Think Portuguese with Bode Reasoning

Gabriel Lino Garcia[1], André da F. Schuck[1], João R. R. Manesco[1], Pedro Henrique Paiola[1], Leandro A. Passos[1], João Paulo Papa[1],

[1] São Paulo State University (UNESP)
Av. Eng. Luís Edmundo Carrijo Coube, 14-01 - Bauru - SP - Brazil ,
Correspondence: gabriel.lino@unesp.br

Abstract

Large Language Models (LLMs) have introduced reasoning capabilities through multi-step problem-solving processes. These models predominantly perform reasoning in English, limiting their effectiveness in other languages. This paper introduces Bode Reasoning, a Portuguese-language reasoning approach built upon fine-tuned Qwen3-4B and Qwen3-4B-Thinking models, and the Bode Reasoning Portuguese Dataset, comprising $13,961$ instances from Brazilian examinations and translated datasets. Through supervised fine-tuning, the proposed approach successfully shifts the reasoning process to Brazilian Portuguese while reducing output verbosity. Experimental evaluation demonstrates that fine-tuned models generate Portuguese reasoning in $86-98.7\%$ of outputs and achieve superior lexical alignment with reference answers. However, this specialization results in moderate mean G-Eval and accuracy degradation across diverse multiple-choice question types, highlighting inherent trade-offs in adapting multilingual reasoning models.

1 Introduction

The last few decades have witnessed an exponential growth of intelligent mechanisms capable of reasoning, interpreting, and creating, which allowed the development of useful tools for virtually any field of application, such as medicine (Oliveira et al., 2023, 2024) and art (Ko et al., 2023). Among such mechanisms, Large Language Models (LLMs) have achieved widespread popularity due to their outstanding and revolutionary accomplishments in Natural Language Processing (NLP) and, more recently, in tasks such as image, video, and music generation.

Regarding LLMs, a model called Qwen introduced a novel capability in its recent version, Qwen3 (Yang et al., 2025), namely reasoning. Such a process involves performing complex, multi-step problem-solving, in which the model considers the problem's intrinsic features to formulate an appropriate answer. In this context, Qwen3 incorporates thinking budgets that allow users to control the level of reasoning effort the model applies during the task. This capability optimizes computational resources and performance, tailoring the model's behavior to varying levels of complexity in real-world applications. Moreover, Qwen3 also expanded multilingual support to 119 languages, allowing global accessibility through cross-lingual understanding and generation.

While many efforts are devoted to developing LLM models specialized in distinct languages, like Bode (Garcia et al., 2024), which was trained in Portuguese, Qwen3 is, in essence, still a model trained primarily in English, and thus its reasoning is performed in English. Moreover, Qwen3's reasoning process often produces long-winded answers, resulting in cumbersome outputs.

To tackle such problems, this paper proposes **Bode Reasoning**, a set of two Qwen3-based Portuguese reasoning approaches that provide more direct, pointed answers in Portuguese. To train this model, this paper also proposes the **Bode Reasoning Portuguese Dataset**, a novel Brazilian-Portuguese dataset comprising almost $14,000$ instances of questions and multiple-choice answers, extracted from several Brazilian exams and translated tests. Therefore, the main contributions of the paper are described as follows:

- To propose the Bode Reasoning Portuguese Dataset, a Portuguese dataset designed for reasoning tasks and composed of $13,961$ instances across multiple question formats;
- To propose Bode Reasoning, a novel reasoning LLM framework in Portuguese; and
- To improve the performance of Qwen3 on Portuguese outputs by fine-tuning the models on the Bode Reasoning Portuguese dataset.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 953–958
April 13-16, 2026 ©2026 Association for Computational Linguistics

