

# P3B3: A Multi-Turn Conversational Benchmark for Measuring European and Brazilian Portuguese Variety Bias in LLMs

Rafael Ferreira<sup>1,2</sup>, Inês Vieira<sup>1</sup>, Inês Calvo<sup>1</sup>, James Furtado<sup>1,2</sup>, Iago Paulo<sup>1,2</sup>,  
Diogo Tavares<sup>1,2</sup>, Diogo Glória-Silva<sup>1,2</sup>, David Semedo<sup>1,2</sup>, João Magalhães<sup>1,2</sup>

<sup>1</sup>NOVA University of Lisbon, Portugal, <sup>2</sup>NOVA LINCS

{rah.ferreira, im.paulo, jh.furtado, dc.tavares, dmgc.silva}@campus.fct.unl.pt

{im.vieira, i.calvo, df.semedo, jmag}@fct.unl.pt

## Abstract

As Large Language Models (LLMs) become embedded in everyday communication, capturing regional linguistic variation is essential for reliable and equitable language use. In Portuguese, European (pt-PT) and Brazilian (pt-BR) varieties remain unevenly represented, with pt-BR dominating in data quantity, while LLM preference for Portuguese variants remains underexplored. To address this gap, we introduce P3B3, an expert-curated language variety agnostic benchmark of conversational prompts, along with an evaluation framework for measuring variety bias and controllability. Experiments on several models show that most LLMs exhibit a strong bias toward pt-BR, with variation in controllability across models. These results highlight the need for more balanced multilingual representation across language varieties.<sup>1</sup>

## 1 Introduction

LLMs are increasingly used in everyday communication, supporting applications such as customer service, creative writing, and educational tools. As these systems become more widely deployed, their ability to capture regional linguistic variation is critical for reliable and equitable language use, as well as for linguistic and cultural diversity preservation (Aeppli et al., 2023; Chiu et al., 2025).

Portuguese is a pluricentric language with over 250 million speakers worldwide. While it shares a common core, the divergence between European Portuguese (pt-PT) and Brazilian Portuguese (pt-BR) is substantial, spanning lexical choices (e.g., *autocarro* vs. *ônibus*), forms of address (e.g., *tu* vs. *você*), and syntax (e.g., clitic placement and gerund usage) (Rio-Torto et al., 2022; Preda et al., 2024). Despite these differences, most training and evaluation frameworks treat Portuguese as a single, uniform language (Team, 2024, 2025c; AI,

<sup>1</sup><https://github.com/Amalia-LLM/p3b3-benchmark>

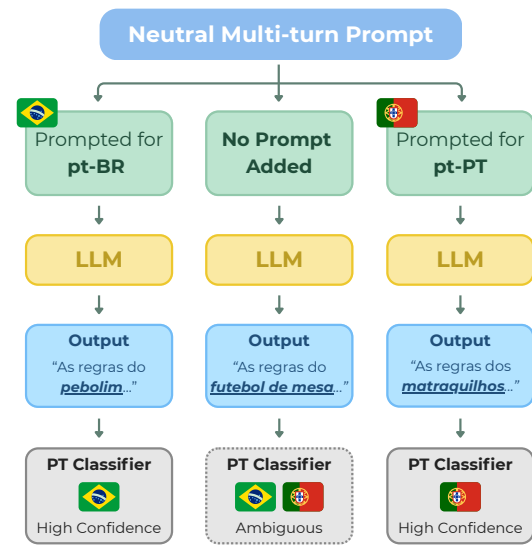


Figure 1: P3B3 example and evaluation methodology. The prompts elicit different Portuguese variants, illustrated by the lexical variations of “table football” in the outputs.

2026). Given the strong prevalence of Brazilian data in large-scale datasets (Penedo et al., 2024; Common Crawl, 2026), this often leads to a silent bias, where models default to pt-BR even when the context suggests pt-PT (Simplício et al., 2026; Riley et al., 2023; González et al., 2026). This behavior can degrade user experience and contribute to linguistic uneven leveling, where regional variation is underrepresented in LLMs.

This imbalance raises two key questions: (i) to what extent do current LLMs exhibit a preference for one Portuguese variant, and (ii) can models be steered toward a target variant.

Facing these questions, our contributions are threefold. First, we introduce the pt-PT/pt-BR Bias Benchmark (P3B3), a benchmark of 74 expert-curated, variety-agnostic multi-turn dialogues designed to elicit Portuguese variety-specific linguistic markers in model responses. Second, we propose a human validated evaluation framework com-

binning classifiers (Sousa et al., 2025) and LLM-as-Judge methods (Gu et al., 2024) to assess variant alignment and controllability. Finally, we conduct an extensive evaluation of diverse LLMs, revealing a bias toward pt-BR and variable success in steering generation toward pt-PT.

