hate speech/not hate speech) was performed by 18 non-expert annotators, achieving low initial agreement, with the final label determined by majority vote. Second, a hierarchical annotation was conducted by an expert using an open coding methodology and a Directed Acyclic Graph (DAG) structure to capture hate speech subtypes (e.g., sexism, racism, homophobia) and their intersections. This detailed annotation achieved substantial agreement between the expert and a second validating annotator on a subset. The leaderboard evaluation focuses on the binary classification ('hate' vs. 'no-hate') using a **25-shot** setup on the `portuguese_hate_speech` task, with the **macro F1-score (f1_macro)** as the metric.
- **TweetSentBR [Brum and Volpe Nunes, 2018]:** This corpus is designed for sentiment analysis in Brazilian Portuguese, containing 15,000 tweets related to Brazilian TV shows collected during the first half of 2017. User-generated content was targeted, excluding retweets and posts from official entities. Tweets were annotated by seven native speakers into three classes: positive, negative, or neutral, following a detailed codebook containing examples, definitions and tips for the annotation process. A 'doubt' option was available to annotators. The final labels were determined by majority vote. Evaluation is performed in a **25-shot** setting using the `tweetsentbr` task, reporting the **macro F1-score (f1_macro)**.

It is important to acknowledge the methodological implications of our evaluation approach. Following the convention of the Open PT-LLM Leaderboard, we use a consolidated average score for a high-level comparison and ranking of models. This practice has inherent limitations, primarily the aggregation of heterogeneous metrics including accuracy, macro F1-score, and the Pearson correlation coefficient—which are not directly commensurable. A single average score, while convenient for summary purposes, can mask important performance variations; for example, a model's strength in classification tasks might obscure its weakness in semantic similarity, or vice-versa. We adopt this method for consistency with the established community benchmark. Presenting the detailed results for all 35 models across all nine benchmarks within this paper would be infeasible due to space constraints. However, for full transparency and to allow for a more granular analysis, we direct readers to the interactive Open PT-LLM Leaderboard[5], where the complete, per-task performance metrics for every model are publicly available.

7 Results

This section delves into the performance evaluation of various Large Language Models (LLMs) fine-tuned for Portuguese, benchmarked against their original base models. The results, summarized across nine distinct Portuguese language tasks, highlight the impact of fine-tuning methodologies and model characteristics (like scale and architecture) on downstream performance within the Portuguese linguistic context. Evaluations were conducted using the framework and datasets specified in the Open PT-LLM Leaderboard.

7.1 Overall Performance Landscape

Table 2 presents the consolidated performance metrics, listing fine-tuned models alongside their base counterparts. Each fine-tuned model is immediately followed by an idented row showing the performance of its base model on the same tasks.

A clear trend emerges where larger models generally achieve higher average scores. The `internlm-chatbode-20b` model leads the pack with an average score of 71.68, demonstrating strong capabilities across the board. Close contenders include the base `LLaMa-3.1-8B-Instruct` (71.24) and several fine-tuned 7B/8B models like `Bode-3.1-8B-Instruct-full` (69.78) and various InternLM 7B variants (around 69).

Conversely, smaller models (e.g., 1.5B, 1.8B, 2B parameters) occupy the lower ranks, with average scores ranging from the low 30s to mid-50s. Models like `phibode_1_5_ultraalpaca` (31.95) and `gembode-2b-ultraalpaca-qlora` (31.19) illustrate the challenges smaller architectures face on these complex benchmarks, although fine-tuning sometimes provides a noticeable uplift compared to their respective base models (e.g., `internlm2-chat-1_8b-ultracabrita` vs. `internlm2-chat-1_8b`). The choice of fine-tuning strategy (Full, LoRA, QLoRA) also influences outcomes, as discussed in subsequent sections.

7.2 Fine-tuning vs. Base Model Performance

Comparing fine-tuned models directly against their base counterparts reveals nuanced impacts. Calculating the average difference across all paired models in Table 2, we observe a modest average improvement of approximately **+0.74 points** for fine-tuned models. However, this average masks significant variation across tasks and model families.

The most substantial average gains from fine-tuning are observed in tasks demanding nuanced understanding of Portuguese sentiment, offensiveness, and cultural context:

- **PT HATE SPEECH**: Average Gain ≈ +4.5 points

[5] `https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard`

- **HATE BR**: Average Gain ≈ +2.2 points
- **tweetSentBR**: Average Gain ≈ +1.8 points
- **ASSIN2 RTE**: Average Gain ≈ +1.4 points
- **FAQUAD NLI**: Average Gain ≈ +1.0 points
- **ASSIN2 STS**: Average Gain ≈ +0.4 points

These results strongly suggest that fine-tuning is particularly effective at adapting models to specific linguistic styles, hate speech patterns, and entailment reasoning prevalent in the target language and datasets.

Conversely, performance slightly decreased on average for tasks potentially requiring broader world knowledge or complex reasoning less emphasized during fine-tuning:

- **BLUEX**: Average Loss ≈ -2.2 points
- **OAB Exams**: Average Loss ≈ -1.3 points
- **ENEM**: Average Loss ≈ -1.1 points

This phenomenon, where base models sometimes outperform their fine-tuned versions (e.g., 'LLaMa-3.1-8B-Instruct' vs. its fine-tuned variants on ENEM/BLUEX/OAB, 'Phi-3-mini-4k-instruct' vs. 'phibode-3-mini-4k-ultraalpaca' on most tasks), could stem from several factors. Fine-tuning might lead to catastrophic forgetting of general knowledge or reasoning skills, or potentially overfit to the specific style or domain of the fine-tuning data (like UltraAlpaca or BODE), hindering performance on broader academic exams like ENEM, BLUEX, and OAB. The specific fine-tuning data and method (Full vs. LoRA/QLoRA) likely play crucial roles in this trade-off.

7.3 Model Specific Analysis

7.3.1 InternLM Variants

InternLM models, particularly the 20B version (`internlm-chatbode-20b`), achieve top-tier performance, indicating strong base capabilities and effective adaptation through fine-tuning (+1.09 Avg vs. base). The 7B variants also perform competitively, often clustered near the top, with fine-tuning generally providing slight improvements (`internlm-chatbode2-7b-beta`: +0.06 Avg; `internlmbode-7b`: +3.65 Avg vs. base). The 1.8B model shows a noticeable gain from fine-tuning with UltraCabrita (`internlm2-chat-1_8b-ultracabrita`: +2.51 Avg vs base), especially on ASSIN2 RTE and HATE BR.

7.3.2 LLaMa Variants

The base `LLaMa-3.1-8B-Instruct` stands out as one of the best-performing models overall, even without specific Portuguese fine-tuning. Its fine-tuned versions (`Bode-3.1-8B...`) show slightly lower average scores (-1.46 Avg for full, -2.43 Avg for LoRA), particularly struggling on ENEM, BLUEX, OAB, and FAQUAD compared to the base. This suggests potential overfitting or knowledge degradation during the fine-tuning process for this specific architecture. Older LLaMa-2 based models (`bode-7b/13b-alpaca-pt-br`) show significant gains over their respective base models (`llama-2-7b/13b-hf`) (+5.92 Avg for 7B, but a decrease of -4.95 Avg for 13B), though their overall scores remain in the mid-to-low range. The 13B fine-tune surprisingly underperforms its base.

7.3.3 Mistral/Zephyr Variants

Mistral-based models show moderate performance. The base `Mistral-7B-v0.1` (61.13 Avg) benefits significantly from fine-tuning using QLoRA and UltraAlpaca (`mistralbode_7b_qlora_ultraalpaca`: +4.22 Avg), particularly on STS, FAQUAD, and HATE tasks. The simpler `mistral-bode` fine-tune shows a decrease (-3.92 Avg). Similarly, `zephyr_7b_beta` (64.47 Avg) sees a slight average improvement when fine-tuned with UltraAlpaca (`zephyr_7b_beta_ultraalpaca`: +0.69 Avg), with gains primarily in HATE BR and PT HATE SPEECH.