Original Dataset	Description	Type	Original Size	Train Split	Test Slipt	Total	Total %
Bode-Reasoning							
ENEM_challenge (Silveira and Mauá, 2017; Nunes et al., 2023)	Brazilian national high school standardized examination covering multiple academic disciplines.	Multiple-choice	1,432	1,263	135	1,398	10.01%
BLUEX (Almeida et al., 2023)	Entrance examinations from top Brazilian universities with multiple-choice questions across various subjects.	Multiple-choice	724	605	87	692	4.96%
OAB_Exams (Delfino et al., 2017)	Multiple choice questions from the Brazilian Bar Association exam.	Multiple-choice	2,210	2,001	80	2,081	14.91%
Edubench (G-Eval $\geq$.7)	Open-ended questions from three Brazilian university entrance exams.	Open-ended	1,917	1,761	156	1,917	13.73%
Math-Reasoning (Solo Tech, 2025)	Synthetic mathematical problems curated datasets, translated to Brazilian Portuguese.	Open-ended	400	320	80	400	2.87%
Reasoning-v1-20m-portuguese (Moro, 2025b)	Portuguese translation of the synthetic reasoning-v1-20m dataset covering social sciences, natural sciences, education, creative writing and general conversations.	Open-ended	20,896,676	4,337	890	5,227	37.44%
Subtotal				***10,287***	***1,428***	***11,715***	***83.91%***
Bode-mix-no-reasoning							
GPT4-500k-Augmented-PTBR-Clean (Moro, 2025a)	Portuguese translation of the Open-Orca/1million-gpt-4 dataset, filtered to exclude programming -related content and non-Latin characters.	Open-ended	565,536	844	170	1,014	7.26%
Edubench (G-Eval $<$ .7)	Open-ended questions from three Brazilian university entrance exams.	Open-ended	1,232	1,156	76	1,232	8.82%
Subtotal				***2,000***	***246***	***2,246***	***16.09%***
Total				***12,287***	***1,674***	***13,961***	***100.00%***

Table 1: Detailed composition of the Bode Reasoning Portuguese Datasets (13,961 instances), including source datasets, question types, original sizes, and train/test split distributions.

2 Bode Reasoning Portuguese Dataset

This section presents the Bode Reasoning Portuguese Dataset, a collection of 13,961 multiple-choice, open-ended, and instruction-following tasks in Portuguese, extracted from seven public and proprietary datasets for reasoning and non-reasoning tasks. Table 1 compiles the dataset details.

2.1 Dataset Generation

Initially, 4,366 multiple-choice questions were extracted from ENEM, BLUEX, and OAB datasets, with recent exam editions reserved for testing and earlier instances for training. Reasoning traces were generated using Gemini-2.5 PRO based on its superior performance on these datasets[1] (Garcia, 2024). Generation quality was validated by comparing Gemini-2.5 PRO outputs against expected answers, removing 195 training instances with divergent responses, yielding 3,869 training samples with reasoning traces and 302 testing instances. Finally, GPT-4o-mini was employed to translate reasoning traces from English to Brazilian Portuguese, balancing cost efficiency and translation quality.

EduBench was incorporated using the same pipeline for training and test split, reasoning trace generation, and translation. Given the essay-based nature of questions, a G-EVAL (Liu et al., 2023) validation with 70% similarity was employed to prune below-threshold instances, thus including 1,716 training and 156 test samples.

Math-Reasoning (Solo Tech, 2025) was included to diversify reasoning traces, with an 80/20 random split, i.e., 320 training and 80 test instances. It comprises English reasoning traces, which were also translated using GPT-4o-mini.

The dataset was further augmented with Reasoning-v1-20m-portuguese Moro (2025b), selecting random instances with reasoning traces up to 500 words to prioritize conciseness, adding 4,337 training and 890 test samples. Additionally, 844 training and 170 test instances were extracted from GPT4-500k-Augmented-PTBR-Clean (Moro, 2025a), alongside 1,156 training and 76 test instances from EduBench below the $<$ 70% similarity G-EVAL threshold. These samples trained the model without intermediate reasoning traces, enabling the use of high-quality essay questions whose reasoning did not meet minimum criteria but whose expected answers remained valuable. The final Bode Reasoning Portuguese Datasets[2] are described in Table 1.

[1] https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard

[2] Available at HuggingFace: https://huggingface.co/datasets/recogna-nlp/Bode-reasoning https://huggingface.co/datasets/recogna-nlp/Bode-mix-no-reasoning