Together, the contributions summarized in Figure 1 enable the systematic analysis of pt-PT and pt-BR language varieties in LLMs and support the development of more equitable multilingual systems.

## 2 Related Work

### LLM Performance across Linguistic Varieties.

A growing body of work shows that LLMs exhibit systematic disparities across linguistic varieties. Pan et al. (2025) and Lee et al. (2025) show that model performance degrades on non-standard linguistic varieties across a wide range of English dialects and evaluation settings. Extending this line of work, Hassan et al. (2025) further find a dialectal preference bias, where models assign more negative sentiment to African American English than to equivalent Standard American English inputs.

**Portuguese Variant Classification.** Given the differences across Portuguese varieties (Rio-Torto et al., 2022), prior work has studied variant identification using classification-based methods (Sousa et al., 2025; Preda et al., 2024). Furthermore, Freitag and Gois (2024) show that LLMs yield inconsistent and unreliable predictions in Brazilian Portuguese dialect profiling.

### Multilingual and Portuguese-Specific Models.

Most multilingual LLMs treat Portuguese as a monolithic language, without explicit modeling of variations (Team, 2025c; AI, 2026; Team, 2024). In response, recent work introduces variant-specific LLMs, including Sabiá (Abonizio et al., 2024; Laitz et al., 2026) and Bode (Garcia et al., 2024) for pt-BR, and Glória (Lopes et al., 2024) and AMALIA (Simplício et al., 2026) for pt-PT, aiming to improve cultural and variety alignment.

While prior work examines variety bias in English or frames Portuguese variation as a classification task, we study generation-time bias across Portuguese variants via the expert-designed P3B3, assessing inherent model preferences and their controllability.



Dialogue	Turn	Message
<b>D1</b> Transport 	T1	O meu carro está na oficina. Que transportes posso usar como alternativa?
	T2	Há uma diferença de tempo muito grande entre ir num transporte público e ir de carro?
	T3	É mais fácil ir para o centro de uma cidade de carro ou de transportes?
<b>D2</b> Beauty Products 	T1	Que tipos de produtos são vendidos numa loja de estética feminina?
	T2	Onde é que posso encontrar lojas desse tipo?
	T3	Qual é o produto de estética mais popular?

Figure 2: Examples of multi-turn variety-agnostic prompts from P3B3. Translation in Figure 4.

## 3 The P3B3 Benchmark

The novelty of P3B3 lies in systematically evaluating whether LLMs exhibit preferences toward Portuguese variants and the controllability of such behavior. Specifically, P3B3 is designed around three guidelines: (i) **Variety-agnostic**, with no explicit references to variants in the input to prevent priming effects; (ii) **Variety Sensitivity**, where prompts elicit natural lexical and grammatical differences between Portuguese varieties; and (iii) **Conversational**, with multi-turn examples that accumulate dialogue history.

These guidelines ensure that variation emerges from generation behavior rather than explicit cues or isolated inputs.

### 3.1 Expert-Based Dialogue Creation

Following MT-Bench (Zheng et al., 2023), P3B3 consists of 74 multi-turn dialogues (203 turns in total), each with 2 to 6 turns. As in MT-Bench, later user inputs build on prior dialogue context without requiring specific responses in earlier turns, enabling isolation of generative behavior while maintaining conversational coherence. This design enables cumulative context effects that may amplify biases not observable in single-turn prompts.

To ensure quality and validity, P3B3 was manually constructed by two language experts holding Master’s degrees in linguistics-related fields and with expertise in Portuguese variants. Having this expert knowledge was essential to pinpoint subtle but systematic variation cues, particularly in vocabulary, orthography, and forms of address (Rio-Torto et al., 2022), while remaining variant-agnostic.

The benchmark focuses on diverse everyday domains (e.g., transportation, shopping, household products), which exhibit high variability across Portuguese varieties. These characteristics make

P3B3 suitable for probing generation-time preferences rather than simple surface-level markers. Additional benchmark statistics are provided in Appendix A.

Figure 2 shows representative examples from the dataset, which are variety-agnostic and elicit both lexical and syntactic variation in pt-PT and pt-BR. For instance, in Dialogue-1, Turn-1, the prompt elicits lexical variation such as *ônibus* (pt-BR) vs. *autocarro* (pt-PT), as well as grammatical differences in the response.

## 4 P3B3 Experimental Setup

### 4.1 Priming for Variant Bias and Controllability

We evaluate inherent bias and controllability using P3B3 under three settings: (i) **No-Prompt**, where no explicit variety specification is provided, measuring implicit bias; (ii) **pt-BR Prompt**; and (iii) **pt-PT Prompt**, where in the first dialogue turn the model is explicitly instructed to respond in the corresponding variant.