7.3.4 Phi Variants

The `Phi-3-mini-4k-instruct` (3.8B) base model performs remarkably well for its size (66.41 Avg). Fine-tuning it with UltraAlpaca (`phibode-3-mini-4k-ultraalpaca`) leads to a significant drop in average performance (-5.72 Avg), underperforming the base on nearly all tasks except ASSIN2 STS and PT HATE. The older Phi-2 (2.7B) and Phi-1.5 (1.5B) models and their fine-tunes (`Phi-Bode`, `phi-bode-2-ultraalpaca`, `phibode_1_5_ultraalpaca`) occupy the lower performance tiers. Fine-tuning provides gains over the base Phi-2 (+7.07 Avg for Phi-Bode, +3.37 Avg for phi-bode-2) and Phi-1.5 (+2.31 Avg), but their absolute scores remain low.

7.3.5 Gemma Variants

Gemma models show varied results. The fine-tuned `gembode-7b` improves notably over the base `gemma-7b` (+2.93 Avg), especially on PT HATE SPEECH. However, the instruct-tuned base `gemma-7b-it` performs poorly (49.61 Avg), though its fine-tuned version `gembode-7b-it` sees a large improvement (+10.99 Avg), particularly lifting performance on ENEM, BLUEX, OAB, HATE BR, and tweetSentBR. The smaller Gemma 2B models show modest gains from UltraAlpaca fine-tuning (`gembode-2b-ultraalpaca`: +1.62 Avg vs base), but QLoRA fine-tuning results in a significant performance drop (`gembode-2b-ultraalpaca-qlora`: -12.88 Avg). The instruct-tuned `gemma-2b-it` and its fine-tuned version `GemBode-2b-it` (+1.87 Avg) both perform near the bottom.

7.3.6 Qwen Variants

The Qwen 1.8B models are among the lowest performers. Fine-tuning the base `Qwen-1_8B` with UltraAlpaca (`qwenbode_1_8b_chat_ultraalpaca`) yields a noticeable improvement (+4.03 Avg), especially on HATE BR and PT HATE SPEECH. However, applying QLoRA during fine-tuning (`qwenbode_1_8b_chat_ultraalpaca_qlora`) drastically reduces performance compared to the base (-2.51 Avg), particularly harming performance on ASSIN2 RTE/STS and FAQUAD NLI.

Table 2. Consolidated Performance: Fine-tuned models vs Base Models

Model Name	Average	ENEM	BLUEX	OAB Exams	ASSIN2 RTE	ASSIN2 STS	FAQUAD NLI	HATE BR	PT HATE SPEECH	tweetSentBR
internlm-chatbode-20b	71.68	65.78	58.69	43.33	91.53	78.95	81.36	81.72	73.66	70.11
internlm2-chat-20b	70.59	67.67	57.86	44.01	91.22	82.75	78.81	79.87	72.82	60.31
internlm-chatbode2-7b-beta	69.65	62.98	52.43	42.55	91.63	82.92	77.98	88.32	62.71	65.31
internlm2-chat-7b	69.59	61.79	50.76	41.82	91.09	82.39	78.55	87.26	70.98	61.67
internlm-chatbode-7b	69.54	63.05	51.46	42.32	91.33	80.69	79.8	87.99	68.09	61.11
internlm2-chat-7b	69.59	61.79	50.76	41.82	91.09	82.39	78.55	87.26	70.98	61.67
internlm-chatbode2-7b-alpha	68.89	57.94	48.68	38.22	91.55	81.51	79.09	85.85	71.79	65.41
internlm2-chat-7b	69.59	61.79	50.76	41.82	91.09	82.39	78.55	87.26	70.98	61.67
Bode-3.1-8B-Instruct-full	69.78	70.26	58.97	50.84	91.08	76.22	71.57	86.2	62.48	60.41
LLaMa-3.1-8B-Instruct	71.24	70.75	59.81	51.53	92.64	77.36	75.36	86.45	64.37	62.9
Bode-3.1-8B-Instruct-lora	68.81	69.84	57.3	47.65	91.97	74.3	65.06	88.41	63.94	60.8
LLaMa-3.1-8B-Instruct	71.24	70.75	59.81	51.53	92.64	77.36	75.36	86.45	64.37	62.9
internlmbode-7b	68.52	60.18	50.07	40.27	90.74	81.74	75.39	87.93	67.51	62.88
internlm2-7b	64.87	60.18	51.88	39.86	88.2	81.15	60.07	67.98	68.24	66.23
gembode-7b	67.11	66.9	57.16	45.47	86.61	71.39	67.4	79.81	63.75	65.49
gemma-7b	64.18	67.04	56.47	42.87	81.34	64.28	69.23	85.69	42.51	68.19
mistralbode_7b_qlora_ultraalpaca	65.35	56.82	47.15	36.31	88.92	76.37	67.17	82.02	69.24	64.19
Mistral-7B-v0.1	61.13	63.89	50.21	43.92	88.92	62	48	76.73	59.64	56.84
mistral-bode	57.21	47.03	39.78	33.76	85.66	62.15	56.45	73.25	63.61	53.17
Mistral-7B-v0.1	61.13	63.89	50.21	43.92	88.92	62	48	76.73	59.64	56.84
zephyr_7b_beta_ultraalpaca	65.16	57.03	44.92	39.64	90.68	69.97	65.14	83.25	70.36	65.45
zephyr_7b_beta	64.47	57.87	47.98	39.32	88.36	66.78	70.18	81.77	66.59	61.42
phibode-3-mini-4k-ultraalpaca	60.69	56.12	40.75	38.5	88.56	69.63	50.65	82.19	68.1	51.67
Phi-3-mini-4k-instruct	66.41	65.22	53.96	45.69	90.64	73.6	56.06	84.34	70.82	57.4
gembode-7b-it	60.6	49.34	36.58	34.76	79.09	64.95	64.67	86.27	63.61	66.17
gemma-7b-it	49.61	36.6	30.32	27.7	81.44	60.84	57.36	72.73	55.99	23.53
bode-7b-alpaca-pt-br	54.82	34.36	28.93	30.84	79.83	43.47	67.45	85.06	65.73	57.67
llama-2-7b-hf	48.9	31.91	31.29	35.44	67.02	31.1	53.87	75.16	55.26	59.06
bode-13b-alpaca-pt-br	52.54	33.66	38.25	36.04	71.22	46.75	51.68	82.21	65.54	47.55
llama-2-13b-hf	57.49	53.74	44.51	39.95	86.32	58.74	43.97	81.08	53.1	56.03
gembode-2b-ultraalpaca	45.69	34.71	25.87	31.71	71.31	34.08	60.09	47.01	57.04	49.37
gemma-2b	44.07	26.45	28.37	28.34	63.53	36.35	44.75	77.82	36.81	54.25
gembode-2b-ultraalpaca-qlora	31.19	32.05	21.56	27.47	33.33	0.87	43	36.41	34.22	51.79
gemma-2b	44.07	26.45	28.37	28.34	63.53	36.35	44.75	77.82	36.81	54.25
Phi-Bode	43.59	33.94	25.31	28.56	68.1	30.57	43.97	60.51	54.6	46.78
phi-2	36.52	34.99	26.98	28.29	38.38	8.87	43.92	59.63	51.23	36.37
phi-bode-2-ultraalpaca	39.89	38.35	25.17	29.61	45.39	24.43	43.97	54.15	54.59	43.34
phi-2	36.52	34.99	26.98	28.29	38.38	8.87	43.92	59.63	51.23	36.37
GemBode-2b-it	36.12	21.69	25.31	26.83	52.71	16.28	52.95	67.52	24.22	37.54
gemma-2b-it	34.25	28.76	25.87	28.38	57.17	5.51	55.2	44.55	23.59	39.2
qwenbode_1_8b_chat_ultraalpaca	40.22	31.21	25.73	24.83	69.07	17.89	40.29	52.88	58.6	41.51
Qwen-1_8B	36.19	30.23	26.15	27.2	64.83	19.53	43.97	33.33	41.23	39.26
qwenbode_1_8b_chat_ultraalpaca_qlora	33.68	31.21	26.01	26.2	40.52	4.64	32.15	60.1	54.14	28.18
Qwen-1_8B	36.19	30.23	26.15	27.2	64.83	19.53	43.97	33.33	41.23	39.26
phibode_1_5_ultraalpaca	31.95	23.58	20.72	24.87	69.07	4.94	43.97	34.94	41.23	24.19
phi-1_5	29.64	21.62	23.5	23.92	33.33	13.02	43.97	33.3	41.23	32.88
internlm2-chat-1_8b-ultracabrita	51.51	35.48	30.74	30.11	84.74	60.31	43.97	72.31	55.21	50.71
internlm2-chat-1_8b	49	32.05	31.15	30.57	75.68	52.56	66.05	58.55	55.82	38.53