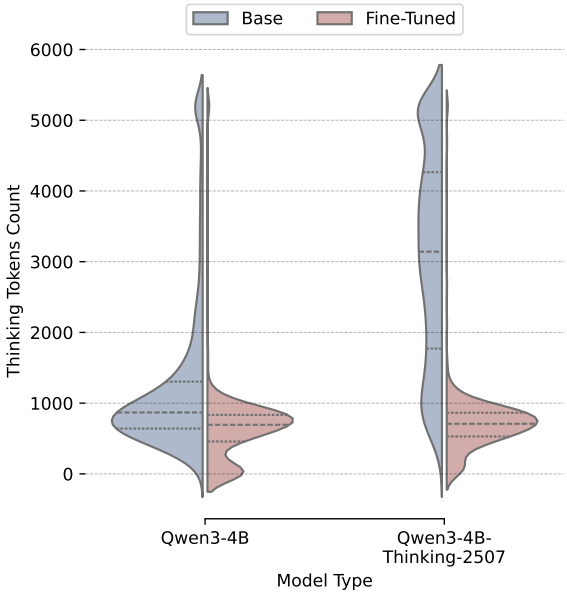


Figure 1: Distribution of thinking token counts comparing base and fine-tuned model types.

3 Methodology

3.1 Fine-Tuning

Bode Reasoning comprises two reasoning approaches, Bode-Reasoning-V0[3] and Bode-Reasoning-Thinking-V0[4], built by fine-tuning the Qwen3-4B and Qwen3-4B-Thinking-2507 SLMs, respectively, for the Portuguese language using the proposed Bode Reasoning Portuguese Dataset.

Fine-tuning was conducted via Supervised Fine-Tuning (SFT) using Low-Rank Adaptation (LoRA) (Hu et al., 2021) applied to all linear projection layers (i.e., query, key, value, output, gate, up, and down projections), with rank $r = 32$, scaling factor $\alpha = 64$, and dropout rate of 0.05. Training was performed for 2 epochs with a per-device batch size of 2 and gradient accumulation over 4 steps, yielding an effective batch size of 8. The AdamW optimizer with 8-bit quantization was employed with a learning rate of 2×10^{-5}, linear scheduling, a warmup ratio of 0.03, and weight decay of 0.01. Due to hardware constrains input sequences were truncated to a maximum length of $7{,}500$ tokens. Gradient checkpointing via Unsloth[5] was enabled to reduce memory consumption. All experiments used a fixed random seed of $3{,}407$ for reproducibility.

[3] https://huggingface.co/recogna-nlp/bode-reasoning-v0

[4] recogna-nlp/bode-reasoning-thinking-v0

[5] https://unsloth.ai/

3.2 Evaluation

The proposal's evaluation is performed by comparing the fine-tuned versions against a baseline composed of the original Qwen3-4B and -4B-Thinking-2507.

The comparison between the fine-tuned versions of Bode Reasoning is performed against Qwen3-4B and Qwen3-4B-Thinking-2507, considering three distinct evaluation metrics: (i) Generative Evaluation (G-Eval) (Liu et al., 2023), an evaluation framework that uses an LLM as a judge to assess the quality of outputs generated by other LLMs using custom, human-defined criteria to score responses; (ii) BERTScore (Zhang et al., 2020), an automatic evaluation metric for text generation that computes the semantic similarity between a text generated by a model and a reference text; and the (iii) Bilingual Evaluation Understudy (BLUE) (Papineni et al., 2002), a metric that measures how closely a machine-generated text matches one or more human-written reference texts. Moreover, the models' accuracies are also compared.

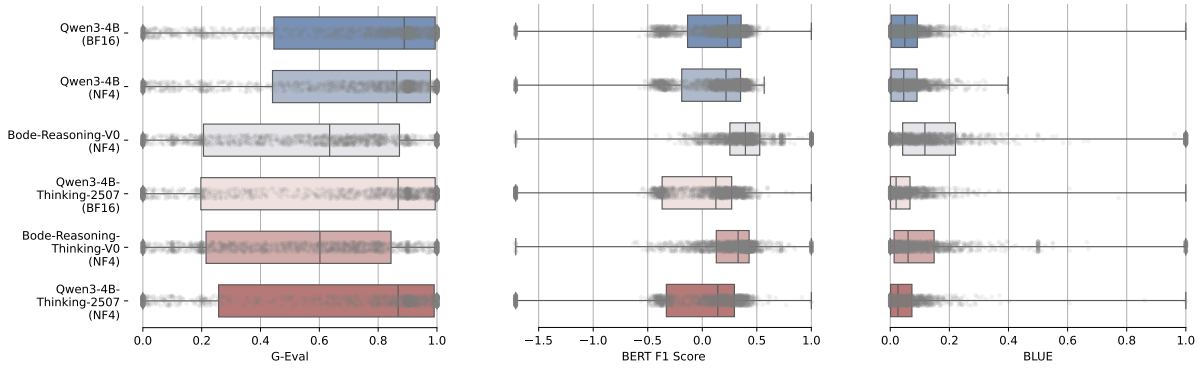
3.3 Experimental Setup

Initially, model responses were generated on Google Colab using an L4 machine (22.5 GB GPU) with generation settings from Yang et al. (2025): Sampling decoding with `temperature=0.6`, `top_p=0.95`, `top_k=20`, `min_p=0`, and `max_new_tokens=5,200` due to hardware constraints.