Following MT-Bench (Zheng et al., 2023), models generate a response at each turn, with the full dialogue history provided as context.

### 4.2 Automatic Evaluation Methods

We employ two evaluation methods.

**Encoder Classifiers.** Model outputs are scored using pt-PT/pt-BR binary classifiers. We use **PeroVaz (PV)** (Bastos, 2024) and **PtBrVid (Vid)** (Sousa et al., 2025), both fine-tuned BERT-based models (Devlin et al., 2019).

The encoder outputs a probability score with lower values indicating pt-BR and higher values indicating pt-PT. Due to input length constraints, we apply a sliding window over each answer and average the scores. While efficient, these models may miss subtle or context-dependent variation, motivating the use of LLM-based evaluation.

**LLM-as-Judge.** We use Gemini-3-Flash (Team, 2025b) to assign a score from 0 (pt-BR) to 10 (pt-PT), along with a chain-of-thought trace (Wei et al., 2022). We use Gemini-3 due to its reliable instruction-following ability and strong understanding of Portuguese (Chang and Arnett, 2025)<sup>2</sup>.

To improve reliability, we vary three factors: prompt language (*EN* or *PT*), evaluation scope

<sup>2</sup>In Appendix E, we show that Gemma-4 (Gemma Team, 2026) can be used a strong open-source alternative.

Method	Weighted $\kappa$ (Quadratic)	Pearson	MAE ↓
Humans	0.69	0.75	2.07
PeroVaz (PV)	0.42	0.50	2.43
PtVid (Vid)	0.57	0.60	2.66
Gemini-3 Flash - EN-S-ST	0.79	0.81	1.79
Gemini-3 Flash - PT-S-ST	0.79	0.82	1.72
Gemini-3 Flash - PT-S-H	0.80	<b>0.83</b>	1.72
Gemini-3 Flash - PT-C-ST	<b>0.81</b>	<b>0.83</b>	<b>1.58</b>
Gemini-3 Flash - PT-C-D	<b>0.81</b>	0.82	1.60

Table 1: Agreement between human annotations and automatic metrics. EN/PT = language; S = Simple; C = Category; ST = Single-Turn; D = Dialogue history.

(Single-Turn, *ST*, or full dialogue history, *D*), and prompt detail (Simple, *S*, or Category-based, *C*). Category prompts include 11 expert-defined linguistic categories added to the prompt as examples, such as pronoun placement, preposition contractions, vocabulary, and spelling.

#### 4.2.1 Validating Automatic Metrics

We validate the metrics using 200 responses annotated by two linguistic experts on a 0 (pt-BR) to 10 (pt-PT) scale, with -1 for invalid outputs. Samples are randomly drawn equally from all settings (No-Prompt, pt-BR, pt-PT) across 12 models. Considering valid responses (88.5%), Table 1 reports agreement with human judgments.

Human annotators show strong agreement, indicating task reliability. Classifiers achieve moderate alignment with humans, with Vid performing best. The LLM-as-a-Judge exhibits strong alignment, performing best under the PT-C-ST setup, which is used in subsequent experiments.

Overall, classifiers provide a low-cost complementary signal, while LLM-based evaluation yields more aligned and interpretable results.

### 4.3 Models and Implementation Details

We evaluate a range of multilingual instruction-tuned LLMs of varying sizes (list in Appendix B). Models are selected based on multilingual and Portuguese performance, public availability, and architectural diversity. All models are evaluated in BF16 using greedy decoding. Additionally, we include closed-source models: Sabiá-4 (Laitz et al., 2026), specialized for Brazilian Portuguese, and Gemini-3-Flash (Team, 2025b) as a frontier model.