7.4 Comparative Analysis with SOTA Portuguese LLMs

To contextualize the Bode family's performance within the landscape of Portuguese Large Language Models (LLMs), we performed a comparative analysis using the Open PT-LLM Leaderboard (results presented in Table 3). Our analysis categorizes models into two approximate parameter size classes: larger models ($\approx$ 16B Parameters) and smaller models ($\approx$ 3B Parameters). This approach facilitates performance comparisons within distinct computational scales. Consistent with the goal of evaluating accessible models, we exclusively selected open-weight LLMs primarily developed or fine-tuned for Portuguese, thereby ensuring reproducibility and community access by excluding proprietary models, despite their potential high performance.

Analysis of Larger Models ($\approx$ 16B): In the higher parameter range, the Bode family demonstrates strong competitiveness, although it doesn't universally occupy the top positions. Our analysis reveals that model performance is dictated not just by parameter count but crucially by the recency of the base architecture, as models like `boto-9B-it` and a Portuguese-specialized LLaMa 3 variant (`LLaMa-3-8B-Dolphin-Portuguese-v0.3`) currently lead this category on the leaderboard, which is largely populated by models in the 8-9B class. Against this backdrop, the `internlm-chatbode-20b` model stands out due to its scale and ranks as the highest-performing Bode variant, showcasing the potential of combining scale with effective Portuguese fine-tuning on the modern InternLM architecture.

The importance of architectural lineage over sheer size becomes evident when comparing newer models with older ones. For instance, the Llama 2-based `bode-13b-alpaca-pt-br` (Avg: 52.54) is significantly outperformed by numerous smaller 7B models built on more recent architectures, such as `internlm-chatbode2-7b-beta` (Avg: 69.65) and `mistralbode_7b_qlora_ultraalpaca` (Avg: 65.35). methodologies. This trend even challenges newer models, as the `internlm-chatbode2-7b-beta` also surpasses the larger `CabraQwen-14B` (Avg: 68.66), further highlighting the critical impact of the underlying architecture and fine-tuning strategy.

Table 3. Performance Comparison of open-source models on the Open PT-LLM Leaderboard.

Model Name	Parameter Size (Approx.)	Average Score (Leaderboard)
≈ 16B Parameter		
boto-9B-it	9B	73.65
LLaMa-3-8B-Dolphin-Portuguese-v0.3	8B	73.15
Internlm-chatbode-20b	20B	71.68
Bode-3.1-8B-Instruct-full	8B	69.78
internlm-chatbode2-7b-beta	7B	69.65
Bode-3.1-8B-Instruct-lora	8B	68.81
CabraQwen14b	14B	68.66
internlmbode-7b	7B	68.52
gembode-7b	7B	67.11
mistralbode_7b_qlora_ultraalpaca	7B	65.35
zephyr_7b_beta_ultraalpaca	7B	65.16
gembode-7b-it	7B	60.60
Vicuna-7B	7B	57.03
mistral-bode	7B	57.21
bode-7b-alpaca-pt-br	7B	54.82
bode-13b-alpaca-pt-br	13B	52.54
Sabiá-7B	7B	47.09
≈ 3B Parameter		
phi-3-portuguese-tom-cat-4k-instruct	3.8B	64.57
cesar-ptbr	1.14B	64.04
phibode-3-mini-4k-ultraalpaca	3.8B	60.69
internlm2-chat-1_8b-ultracabrita	1.8B	51.51
gembode-2b-ultraalpaca	2B	45.69
Phi-Bode	2.7B	43.59
qwenbode_1_8b_chat_ultraalpaca	1.8B	40.22
phi-bode-2-ultraalpaca	2.7B	39.89
GemBode-2b-it	2B	36.12
qwenbode_1_8b_chat_ultraalpaca_qlora	1.8B	33.68
phibode_1_5_ultraalpaca	1.5B	31.95
gembode-2b-ultraalpaca-qlora	2B	31.19
GlórIA 1.3B	1.3B	5.44

The 8B LLaMa 3 Bode variants (`Bode-3.1 8B Instruct full` and `lora`) also achieve high rankings, placing competitively near the top. Notably, as observed in Section 7, the full fine-tune slightly outperforms the LoRA version, but both lag marginally behind their highly capable base model (`LLaMa-3.1-8B-Instruct`, Avg: 71.24 from Table 2) and the specialized Dolphin fine-tune on this benchmark average. This reinforces the finding that while fine-tuning adapts the model to Portuguese nuances (improving on specific tasks like Hate Speech detection), it can sometimes slightly hinder performance on broader benchmarks compared to a very strong, well-aligned multilingual base or alternative fine-tuning approaches.

Several 7B Bode models, particularly those based on InternLM (`internlm-chatbode2-7b-beta`), Mistral (`mistralbode_7b_qlora_ultraalpaca`), and Zephyr (`zephyr_7b_beta_ultraalpaca`), occupy solid positions in the upper-mid tier, generally outperforming baselines like `Vicuna-7B` and earlier Portuguese models such as `Sabiá-7B`. This highlights the effectiveness of the UltraAlpaca dataset and QLoRA/LoRA fine-tuning on more recent architectures.

Analysis of Smaller Models (≈ 3B): In the smaller model category, the landscape is diverse. The Bode family offers several contenders in this space. The `phibode-3-mini-4k-ultraalpaca` is the highest-ranking smaller Bode model, leveraging the capable Phi-3 Mini base. However, similar to the LLaMa 3 case, it scores lower than both its base model (‘Phi-3-mini-4k-instruct‘, Avg: 66.41) and the alternative `tom-cat` fine-tune on the leaderboard average, suggesting potential for further optimization in adapting this specific architecture.

While our analysis focuses on open-weight models to ensure reproducibility, it is helpful to contextualize their performance against state-of-the-art proprietary systems. For instance, the top five positions in the Open PT-LLM Leaderboard are occupied by proprietary models: `Gemini 2.5 Pro [2025-04-03]`, `Claude 3.7 Sonnet [2025-04-04]`, `GPT 4o [2025-04-09]`, `Sabiá 3 [2024-08-20]` and `Gemini 2.0 Flash [2025-04-03]` respectively. These models achieved average scores of 88.37, 84.49, 83.81, 82.32 and 82.3, respectively. Including this information provides a reference point that highlights the performance ceiling that open-source efforts aim to reach while situating the accomplishments of the Bode models within a broader context.