The responses were evaluated on a T4 machine (15 GB GPU). The G-Eval computation employed GPT-4.1 as the LLM evaluator, assigning zero scores to questions exceeding the $5,200$-token limit. Multiple-choice accuracy was computed using unsloth/Qwen3-8B-bnb-4bit to parse responses and identify selected alternatives from the raw text, returning zero when no option was detected. The evaluation compared generated outputs directly against reference responses, excluding reasoning trace quality and correctness from assessment.

4 Experimental Results

Unlike base models that generate English reasoning, fine-tuned models predominantly produced Brazilian Portuguese reasoning: 86% of all reasoning traces generated by Bode-Reasoning-V0 and 98.7% by Bode-Reasoning-Thinking-V0 were in Portuguese. The remaining outputs contained no reasoning traces, indicating that the models by-



(a) G-Eval. (b) BERT F1 Score. (c) BLUE.

Figure 2: Boxplots distributions of evaluation metrics across models: (a) G-Eval, (b) BERT F1 Score, and (c) BLEU.

passed the reasoning stage entirely and produced final answers directly. This selective omission of reasoning is consistent with the fine-tuning dataset composition, in which approximately 16% of instances (the Bode-mix-no-reasoning subset) consist exclusively of direct question–answer pairs without intermediate reasoning stages, as illustrated in Figure 1.

Moreover, Figure 1 also shows that fine-tuning substantially reduced the reasoning length, generating more concise reasoning, particularly evident in Qwen3-4B-Thinking-2507, which shifted from dispersed, high-token-count distributions to compact reasoning patterns. This behavior also reflects the composition of the training dataset, i.e., 35% of instances have reasoning limited to 500 words.

Table 2 shows that the fine-tuned models obtained a slight degradation on the G-Eval mean performance (-0.117) compared to base models. Both Bode-Reasoning-V0 and Bode-Reasoning-Thinking-V0 exhibited higher concentration of intermediate G-Eval scores (0.4–0.8), while the base models demonstrated a more left-skewed distribution (0.8–1.0), as seen in Figure 2. Conversely, such results demonstrate that fine-tuned models achieved superior BERT F1 and BLEU scores, indicating stronger lexical alignment with reference answers.

Fine-tuned models accounted for 94% (128/136) of perfect BERT F1 scores and 80% (276/343) of perfect BLEU scores, translating to substantial average improvements of $+0.385$ in BERT F1 and $+0.139$ in BLEU over base models. Notably, Bode-Reasoning-V0 accounted for the majority of these scores: 84% of BERT F1 and 76% of BLEU maximal values.

The performance of fine-tuned models was also evaluated in terms of accuracy on multiple-choice question responses. Table 3 revealed an average 11% degradation in fine-tuned models versus the base model in this context. All models achieved the lowest accuracies on the OAB dataset, where the Bode-Reasoning-V0 model showed a marginal 3% improvement.

Model	Precision	G-Eval		BERT F1		BLUE	
		$\bar{X}$	$\tilde{X}$	$\bar{X}$	$\tilde{X}$	$\bar{X}$	$\tilde{X}$
Qwen3-4B	BF16	**0.699**	**0.889**	0.076	0.232	0.059	0.049
	NF4	0.682	0.863	0.050	0.218	0.058	0.046
Bode-Reasoning-V0	NF4	0.556	0.635	**0.420**	**0.396**	**0.249**	**0.117**
Qwen3-4B-Thinking-2507	BF16	0.634	0.868	-0.172	0.125	0.044	0.020
	NF4	0.654	0.868	-0.106	0.142	0.048	0.026
Bode-Reasoning-Thinking-V0	NF4	0.544	0.602	0.273	0.329	0.133	0.060

Table 2: Model performance across G-Eval, BERT F1 Score, and BLEU. Bold indicates best performance; underline indicates second-best. $\bar{X}$ = Mean; $\tilde{X}$ = Median.