Model	No-Prompt			pt-BR (↓)			pt-PT (↑)		
	PV	Vid	LLM	PV	Vid	LLM	PV	Vid	LLM
<b>Fully Open</b>									
Salamandra-7B	47.9	32.8	34.5	44.9	22.1	20.7	52.0	43.7	52.8
OLMo-3-7B	38.6	37.0	17.7	37.5	29.4	15.8	45.8	41.0	23.3
OLMo-3.1-32B	38.9	22.2	16.6	32.9	8.0	10.5	58.9	51.9	42.8
EuroLLM-22B	38.1	25.3	22.5	35.6	10.7	9.6	70.3	87.7	88.5
Apertus-8B	43.2	26.8	20.3	41.9	18.9	14.7	59.6	64.1	61.0
Apertus-70B	41.0	25.7	16.1	36.5	13.6	8.6	68.8	85.7	86.3
AMALIA-9B	66.9	85.6	91.2	45.9	33.5	34.8	71.3	91.8	95.9
<b>Open Weight</b>									
Llama-3.1-8B	41.0	24.1	12.6	38.6	17.0	10.6	57.1	69.3	47.7
Llama-3.3-70B	40.9	25.6	15.3	34.6	13.4	6.7	61.6	81.5	67.8
Gervásio-8B-ptpt	39.7	25.7	12.7	38.6	17.8	11.1	55.3	65.5	46.6
Gervásio-70B-ptpt	47.4	32.4	22.8	36.8	16.7	10.8	64.6	84.8	75.2
Ministral-3-14B	43.6	23.3	17.8	32.5	6.9	4.2	74.8	95.2	90.7
Qwen3-8B	42.0	15.7	10.4	36.4	9.1	7.9	48.1	41.0	32.7
Qwen3.5-9B	42.9	42.0	35.1	28.7	12.8	12.2	76.0	93.6	86.1
Qwen3.5-27B	47.5	47.4	44.5	28.5	11.6	11.9	78.4	98.1	93.3
Gemma-3-12b	38.9	26.5	23.4	30.3	6.7	4.9	72.2	95.0	92.9
Gemma-4-E4B	40.4	29.3	21.9	30.1	9.3	4.2	70.1	96.4	94.4
Gemma-4-31B	49.2	52.3	49.2	28.4	8.3	7.3	69.7	95.6	98.0
<b>Close Source</b>									
Sabiá-4	34.5	7.0	8.1	32.5	4.5	4.6	67.0	90.4	87.2
Gemini-3-Flash	55.5	63.1	63.7	23.6	5.8	4.1	79.7	99.3	99.8

Table 2: Results across models and prompting settings in P3B3. All scores are scaled to 0 (pt-BR) – 100 (pt-PT) measuring variant alignment.

## 5 Results and Discussion

### 5.1 Bias and Controllability Analysis

Table 2 reports model results across three settings. We note that we measure *variant preference* in generation, rather than response quality; therefore, lower or higher scores indicate a tendency toward pt-BR or pt-PT, respectively, rather than differences in output quality.

**No-Prompt.** This setting measures inherent model bias, showing that most models exhibit a consistent preference for pt-BR, likely reflecting training data imbalance. AMALIA, which is specifically pre-trained and fine-tuned for pt-PT, is the only model consistently biased toward pt-PT (91.2). In contrast, Gervásio, a LLaMA model fine-tuned for pt-PT, still exhibits a pt-BR bias, suggesting that pre-training data and scale play a critical role in capturing variety-specific behavior. Interestingly, larger and more recent models such as Qwen-3.5 27B, Gemma-4-31B, and Gemini-3-Flash show more balanced behavior across varieties (44.5 to 63.7).

**pt-BR.** When prompted for pt-BR, most models maintain or increase their use of the pt-BR variety, indicating that models are able to follow explicit variety instructions, particularly at larger model sizes. In contrast, AMALIA is an outlier showing a higher tendency toward pt-PT even in this setting, consistent with its pt-PT specialized training setup.

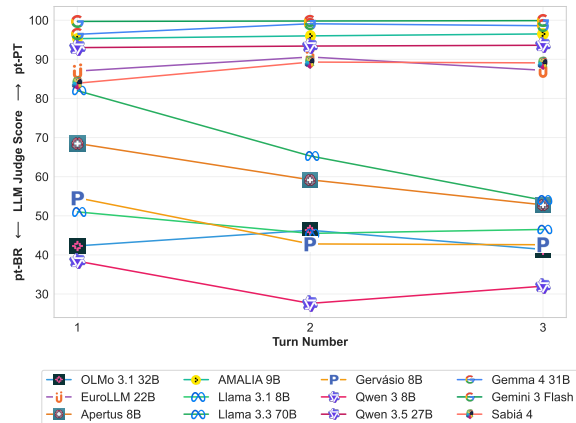


Figure 3: Alignment across turns for pt-PT prompt.

**pt-PT.** Results for pt-PT are more variable. Strong alignment is observed from EuroLLM, Apertus-70B, and AMALIA ( $\geq 86$ ), while LLaMA-based models struggle to produce consistent pt-PT outputs. All Gemma models perform well in this setting ( $\geq 92$ ). Notably, Qwen-3.5 shows a substantial improvement over Qwen-3 (32.7 to 86.1), suggesting that changes in training significantly improved pt-PT handling. Despite being pt-BR-oriented, Sabiá-4 generalizes well to pt-PT (87.2).

**Classifiers vs. LLM Scores.** Classifier-based and LLM-based evaluations yield broadly consistent rankings. Overall, classifiers provide a low-cost complementary signal, while LLM-based evaluation serves as the primary analysis due to higher accuracy and interpretability.