It's also important to note that these proprietary models are significantly bigger than the models discussed and proposed in this paper.

In conclusion, the Bode models represent a significant contribution to the Portuguese NLP ecosystem, offering competitive performance and, crucially, providing a broad comparative benchmark across multiple modern LLM architectures adapted for the language. While specialized models might top specific leaderboards, the Bode family provides a robust and diverse set of tools and insights for researchers and practitioners working with Brazilian Portuguese.

7.5 Qualitative Evaluation

To isolate the benefits of Brazilian Portuguese adaptation, and better understand the big disparity in score between ChatBode and InternLM2, we analyze cases where ChatBode (our best performing model) correctly classifies sentiment in the tweetSentBR dataset, while InternLM2 (it's multilingual counterpart) makes mistakes. Of the 374 instances, 99.2% of InternLM's errors involve misclassifying sentiments as Neutral, reflecting systemic struggles with BP's informal and culturally nuanced language. To better understand that, we selected three particular cases for qualitative evaluation.

The first case reads "quero que saía rolando a glória pires" (literal translation to "I want Gloria Pires to roll out"). This case employs the Portuguese verb "rolar", a meaning that has been repurposed in BP slang to mean abrupt dismissal or rejection. ChatBode correctly identifies the sarcastic negative sentiment, while InternLM defaults to neutral, interpreting the phrase literally.

Another such instance happens with the expression "vergonha alheia" (literally translated to "other people's shame"). While neutral in literal translation, it conveys secondhand embarrassment ("cringe") in practice. ChatBode's negative classification aligns with this cultural nuance, whereas InternLM's neutral label reveals its inability to disambiguate idioms from their literal counterparts.

A subtler error emerges in factual statements devoid of explicit sentiment, such as "meu marido tem NUMBER anos e eu estou com NUMBER estamos juntos à NUMBER anos" ("My husband is NUMBER years old, I am NUMBER, we've been together for NUMBER years"). Here, the absence of emotional markers renders the sentiment neutral. ChatBode correctly identifies this neutrality, while InternLM erroneously assigns a positive, likely inferring positivity from keywords like "marido" (husband) and "juntos" (together).

These cases reveal an interesting trend: InternLM's errors arise from an overreliance on surface-level lexical cues (e.g., explicit sentiment words like "rage" or "happiness") and insufficient training on BP's implicit linguistic strategies. Showcasing that our fine-tuning strategies can deal with such a gap by exposing the model to annotated BP slang, idioms, and pragmatic norms, enabling it to decode sarcasm and cultural expressions.

7.5.1 Culturally Nuanced Evaluation with ENEM Questions

In order to evaluate the quality of language-specific fine-tuning, we assessed how each model performs on a subset of 7 carefully selected questions from ENEM (Brazilian National High School Examination), which were chosen to highlight cultural and linguistic nuances of Brazilian Portuguese. These questions require not only linguistic competence but also cultural awareness and sensitivity to Brazilian literary contexts. Table 4 presents the results of this evaluation across distinct base models and their Bode-adapted variants.

Across 12 model families, Bode-adapted variants matched or outperformed their base counterparts in 67% of cases (8/12), demonstrating the general efficacy of targeted adaptation. The most striking improvement occurred in the Mistral family: while the base model scored zero correct answers, its Bode-adapted variant achieved three, suggesting that adaptation unlocked previously inaccessible cultural reasoning capabilities.

Three models, LLaMa 3.1, Bode LLaMa 3.1, and ChatBode 20B, are tied for the highest score of 4 correct answers. It is noteworthy that two of these top performers are Bode-adapted models. The fact that LLaMa 3.1 performed equally well with or without adaptation suggests that this base model may already possess robust cross-lingual capabilities that transfer effectively to the Brazilian context. At the other extreme, LLaMa 2 and Mistral base models failed entirely without adaptation, underscoring the necessity of language-specific tuning for architectures lacking multilingual pretraining breadth.

The distribution of correct answers across questions provides valuable insights into model capabilities. Questions 4 and 5 were the most frequently answered correctly (each by 41.7% of models), while Question 6 proved exceptionally challenging, with only GemBode 2 providing the correct answer. This particular question required empathy analysis toward young readers in a conversational excerpt from a Brazilian novel, a task demanding deep cultural awareness beyond mere language translation. The fact that only an adapted model succeeded on this question reveals the value of cultural adaptation for capturing nuanced communicative styles.

Notably, all models operated in a zero-shot setting without task-specific training, measuring their ability to transfer general BP knowledge to cultural reasoning. While this tests broader generalization, it may underestimate potential performance with dedicated fine-tuning. Nevertheless, the results carry critical implications: the dominance of adapted models among top performers demonstrates that cultural sensitivity is not an emergent property of scale alone but requires deliberate, language-first adaptation.

7.6 Environmental Impact Analysis

In an effort to address the environmental impact of large-scale model training, we estimated the carbon footprint of our fine-tuning experiments. The methodology is based on the work of Lacoste *et al.* [2019] and utilizes the Machine Learning Impact calculator[6]. The calculation is based on an estimated cumulative training time of 780 hours across

[6] `https://mlco2.github.io/impact/`

Table 4. Qualitative evaluation results for different model families based on seven questions. A checkmark (✓) indicates a correct answer, while a cross (×) indicates an incorrect answer. The total column shows the number of correct answers per model.

Family	**Model**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Total**
LLaMa 2	LLaMa 2	×	×	×	×	×	×	×	0
	Bode 7B	✓	×	✓	×	×	×	×	2
LLaMa 3.1	LLaMa 3.1	✓	×	✓	✓	✓	×	×	4
	Bode LLaMa 3.1	✓	×	✓	✓	✓	×	×	4
Phi1.5	Phi1.5	×	✓	×	×	×	×	×	1
	PhiBode1.5	×	✓	×	×	×	×	×	1
Phi2	Phi2	×	✓	×	×	×	×	×	1
	Phi2 Bode	×	✓	×	✓	×	×	×	2
Phi3	Phi3	✓	×	×	✓	✓	×	✓	4
	Phi3 Bode	×	×	✓	✓	✓	×	✓	4
Gemma 7B	Gemma 7B	×	✓	×	×	×	×	×	1
	Gembode 7B	×	✓	×	×	×	×	×	1
Gemma 2	Gemma 2	×	✓	×	×	×	×	×	1
	GemBode 2	×	×	×	×	✓	✓	✓	3
Mistral	Mistral	×	×	×	×	×	×	×	0
	Mistral Bode	✓	×	×	✓	✓	×	×	3
Qwen	Qwen	×	✓	×	×	×	×	×	1
	Qwen Bode	×	✓	×	×	×	×	×	1
Zephyr	Zephyr	×	×	×	✓	×	×	✓	2
	Zephyr Bode	✓	×	×	×	×	×	×	1
InternLM 7B	InternLM 7B	✓	×	×	✓	✓	×	×	3
	ChatBode 7B	✓	×	×	✓	✓	×	×	3
InternLM 20B	InternLM 20B	×	×	✓	×	✓	×	×	2
	ChatBode 20B	✓	×	✓	✓	✓	×	×	4

all 35 models. This estimation was derived from developer recollection and model scale: approximately 12 hours for small-scale models (≈ 3B parameters), 24 hours for medium-scale models (≈ 7B parameters), and 48 hours for larger-scale models (≈ 16B parameters). This cumulative time was paired with the hardware specifications, specifically the use of NVIDIA V100 GPUs and the geographical location of the compute infrastructure (Brazil) to account for the local power grid's carbon intensity.