Model	Precision	ENEM	BLUEX	OAB	Total
Qwen3-4B	BF16	0.881	**0.885**	0.463	**0.772**
	NF4	0.867	**0.885**	0.475	0.768
Bode-Reasoning-V0	NF4	0.770	0.736	**0.488**	0.685
Qwen3-4B-Thinking-2507	BF16	**0.904**	**0.885**	0.338	0.748
	NF4	0.889	**0.885**	0.425	0.765
Bode-Reasoning-Thinking-V0	NF4	0.785	0.747	0.425	0.679

Table 3: Accuracy per model on multiple-choice question datasets.

5 Conclusion

This paper introduced Bode-Reasoning-V0 and Bode-Reasoning-Thinking-V0, two novel reasoning approaches for Portuguese built on fine-tuning Qwen3-4B and Qwen3-4B-Thinking-2507. It also proposed the Bode Reasoning Portuguese Dataset, comprising 13,961 instances across multiple ques-

tion formats. The experimental evaluation demonstrated that fine-tuning successfully shifted the reasoning process from English to Brazilian Portuguese while substantially reducing verbosity.

The experimental results also revealed distinct performance characteristics across evaluation dimensions. Fine-tuned models excelled in lexical precision for structured tasks, providing direct, format-compliant responses. This specialization came at the cost of reduced semantic quality, as measured by G-Eval, and reduced accuracy across diverse question types. The selective omission of reasoning traces in some outputs suggests that the models acquired adaptive behavior, balancing computational efficiency with task requirements.

These findings underscore the inherent challenges in developing multilingual reasoning models, particularly the tension between linguistic adaptation and the preservation of reasoning capabilities. Future work should prioritize mitigation strategies for performance degradation, specifically addressing catastrophic forgetting—a phenomenon that disproportionately affects SLMs (Luo et al., 2025).

Promising research directions include: (i) augmenting the training dataset with higher-quality, more diverse reasoning traces; (ii) developing Portuguese-specific evaluation frameworks that better capture nuanced reasoning quality beyond direct translation of English metrics; and (iii) investigating continual learning techniques that balance multilingual adaptation with knowledge retention.

Acknowledgments

This study was partially funded by the São Paulo Research Foundation (FAPESP), Brazil, under Grant Nos. #2025/13172-1, #2024/22853-0, and #2023/14427-8. This work was also supported by Petrobras through research grants No. #2025/00607-0 and No. #2023/00466-1.

References

Thales Sales Almeida, Thiago Laitz, Giovana K. Bonás, and 1 others. 2023. Bluex: A benchmark based on brazilian leading universities entrance exams. *Preprint*, arXiv:2307.05410.

Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and 1 others. 2017. Passing the brazilian oab exam: data preparation and some experiments. https://arxiv.org/abs/1712.05128. *Preprint*, arXiv:1712.05128.

Eduardo A. S. Garcia. 2024. Open portuguese llm leaderboard. https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024. Introducing bode: a fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Carlo Moro. 2025a. Gpt4-500k-augmented-ptbr-clean.

Carlo Moro. 2025b. reasoning-v1-20m-portuguese.

Desnes Nunes, Ricardo Primi, Ramon Pires, and 1 others. 2023. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. https://arxiv.org/abs/2303.17003. *Preprint*, arXiv:2303.17003.

Guilherme C Oliveira, Quoc C Ngo, Leandro A Passos, Joao P Papa, Danilo S Jodas, and Dinesh Kumar. 2023. Tabular data augmentation for video-based detection of hypomimia in parkinson's disease. *Computer Methods and Programs in Biomedicine*, 240:107713.

Guilherme C Oliveira, Gustavo H Rosa, Daniel CG Pedronette, João P Papa, Himeesh Kumar, Leandro A Passos, and Dinesh Kumar. 2024. Robust deep learning for eye fundus images: Bridging real and synthetic data for enhancing generalization. *Biomedical Signal Processing and Control*, 94:106263.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Igor Cataneo Silveira and Denis Deratani Mauá. 2017. University entrance exam as a guiding test for artificial intelligence. In *Proceedings of the 6th Brazilian Conference on Intelligent Systems*, BRACIS, pages 426–431.

Solo Tech. 2025. Math-reasoning.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.