### 5.2 Turn-level Analysis

Figure 3 shows LLM scores across the first three dialogue turns (later turns are excluded due to sample size). We focus on pt-PT as it has the most variance. Except for Llama, we observe that models with strong initial pt-PT alignment generally maintain consistent behavior across turns. Models with lower initial tendency to pt-PT tend to drift toward pt-BR over time. This suggests that a single initial instruction is often insufficient to sustain language variety adherence across multiple turns, indicating that controllability degrades in longer contexts. For the No-Prompt and pt-BR settings, behavior remains stable across turns (Appendix C).

### 5.3 Human Analysis

Linguistic experts evaluated a random subset of various model outputs in the more challenging pt-PT setting. As observed in the automatic metrics,

weaker models tend to drift toward pt-BR, exhibiting proclisis, gerund constructions, pronominal usage (notably *você*), and pt-BR lexical choices (e.g. *geladeira* and *ônibus*). They also show inconsistencies in orthography and accentuation, often producing hybrid outputs that mix pt-BR grammar with partial pt-PT vocabulary (Rio-Torto et al., 2022). In contrast, stronger models largely follow pt-PT conventions, indicating better syntactic and morphological control. We provide examples of model and judge outputs in Appendix F.

## 6 Conclusion

We introduced P3B3, a multi-turn benchmark for evaluating bias and controllability in LLM generation across Portuguese variants. Most models exhibit a strong bias toward pt-BR, and while LLMs can be steered toward different varieties, larger and newer models show greater controllability. However, pt-PT performance remains less consistent across models, highlighting the need for better representation of linguistic variation in training and evaluation.

## Limitations

P3B3 provides a systematic evaluation of LLM adherence to Portuguese varieties in conversational domains. The current setup does not include technical or specialized domains, focusing on everyday topics. Moreover, while P3B3 supports analysis of pt-BR and pt-PT, extending it to other Portuguese varieties spoken in countries such as Angola, Mozambique, and Cape Verde is a promising direction for future work.

## Acknowledgments

This work was supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência. We also acknowledge the support of Fundação para a Ciência e Tecnologia (FCT) and the NOVA LINC project (UID/04516/2025). Finally, we thank the Barcelona Supercomputing Center (BSC) for providing the computational resources that made this work possible.

## References

Hugo Queiroz Abonizio, Thales Sales Almeida, Thiago Soares Laitz, Roseval Malaquias Junior, Giovana K. Bonás, Rodrigo Nogueira, and Ramon

Pires. 2024. [Sabiá-3 technical report](#). *CoRR*, abs/2410.12049.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. [Findings of the VarDial evaluation campaign 2023](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261.

Mistral AI. 2026. [Ministral 3](#). *CoRR*, abs/2601.08584.

Project Apertus. 2025. [Apertus: Democratizing open and compliant llms for global language environments](#). *CoRR*, abs/2509.14233.

Fábio Bastos. 2024. [Perovaz pt-br classifier](#).

Tyler A. Chang and Catherine Arnett. 2025. [Global PIQA: evaluating physical commonsense reasoning across 100+ languages and cultures](#). *CoRR*, abs/2510.24081.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Common Crawl. 2026. Common crawl dataset. <https://commoncrawl.org/>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, and 57 others. 2025. [Olmo 3](#). *CoRR*, abs/2512.13961.

Raquel M. Ko Freitag and Túlio Sousa de Gois. 2024. [Performance in a dialectal profiling task of llms for varieties of brazilian portuguese](#). In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*, STIL 2024, page 317–326. Sociedade Brasileira de Computação.

Gabriel Garcia, Ricardo Rodrigues, Diego Silva, and 1 others. 2024. [Bode: A LLaMA-based large language model for Portuguese](#). *arXiv preprint arXiv:2401.00000*.