The carbon efficiency for Brazil's National Interconnected System (SIN) in 2024 is, on average, 0.0545 tCO^2/MWh, equivalent to 0.0545 kg/kWh. This factor is published by the Brazilian Ministry of Science, Technology, and Innovation (MCTI) for use in emission inventories[7]. The MCTI emission factor used in this estimation is based on a recently improved data collection process from the National Electrical System Operator (ONS), initiated in January 2025. This enhancement expands the calculation base to include additional zero-emission sources, such as biomass plants and entire solar and wind farm complexes, which were previously not fully accounted for, making the data more precise. While this data expansion leads to more accurate factors, the core calculation methodology remains unchanged.

Based on these parameters, the total emissions for the cumulative 3,120 GPU-hours of computation (780 hours on nodes with 4 GPUs each) are estimated to be 51.01 $kgCO_2eq$, of which 0 percent were directly offset. To contextualize this value, it is equivalent to the emissions from driving an average internal combustion engine car for 206 km, burning 25.5 kg of coal, or the carbon sequestered by approximately one tree seedling over 10 years.

This analysis, grounded in official national data, underscores the non-trivial environmental cost associated with large-scale experimentation in NLP. However, the relatively low emissions also highlight the benefits of using an energy grid with low carbon intensity and, more importantly, the value of the PEFT methods investigated in this paper. The use of Q-LoRA requires substantially less training time compared to full fine-tuning, directly translating to a lower carbon footprint. This positions PEFT not only as a computationally efficient strategy but also as a more sustainable approach for adapting LLMs. The clear trade-off between model size and environmental impact further reinforces the importance of developing smaller, specialized models that can achieve competitive performance without the high environmental cost of their larger counterparts.

[7] `https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/sirene/dados-e-ferramentas/fatores-de-emissao`. Accessed on June 26th, 2025

7.7 Bias, Fairness and Limitations

A primary limitation of our work stems from the dual sources of bias inherent in our data preparation strategy. Firstly, the reliance on source datasets created in English, such as Alpaca, means we risk importing the cultural norms and societal biases of their origin, as direct translation can fail to capture subtle cultural nuances. Secondly, the use of GPT-3 family models for translation introduces a vector for propagating the model's own latent biases, which can lead to culturally insensitive outputs. However, we contend that this approach is a pragmatic necessity, as it is currently the most viable method for obtaining the sheer volume of instruction-following examples required for robust model training.

Despite the challenges of data-induced bias, previous works have shown that language-specific fine-tuning can yield a critical benefit. Adapting models to the specific linguistic structures of a target language, such as Brazilian Portuguese, can markedly reduce their propensity for hallucination. In fact, it has been demonstrated that Portuguese-specific fine-tuning makes models less prone to hallucination even when trained on translated datasets Jodas *et al*. [2024]. This suggests that language-specific adaptation makes the models more robust and reliable in their outputs. Nevertheless, this improvement in factuality does not erase the impact on fairness; the models may still produce culturally insensitive or unfair content, posing a risk in real-world applications.

The core limitation of the current approach is, therefore, a trade-off between reduced hallucination and inherited bias. While the Bode family represents a significant step forward in providing accessible and proficient Portuguese models, our findings highlight that the most crucial future perspective is to break the cycle of dependency on English-centric resources. The development of large-scale, high-quality, "Portuguese-first" instruction datasets remains an important next step for the research community. Such an initiative is essential to mitigate propagated biases and build models that are not only linguistically proficient but also truly equitable and culturally aware.

8 Conclusion

The Bode family represents a significant step toward advancing LLM capabilities in Brazilian Portuguese. Our findings illuminate several key insights for future development: (1) the necessity of tailored adaptation strategies that balance language-specific optimization with general reasoning preservation, (2) the importance of robust evaluation methodologies that capture both linguistic competence and cultural relevance, and (3) the value of culturally aware model development to ensure these technologies effectively serve Brazil's diverse linguistic needs.

Through systematic experimentation, we've demonstrated that while language-specific fine-tuning yields substantial improvements for Portuguese-language tasks, it requires careful implementation to avoid compromising general capabilities. The observed trade-offs between specialized adaptation and broad competence highlight the complexity of developing truly multilingual AI systems.

These results have immediate practical implications for NLP practitioners working with Portuguese and other underrepresented languages. The released Bode models and associated benchmarks provide both a foundation for future research and a cautionary template for language-specific adaptation efforts. Our work particularly underscores how architectural choices, scaling effects, and fine-tuning approaches differentially impact model performance across various linguistic tasks.

Looking forward, we emphasize that realizing the full potential of LLMs for Brazilian Portuguese will require sustained, collaborative efforts across multiple dimensions: from creating more comprehensive evaluation frameworks to developing novel adaptation techniques that better preserve model versatility. The challenges identified in this study - particularly regarding knowledge retention and reasoning preservation - suggest fruitful directions for both theoretical and applied research in multilingual NLP.

Ultimately, this investigation contributes to the broader goal of democratizing language technologies by providing both concrete resources (the Bode model family) and empirical insights that can guide future development. As the field progresses, we hope these efforts will help bridge the current gaps in LLM capabilities for Portuguese and other global languages, moving us closer to truly equitable language technology access.

Declarations

Acknowledgements

The authors would like to thank FAPESP and CNPq for the funding of this paper.

Funding

The authors are grateful to the São Paulo Research Foundation (FAPESP) grants 2013/07375-0, 2023/14427-8, and 2024/00789-8, and also to the Brazilian National Council for Scientific and Technological Development (CNPq) grants 308529/2021-9 and 400756/2024-2.

Authors' Contributions

Pedro Henrique Paiola and Gabriel Lino Garcia contributed to the conception of this study. Pedro Henrique Paiola and Gabriel Lino Garcia performed the experiments. João Vitor Mariano Correia, João Renato Ribeiro Manesco, and Ana Lara Alves Garcia wrote the initial draft of this paper. João Paulo Papa was responsible for supervision and funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors declare that they don't have competing interests.

Availability of data and materials

The models and datasets generated and used during the current study are publicly available in HuggingFace:

- Models: `https://huggingface.co/recogna-nlp`
- Datasets:
  - ENEM: `https://huggingface.co/datasets/maritaca-ai/enem`
  - BLUEX: `https://huggingface.co/datasets/portuguese-benchmark-datasets/BLUEX`
  - OAB Exams: `https://huggingface.co/datasets/eduagarcia/oab_exams`
  - ASSIN2: `https://huggingface.co/datasets/nilc-nlp/assin2`
  - FAQUAD-NLI: `https://huggingface.co/datasets/ruanchaves/faquad-nli`
  - PT Hate Speech: `https://huggingface.co/datasets/eduagarcia/portuguese_benchmark`
  - HateBR: `https://huggingface.co/datasets/ruanchaves/hatebr`
  - TweetSentBR: `https://huggingface.co/datasets/eduagarcia/tweetsentbr_fewshot`

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. DOI: 10.18653/v1/2023.emnlp-main.298.

Almeida, T. S., Laitz, T., Bonás, G. K., and Nogueira, R. (2023). BLUEX: A benchmark based on Brazilian Leading Universities Entrance eXams. arXiv. DOI: 10.48550/arXiv.2307.05410.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. DOI: 10.48550/arxiv.2004.05150.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

Brum, H. and Volpe Nunes, M. d. G. (2018). 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). DOI: 10.48550/arXiv.1712.08917.

Chaudhary, S. (2023). Code alpaca: An instruction-following llama model for code generation. Available at:`https://github.com/sahil280114/codealpaca`.

Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. DOI: 10.48550/arxiv.1904.10509.