- Gemma Team. 2026. [Gemma 4](#).
- José Ángel González, Ian Borrego-Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Rios, Angelo Basile, and Marc Franco-Salvador. 2026. [Iberbench: LLM evaluation on iberian languages](#). *Comput. Speech Lang.*, 96:101899.
- Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and 14 others. 2025. [Salamandra technical report](#). *CoRR*, abs/2502.08489.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 2 others. 2024. [A survey on llm-as-a-judge](#). *CoRR*, abs/2411.15594.
- Muhammad Hassan, Faiza Khattak, and Laleh Seyyed-Kalantari. 2025. [Dialectic preference bias in large language models](#). *Proceedings of the AAAI Symposium Series*, 5:365–369.
- Thiago Laitz, Thales Sales Almeida, Hugo Abonizio, Roseval Malaquias Junior, Giovana Kerche Bonás, Marcos Piau, Celio Larcher, Ramon Pires, and Rodrigo Nogueira. 2026. [Sabia-4 technical report](#). *Preprint*, arXiv:2603.10213.
- Jiyoung Lee, Seungho Kim, Jieun Han, Jun-Min Lee, Kitaek Kim, Alice Oh, and Edward Choi. 2025. [Trans-env: A framework for evaluating the linguistic robustness of llms against english varieties](#). *Preprint*, arXiv:2505.20875.
- Ricardo Lopes, João Magalhães, and David Semedo. 2024. [Glória: A generative and open large language model for portuguese](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, March 12-15, 2024, Volume 1*, pages 441–453. ACL.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 5 others. 2024. [Eurollm: Multilingual language models for europe](#). *CoRR*, abs/2409.16235.
- Eileen Pan, Anna Seo Gyeong Choi, Maartje ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. [Analyzing dialectal biases in llms for knowledge and reasoning benchmarks](#). *Preprint*, arXiv:2510.00962.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: decanting the web for the finest text data at scale](#). NIPS '24. Curran Associates Inc.
- David Preda, Tomás Freitas Osório, and Henrique Lopes Cardoso. 2024. [Across the atlantic: Distinguishing between european and brazilian portuguese dialects](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, March 12-15, 2024, Volume 1*, pages 353–363. Association for Computational Linguistics.
- Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. [FRMT: A benchmark for few-shot region-aware machine translation](#). *Trans. Assoc. Comput. Linguistics*, 11:671–685.
- Graça Rio-Torto, Tânia Ferreira, Zusana Greksakova, Ana Isabel Guerra, and Yunfeng Zhang. 2022. [Português brasileiro e português europeu: um diálogo de séculos](#). Universidade Politécnica de Macau.
- Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. [Advancing generative AI for portuguese with open decoder gervásio PT](#). *CoRR*, abs/2402.18766.
- Afonso Simplicio, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, and 12 others. 2026. [AMALIA: A fully open large language model for European Portuguese](#). In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 380–391, Salvador, Brazil.
- Hugo O. Sousa, Rúben Almeida, Purificação Silvano, Inês Cantante, Ricardo Campos, and Alípio Mário Jorge. 2025. [Enhancing portuguese variety identification with cross-domain approaches](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 25192–25200. AAAI Press.
- Gemma Team. 2025a. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Google Team. 2025b. [Gemini 3 flash model card: Frontier-class performance at scale](#). Technical report, Google DeepMind. Accessed: 2026-05-01.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Qwen Team. 2025c. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *NeurIPS 2022*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 3 others. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

## A P3B3 Benchmark Statistics

P3B3 was constructed through a manual process conducted by two language experts with expertise in Portuguese and comprises of 74 conversations and 203 turns.

**Structural Distribution.** The dataset targets short-form multi-turn interactions with dialogue turns distributed as follows: 31 dialogues have 2 turns, 35 have 3, 5 have 4, and 3 have 5–6 turns.

**Lexical and Sequence Diversity.** The dataset shows strong lexical diversity. Per-conversation TTR is 0.73, with Distinct-2 and Distinct-3 scores of 0.71 and 0.88, indicating high phrasal variation. Lower Distinct-1 (0.26) and global TTR (0.25) reflect expected function word frequency. Entropy increases from 7.7 (1-gram) to 10.9 (3-gram), confirming linguistic variability via different conversational domains.

**Length Statistics.** Turns average 14.1 words (SD=6.0), with word counts averaging 38.7 per dialogue and ranging from 14 to 74 words, consistent with the benchmark’s focused design.

## B Detailed Model List

Table 3 provides a comprehensive list of the models evaluated in this work, along with their respective model identifiers.

Model Provider & Name	Model Identifier
<b>BSC-LT</b>	
Salamandra 7B (Gonzalez-Agirre et al., 2025)	salamandra-7b-instruct
<b>allenai</b>	
OLMo 3 7B (Ettinger et al., 2025)	0lmo-3-7b-instruct
OLMo 3.1 32B (Ettinger et al., 2025)	0lmo-3.1-32b-instruct
<b>utfer-projekt</b>	
EuroLLM 22B (Martins et al., 2024)	EuroLLM-22B-Instruct-2512
<b>swiss-ai</b>	
Apertus 8B (Apertus, 2025)	Apertus-8B-Instruct-2509
Apertus 70B (Apertus, 2025)	Apertus-70B-Instruct-2509
<b>AMALIA-LLM</b>	
AMALIA 9B (Simplicio et al., 2026)	AMALIA-9B-50-1225-DPO
<b>meta-llama</b>	
Llama 3.1 8B Instruct (Team, 2024)	Llama-3.1-8B-Instruct
Llama 3.3 70B (Team, 2024)	Llama-3.3-70B-Instruct
<b>PORTULAN</b>	
Gervasio 8B-pt-pt (Santos et al., 2024)	gervasio-8b-portuguese-ptpt-decoder
Gervasio 70B-pt-pt (Santos et al., 2024)	gervasio-70b-portuguese-ptpt-decoder
<b>mistralai</b>	
Ministral 3 14B (AI, 2026)	Ministral-3-14B-Instruct-2512
<b>Qwen</b>	
Qwen 3 8B (Team, 2025c)	Qwen3-8B
Qwen 3.5 9B	Qwen3.5-9B
Qwen 3.5 27B	Qwen3.5-27B
<b>google</b>	
Gemma 3 12B (Team, 2025a)	gemma-3-12b-it
Gemma 4 E4B (Gemma Team, 2026)	gemma-4-E4B-it
Gemma 4 31B (Gemma Team, 2026)	gemma-4-31B-it
<b>API-only</b>	
Gemini 3 Flash Preview (Team, 2025b)	Google API - gemini-3-flash-preview
Sabiá 4 (Laitz et al., 2026)	Maritaca API - sabia-4

Table 3: Model names and identifiers used in this work.



Dialogue	Turn	Message
D1 Transport 	T1	My car is in the mechanic's shop. What alternative transportation can I use?
	T2	Is there a very big time difference between taking public transport and going by car?
	T3	Is it easier to go to a city center by car or by public transport?
D2 Beauty Products 	T1	What types of products are sold in a women's beauty shop?
	T2	Where can I find these types of shops?
	T3	What is the most popular beauty product?

Figure 4: Translated examples from P3B3 benchmark. Original Portuguese version in Figure 2.

## C Turn-level Analysis

Figures 5 and 6 show LLM scores across the first three dialogue turns for the No-Prompt and pt-BR settings, respectively. In contrast to the pt-PT setting (Figure 3), where alignment progressively shifts toward pt-BR, these configurations exhibit largely stable behavior across turns. This consistency suggests a persistent model bias toward pt-BR, which remains unchanged throughout the interaction, with no tendency to move toward pt-PT.

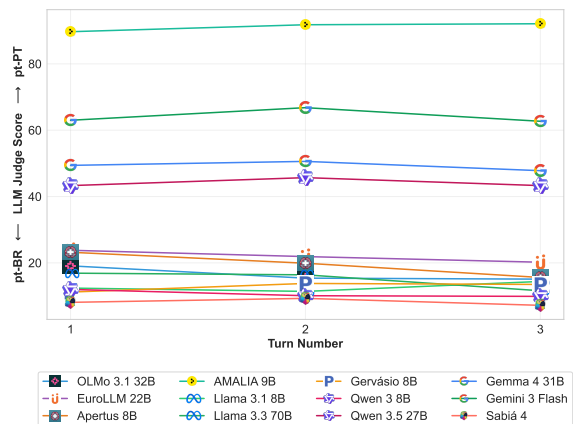


Figure 5: Alignment across turns for No-Prompt.

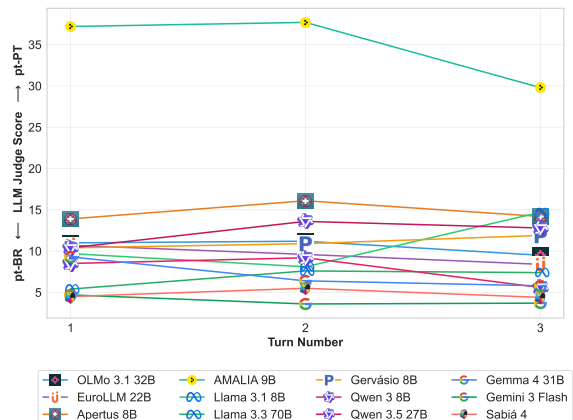


Figure 6: Alignment across turns for pt-BR prompt.

## D Portuguese Output Consistency

We used Gemini-3 Flash to classify whether outputs were predominantly in Portuguese, regardless of variant. Most models consistently generate outputs in Portuguese (99%). The only exception is OLMo-3-7B, which still produces Portuguese in 94% of cases, with occasional outputs in Spanish.

## E Extending the LLM-as-Judge Setup

We additionally include Gemma-4-31B (Gemma Team, 2026) in our LLM-as-Judge setup in Table 4. Gemma-4-31B performs strongly across all settings and metrics. While it trails Gemini-3-Flash, it significantly improves over the classifiers and serves as a strong open-source alternative for this task, with competitive performance.

With respect to the results, Table 5 compares the evaluations produced by Gemini-3-Flash and Gemma-4-31B across all models and prompting settings in P3B3. We observe a strong agreement between the two judges, both in ranking order and overall score ranges, with only a few discrepancies in specific settings. These findings further validate our results and highlight the potential of using open-weight LLM judges to reliably measure Portuguese variant alignment.

## F Example Model and Judge Outputs

Figures 7, 8, and 9 show example model and judge outputs across different settings.