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and De Oliveira, N. (2024). Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *Machine Learning with Applications*, 16:100558. DOI: 10.1016/j.mlwa.2024.100558.

Delfino Pedro, Cuconato Bruno, Haeusler Edward Hermann, and Rademaker Alexandre (2017). Passing the Brazilian OAB Exam: Data Preparation and Some Experiments. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. DOI: 10.48550/arXiv.1712.05128.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115. DOI: 10.48550/arxiv.2305.14314.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., and Zhou, B. (2023). Enhancing chat language models by scaling high-quality instructional conversations. DOI: 10.18653/v1/2023.emnlp-main.183.

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., and Yang, M. (2024). Longrope: Extending llm context window beyond 2 million tokens. DOI: 10.48550/arxiv.2402.13753.

Du, J., Grave, É., Gunel, B., Chaudhary, V., Çelebi, O., Auli, M., Stoyanov, V., and Conneau, A. (2021). Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418. DOI: 10.18653/v1/2021.naacl-main.426.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/W19-3510.

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

Garcia, G. L., Paiola, P. H., Garcia, E., Manesco, J. R. R., and Papa, J. P. (2025). GemBode and PhiBode: Adapting Small Language Models to Brazilian Portuguese. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 228–243, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-76607-$7_1$7.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing bode: a fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*. DOI: 10.48550/arxiv.2401.02909.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. DOI: 10.48550/arXiv.2407.21783.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. (2023). Textbooks Are All You Need. arXiv. DOI: 10.48550/arXiv.2306.11644.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. DOI: 10.48550/arxiv.2106.09685.

Hughes, A. (2023). Phi-2: The surprising power of small language models. Available at:`https:`

//www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

InternLM Team (2023). Internlm: A multilingual language model with progressively enhanced capabilities. Available at:https://github.com/InternLM/InternLM-techreport.

Jain, S. (2022). tiktoken: A fast bpe tokeniser for use with openai's models. Available at:https://github.com/openai/tiktoken/.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. arXiv. DOI: 10.48550/arXiv.2310.06825.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts. DOI: 10.48550/arxiv.2401.04088.

Jodas, D. S., Garcia, G. L., Paiola, P. H., Ribeiro Manesco, J. R., and Papa, J. P. (2024). Impact of quantization on large language models for portuguese classification tasks. In *Iberoamerican Congress on Pattern Recognition*, pages 213–227. Springer. DOI: 10.1007/978-3-031-76607-7_16.

Kumar, P. (2024). Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260. DOI: 10.1007/s10462-024-10888-y.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. (2023). Openassistant conversations – democratizing large language model alignment. Available at:https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. DOI: https://doi.org/10.48550/arxiv.1910.09700.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*. DOI: 10.1016/j.aiopen.2024.09.002.

Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., and Caridá, V. (2023). Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*. DOI: https://doi.org/10.48550/arxiv.2308.11878.

Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. (2024). Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838. DOI: 10.48550/arxiv.2402.10946.

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., and Lee, Y. T. (2023). Textbooks Are All You Need II: phi-1.5 technical report. arXiv. DOI: 10.48550/arXiv.2309.05463.

Lopes, R., Magalhães, J., and Semedo, D. (2024). Glória–a generative and open large language model for portuguese. DOI: 10.48550/arXiv.2402.12969.

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. (2023). Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. arXiv. DOI: 10.48550/arXiv.2303.17003.

Paiola, P. H., Garcia, G. L., Manesco, J. R. R., Roder, M., Rodrigues, D., and Papa, J. P. (2024). Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *arXiv preprint arXiv:2410.00163*. DOI: 10.24132/csrn.2025-37.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer. DOI: 10.1007/978-3-031-45392-2_15.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al.* (2018). Improving language understanding by generative pre-training. Available at:https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc. Available at:https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67. DOI: 10.48550/arxiv.1910.10683.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). 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*, pages 406–412, Berlin, Heidelberg. Springer-Verlag. DOI: 10.1007/978-3-030-41505-1_39.

Sayama, H. F., Araujo, A. V., and Fernandes, E. R. (2019). FaQuAD: Reading Comprehension Dataset in the Domain of Brazilian Higher Education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. DOI: 10.1109/BRACIS.2019.00084.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, abs/1707.06347. DOI: 10.48550/arxiv.1707.06347.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL). DOI: 10.18653/v1/p16-1009.

Shazeer, N. (2020). Glu variants improve transformer. DOI:

10.48550/arxiv.2002.05202.

Singh, S., Vargus, F., Dsouza, D., Karlsson, B. F., Mahendiran, A., Ko, W.-Y., Shandilya, H., Patel, J., Mataciunas, D., OMahony, L., Zhang, M., Hettiarachchi, R., Wilson, J., Machado, M., Moura, L. S., Krzemiński, D., Fadaei, H., Ergün, I., Okoh, I., Alaagib, A., Mudannayake, O., Alyafeai, Z., Chien, V. M., Ruder, S., Guthikonda, S., Alghamdi, E. A., Gehrmann, S., Muennighoff, N., Bartolo, M., Kreutzer, J., Üstün, A., Fadaee, M., and Hooker, S. (2024). Aya dataset: An open-access collection for multilingual instruction tuning. DOI: 10.18653/v1/2024.acl-long.620.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer. DOI: 10.1007/978-3-030-61377-$8_2$8.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. (2023). Roformer: Enhanced transformer with rotary position embedding. DOI: 10.1016/j.neucom.2023.127063.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and Efficient Foundation Language Models. arXiv. DOI: 10.48550/arXiv.2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., *et al.* (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. (2023). Zephyr: Direct distillation of lm alignment. DOI: 10.48550/arxiv.2310.16944.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association. DOI: 10.48550/arXiv.2103.14972.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. Available at:`https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

XTuner Contributors (2023). Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. (2024). Metamath: Bootstrap your own mathematical questions for large language models. DOI: 10.48550/arxiv.2309.12284.

Zhang, B. and Sennrich, R. (2019). Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. Available at:`https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf`.

Zhou, H., Liu, F., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S. S., Zhou, P., Liu, J., *et al.* (2023). A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*. DOI: 10.48550/arXiv.2311.05112.

Appendix A: ENEM Questions with English Translations Used for Qualitative Evaluation

This appendix presents the seven questions from the Brazilian National High School Examination (ENEM) used in our evaluation of language models, along with their English translations. These questions were specifically selected to assess the models' ability to understand culturally nuanced aspects of Brazilian Portuguese literature and language.

Question 1: Literary Expression in Rural Brazil

Original:

> Firmo, o vaqueiro
>
> No dia seguinte, à hora em que saia o gado, estava eu debruçado à varanda quando o vaqueijo que preparava o animal viajeiro: – Raimundinho, como vai ele?... De longe apontou a palhoça. – Sim. O braço caiu-lhe, olhou-me algum tempo comovido; depois, saltando para o animal, levou o polegar à boca fazendo estalar a unha nos dentes: 'Às quatro horas da manhã... Mirei um verso e disse, para bulir com ele: Pesa, velho! Não respondeu... Tio Firmo, mesmo velho e doente, não era homem para deixar um verso no chão...' Fui ver, coitado... estava morto'. E dei de esporas para que eu não lhe visse as lágrimas. A passagem registra um momento em que a expressividade lírica é reforçada pela:

English Translation:

> Firmo, the cowboy
>
> The next day, at the time when the cattle were leaving, I was leaning on the veranda when the cowboy who was preparing the traveling animal: – Raimundinho, how is he?... From a distance he pointed to the hut. – Yes. His arm fell, he looked at me moved for some time; then, jumping onto the animal, he took his thumb to his mouth making his nail crack against his teeth: 'At four in the morning... I looked at a verse and said, to tease him: Weigh, old man! He didn't answer... Uncle Firmo, even old and sick, wasn't a man to leave a verse on the ground...' I went to see, poor thing... he was dead'. And he spurred on so that I wouldn't see his tears. The passage records a moment in which lyrical expressiveness is reinforced by:

Answer Options:

A. plasticity of the gathered herd image
B. suggestion of the firmness of the backwoodsman when saddling the horse
C. situation of generalized poverty in the Brazilian backlands
D. affection demonstrated when reporting the death of the singer
E. cowboy's concern to demonstrate his virility

Question 2: Modern Life and Relationships

Original:

> Se tudo é para ontem, se a vida engata uma primeira e sai em disparada, se não há mais tempo para paradas estratégicas, caímos fatalmente no vício de querer que os amores sejam igualmente resolvidos num átimo de segundo. Temos pressa para ouvir 'eu te amo'. Não vemos a hora de que fiquem estabelecidas as regras de convivência: somos namorados, ficantes, casados, amantes? Urgência emocional. Uma cilada. Associações diversas palavras ao AMOR: paixão, romance, sexo, adrenalina, palpitação. Esquecemos, no entanto, de palavras tão essenciais quanto 'paciência'. Amor sem paciência não vinga. Amor não resiste à angústia do engolido com cólera, na intransigência de um querer feito de exigências. Amor sem paciência é uma refeição que pode durar uma vida.
>
> Nesse texto de opinião, as marcas linguísticas revelam uma situação tensa e de pouca formalidade, o que se evidencia pelo(a):

English Translation:

> If everything is for yesterday, if life shifts into first gear and races away, if there is no more time for strategic stops, we inevitably fall into the addiction of wanting love to be equally resolved in the blink of an eye. We're in a hurry to hear 'I love you'. We can't wait for the rules of coexistence to be established: are we dating, hooking up, married, lovers? Emotional urgency. A trap. We associate various words with LOVE: passion, romance, sex, adrenaline, palpitation. We forget, however, words as essential as 'patience'. Love without patience doesn't thrive. Love doesn't resist the anguish of being swallowed with anger, in the intransigence of a desire made of demands. Love without patience is a meal that can last a lifetime.
>
> In this opinion text, the linguistic marks reveal a tense situation with little formality, which is evidenced by:

Answer Options:

A. impersonalization throughout the text, as in: 'if there is no more time'
B. construction of an atmosphere of urgency, in words such as: 'hurry'
C. repetition of a certain syntactic structure, as in: 'If everything is for yesterday'
D. reflection in the use of hyperbole, as in: 'a meal that can last a lifetime'
E. use of metaphors, as in: 'life shifts into first gear and races away'

Question 3: Gender Roles in Media and Society

Original:

> Texto 1 Zapeei os canais, como há dezenas de anos faço, e pá: parei num que exibia um episódio daquela velha família do futuro, Os Jetsons.
>
> Nesse episódio em particular, a Jane Jetson, esposa do George, tratava de dirigir aquele veículo voador deles. Meu queixo foi caindo à medida que as piadinhas machistas sobre mulheres dirigindo foram se acumulando. Impressionante! Que futuro careta aqueles roteiristas imaginavam! Seriam incapazes de projetar algo melhor, e ainda mais em termos de tecnologias, robôs e carros voadores? Será que nossa máxima visão de futuro só atinge as coisas, e jamais as pessoas? Como a Jane, uma mulher de 33 anos no desenho, poderia ser o que foram as minhas bisavós?
>
> O futuro, naquele desenho, se assemelha ao de ser melhor nas relações entre as pessoas. Aliás... tão pouco como a vida.
>
> Fiquei de cara, como dizemos aqui, ou me dizem minhas dúvidas na minha adolescência, pobre adolescência, aprendendo a sem querer e sem muita defesa, um futuro tão besta quanto o passado.
>
> Texto 2 Masculino e feminino são campos escorregadios que só se definem por oposição, sempre incompleta, um do outro. São formações imaginárias que buscam produzir uma diferença radical e complementar onde só existem, de fato, mínimas diferenças. O resto é questão de estilo. Até pelo menos a segunda metade do século 19, o divisor de águas era claro: os homens ocupavam o espaço público. As mulheres tratavam das atividades da privada. Privada de quê? De visibilidade, diria Hannah Arendt. De visibilidade pública. E mesmo que tentassem reverter a situação ao século 20 foi de presença pública manifesta não em imagem, mas em ação. A palavra feminina, reservada ao universo doméstico, não produzia diferença na vida social.
>
> Os dois textos acima têm em comum a crítica:

English Translation:

> Text 1 I zapped through the channels, as I've been doing for dozens of years, and bam: I stopped at one showing an episode of that old family of the future, The Jetsons.
>
> In this particular episode, Jane Jetson, George's wife, was driving their flying vehicle. My jaw dropped as the sexist jokes about women driving

accumulated. Impressive! What a square future those writers imagined! Would they be incapable of projecting something better, especially in terms of technologies, robots, and flying cars? Is our maximum vision of the future only about things, and never about people? How could Jane, a 33-year-old woman in the cartoon, be what my great-grandmothers were?

The future, in that cartoon, resembles being better in relationships between people. In fact... as little as life itself.

I was shocked, as we say here, or as my doubts tell me in my adolescence, poor adolescence, learning unwillingly and without much defense, a future as stupid as the past.

Text 2 Masculine and feminine are slippery fields that only define themselves by opposition, always incomplete, to each other. They are imaginary formations that seek to produce a radical and complementary difference where, in fact, there are only minimal differences. The rest is a matter of style. At least until the second half of the 19th century, the watershed was clear: men occupied the public space. Women dealt with private activities. Private from what? From visibility, Hannah Arendt would say. From public visibility. And even if they tried to reverse the situation, the 20th century was one of public presence manifested not in image, but in action. The feminine word, reserved for the domestic universe, did not produce a difference in social life.

The two texts above have in common the critique of:

Answer Options:

A. forms of feminine expression
B. absence of the feminine figure in public life
C. imaginary constructions crystallized in society
D. limitations inherent to feminine and masculine figures
E. difficulty in attributing masculine and feminine roles

Question 4: Political Allegory in Brazilian Literature

Original:

Enquanto estivemos entretidos com os urubus outras coisas andaram acontecendo na cidade. A Companhia baixou novas proibições, umas inteiramente bobocas, só pelo prazer de proibir (ninguém podia cuspir pra cima, nem carregar água em jacá, nem tapar o sol com peneira, como se todo mundo estivesse abusando dessas esquisitices); mas outras bem irritantes, como a de pular muro pra cortar caminho, tática que quase todo mundo que não sofria de reumatismo vinha adotando ultimamente, principalmente os meninos. E não confiando na proibição só, nem na força dos castigos, que eram rigorosos, a Companhia ainda mandou fincar cacos de garrafa nos muros. Achei isso um exagero, e comentei o assunto com mamãe. Meu pai ouviu lá do quarto e veio explicar. Disse que em épocas normais bastava uma coisa ou outra; mas agora a Companhia não podia admitir nenhuma brecha em suas ordens; se alguém desobedecesse à proibição podia se cortar nos cacos; se alguém conseguisse pular um muro quebrando era apanhado pela proibição, nhoc – e fez o gesto de quem torce o pescoço de um frango.

Sob a perspectiva do menino que narra, os fatos ficcionais oferecem um esboço do momento político vigente na década de 1970, aqui representado pelo:

English Translation:

While we were entertained with the vultures, other things were happening in the city. The Company imposed new prohibitions, some entirely silly, just for the pleasure of prohibiting (no one could spit upward, nor carry water in a wicker basket, nor cover the sun with a sieve, as if everyone was abusing these oddities); but others quite irritating, like jumping over walls to take shortcuts, a tactic that almost everyone who didn't suffer from rheumatism had been adopting lately, especially the boys. And not trusting in the prohibition alone, nor in the force of punishments, which were rigorous, the Company even ordered shards of glass to be placed on the walls. I thought this was an exaggeration, and I commented on the matter with Mom. My father heard from the bedroom and came to explain. He said that in normal times one thing or another was enough; but now the Company could not admit any gaps in its orders; if someone disobeyed the prohibition, they could cut themselves on the shards; if someone managed to jump over a wall breaking it, they would be caught by the prohibition, nhoc - and he made the gesture of someone twisting a chicken's neck.

From the perspective of the narrating boy, the fictional facts offer a sketch of the political moment in force in the 1970s, represented here by:

Answer Options:

A. the cult of fear, infiltrated in everyday situations
B. feeling of doubt about the veracity of information
C. dream environment, outlined by disturbing images
D. incentive for economic development with private initiative
E. urban space marked by a policy of isolation of children

Question 5: Environmental Impact of Urban Development

Original:

O masseiro, a mulher, e quatro filhos, dormindo numa tapera de quatro paredes de caixão, coberta de zinco. A água do mangue, na maré cheia, ia dentro de casa. Os maruins de noite encalombavam o corpo dos meninos. O mangue tinha ocasião que fedia, e os urubus faziam ponto por ali atrás dos petiscos. Perto da rua lavavam couro de boi, pele de bode para o curtume de um espanhol. Morria peixe envenenado e quando a maré secava os urubus enchiam o papo, ciscavam a lama, passeando banzeiros pelas biqueiras dos mocambos no Recife.

A aglomeração urbana representada no texto resulta em:

English Translation:

The dough maker, his wife, and four children, sleeping in a shack with four coffin walls, covered with zinc. The mangrove water, at high tide, went inside the house. The midges at night swelled the boys' bodies. The mangrove had occasions when it stank, and the vultures made their spot there looking for tidbits. Near the street, they washed cow leather, goat skin for a Spaniard's tannery. Poisoned fish died, and when the tide ebbed, the vultures filled their crops, scratched the mud, strolling languidly along the eaves of the shanties in Recife.

The urban agglomeration represented in the text results in:

Answer Options:

A. conservation of the rural environment
B. growth of riparian vegetation
C. interference with the geographical space
D. balance of the city environment
E. control of animal proliferation

Question 6: Empathy in Literary Communication

Original:

Volta e meia recebo cartinhas de fãs, e alguns são bem jovens, contando como meu trabalho com a música mudou a vida deles. Fico no céu lendo essas coisas e me emociono quando escrevem que nao sao aceitos pelos pais por serem diferentes, e como minhas músicas são uma companhia e os libertam nessas horas de solidão. Sinto que é mais complicado ser jovem hoje, já que nunca tivemos essa superpopulação no planeta: haja competitividade, culto à beleza, ter filho ou não, estudar, ralar para arranjar trabalho, ser mal remunerado, ser bombardeado com trocentas informações, lavagens cerebrais... Queria dar beijinhos e carinhos sem ter fim nessa moçada e dizer a ela que a barra é pesada mesmo, mas que a juventude está a seu favor e, de repente, a maré de tempestade muda. Diria também um monte de clichê: que vale a pena estudar mais, pesquisar mais, ler mais. Diria que não é sinal de saúde estar bem-adaptado a uma sociedade doente, que o que é normal para uma aranha é o caos para uma mosca. Meninada, sintam-se beijados pela vovó Rita.

Como estratégia para se aproximar de seu leitor, a autora usa uma postura de empatia explicitada em:

English Translation:

Every now and then I receive letters from fans, and some are quite young, telling how my work with music changed their lives. I feel on cloud nine reading these things and I get emotional when they write that they are not accepted by their parents for being different, and how my songs are a companion and free them in these hours of loneliness. I feel that it's more complicated to be young today, since we've never had this overpopulation on the planet: there's competitiveness, cult of beauty, whether to have children or not, study, struggle to find work, be poorly paid, be bombarded with countless pieces of information, brainwashing... I would like to give endless kisses and affection to these young people and tell them that the burden is indeed heavy, but that youth is in their favor and, suddenly, the storm tide changes. I would also say a bunch of clichés: that it's worth studying more, researching more, reading more. I would say that it is not a sign of health to be well-adapted to a sick society, that what is normal for a spider is chaos for a fly. Kids, feel yourselves kissed by grandma Rita.

As a strategy to get closer to her reader, the author uses a posture of empathy made explicit in:

Answer Options:

A. 'Every now and then I receive letters from fans, and some are quite young'
B. 'I feel on cloud nine reading these things'
C. 'I feel that it's more complicated to be young today'
D. 'I would like to give endless kisses and affection to these young people'
E. 'I would say that it is not a sign of health to be well-adapted to a sick society'

Question 7: Intertextuality in Brazilian Writing

Original:

Se você é feito de música, este texto é pra você

As vezes, no silêncio da noite, eu fico imaginando: que graça teria a vida sem música? Sem ela não há paz, não há beleza. Nos dias de festa e nas madrugadas de pranto, nas trilhas dos filmes e nas corridas no parque, o que seria de nós sem as canções que enfeitam o cotidiano com ritmo e verso? Quem nunca curou uma dor de cotovelo dançando lambada ou terminou de se afundar ouvindo sertanejo sofrência? Quantos já criticaram funk e fecharam a noite descendo até o chão? Tudo bem... Raul nos ensinou que é preferível ser essa metamorfose ambulante do que ter aquela velha opinião formada sobre tudo. Já somos castigados com o peso das tragédias, o barulho das buzinas, os ruídos dos conflitos. É pau, é pedra, é o fim do caminho. Há uma nuvem de lágrimas sobre os olhos, você está na lanterna dos afogados, o coração despedaçado. Mas, como um sopro, da janela do vizinho, entra o samba que reanima a mente. Floresce do fundo do nosso quintal a batida que ressuscita o ânimo, sintoniza a alegria e equaliza o fôlego. Levanta, sacode a poeira, dá a volta por cima.

Defendendo a importância da música para o bem-estar e o equilíbrio emocional das pessoas, a autora usa, como recurso persuasivo, a:

English Translation:

If you are made of music, this text is for you

Sometimes, in the silence of the night, I wonder: what joy would life have without music? Without it there is no peace, no beauty. In the days of celebration and in the dawns of weeping, in the movie soundtracks and in the runs in the park, what would become of us without the songs that adorn everyday life with rhythm and verse? Who has never cured a heartbreak by dancing lambada or ended up sinking further listening to sertanejo suffering? How many have criticized funk and ended the night dancing all the way down? It's okay... Raul taught us that it's preferable to be this walking metamorphosis than to have that old opinion formed about everything. We are already punished with the weight of tragedies, the noise of horns, the sounds of conflicts. It's wood, it's stone, it's the end of the road. There is a cloud of tears over the eyes, you are in the lantern of the drowned, the heart shattered. But, like a breath, from the neighbor's window, comes the samba that revives the mind. From the bottom of our backyard blooms the beat that resurrects the spirit, tunes in joy, and equalizes breath. Get up, shake off the dust, turn around.

Defending the importance of music for people's well-being and emotional balance, the author uses, as a persuasive resource:

Answer Options:

A. contradiction, by associating the shattered heart with joy
B. metaphor, by citing the image of the walking metamorphosis
C. intertextuality, by rescuing verses from song lyrics
D. enumeration, by mentioning different musical rhythms
E. hyperbole, by talking about 'suffering', 'tragedies', and 'drowned'