Method	Weighted $\kappa$ (Quadratic)	Pearson	MAE $\downarrow$
Humans	0.69	0.75	2.07
PeroVaz (PV)	0.42	0.50	2.43
PtVid (Vid)	0.57	0.60	2.66
Gemini-3 Flash - EN-S-ST	0.79	0.81	1.79
Gemini-3 Flash - PT-S-ST	0.79	0.82	1.72
Gemini-3 Flash - PT-S-H	0.80	<b>0.83</b>	1.72
Gemini-3 Flash - PT-C-ST	<b>0.81</b>	<b>0.83</b>	<b>1.58</b>
Gemini-3 Flash - PT-C-D	<b>0.81</b>	0.82	1.60
Gemma-4-31B - EN-S-ST	0.75	0.78	1.95
Gemma-4-31B - PT-S-ST	0.78	0.80	1.73
Gemma-4-31B - PT-S-H	0.78	0.80	1.74
Gemma-4-31B - PT-C-ST	0.78	0.79	1.72
Gemma-4-31B - PT-C-D	0.78	0.80	1.71

Table 4: Agreement between human annotations and LLM-based metrics. EN/PT = language; S = Simple; C = Category; ST = Single-Turn; D = Dialogue history.

Model	No-Prompt		pt-BR ( $\downarrow$ )		pt-PT ( $\uparrow$ )	
	Gemma-4	Gemini-3	Gemma-4	Gemini-3	Gemma-4	Gemini-3
<i>Fully Open</i>						
Salamandra-7B	35.5	34.5	24.4	20.7	54.4	52.8
OLMo-3-7B	23.1	17.7	18.1	15.8	25.6	23.3
OLMo-3.1-32B	17.0	16.6	13.3	10.5	47.7	42.8
EuroLLM-22B	23.4	22.5	11.8	9.6	85.0	88.5
Apertus-8B	20.9	20.3	15.8	14.7	59.1	61.0
Apertus-70B	17.7	16.1	12.1	8.6	82.1	86.3
AMALIA-9B	87.5	91.2	33.8	34.8	91.8	95.9
<i>Open Weight</i>						
Llama-3.1-8B	14.1	12.6	10.8	10.6	47.8	47.7
Llama-3.3-70B	16.6	15.3	8.8	6.7	69.5	67.8
Gervásio-8B-ptpt	14.9	12.7	12.4	11.1	46.3	46.6
Gervásio-70B-ptpt	23.3	22.8	12.0	10.8	76.4	75.2
Ministral-3-14B	19.6	17.8	4.2	4.2	93.5	90.7
Qwen3-8B	12.0	10.4	9.2	7.9	33.2	32.7
Qwen3.5-9B	36.1	35.1	13.3	12.2	89.1	86.1
Qwen3.5-27B	44.4	44.5	12.4	11.9	95.1	93.3
Gemma-3-12b	25.6	23.4	6.8	4.9	93.1	92.9
Gemma-4-E4B	23.1	21.9	6.2	4.2	92.9	94.4
Gemma-4-31B	48.4	49.2	8.9	7.3	95.8	98.0
<i>Close Source</i>						
Sabidá-4	9.4	8.1	7.9	4.6	86.9	87.2
Gemini-3-Flash	61.2	63.7	4.7	4.1	98.7	99.8

Table 5: LLM-as-a-Judge comparison between Gemma-4-31B and Gemini-3-Flash across models and prompting strategies in P3B3. Scores are normalized to a 0–100 scale, where 0 corresponds to pt-BR and 100 to pt-PT, measuring alignment with each Portuguese variant.

**No Prompt Added**
**Turn 1**

---

**Question:**  
O meu carro está na oficina. Que transportes posso usar como alternativa?

**Answer:**  
Tens várias opções! Podes apanhar um táxi, usar transportes públicos como o autocarro ou metro, ou pedir um serviço de boleias a amigos ou colegas. Se preferires algo mais confortável, podes considerar alugar um carro por um dia ou usar uma aplicação de partilha de carros. Depende um pouco de onde estás e do teu orçamento, mas há sempre alternativas.

10

O texto apresenta uma marcação total de Português Europeu (PT-PT), sem qualquer vestígio de Português do Brasil (PT-BR).

1. **Forma de Tratamento** (...)
2. **Léxico e Vocabulário** (...)
3. **Verbo Associado ao Nome** (...)
4. **Pronominalização do Sujeito** (...)
5. **Ortografia** (...)

Figure 7: Example of the LLM Judge score and reasoning given to an answer from AMALIA-9B for No-Prompt setting. The colors represent the points in the answer the judge’s reasoning is addressing.

**EN Translation:** *Question:* My car is in the shop. What alternative transportation can I use? *Answer:* You have several options! You can catch a taxi, use public transport such as the bus or metro, or ask friends or colleagues for a ride. If you prefer something more comfortable, you can consider renting a car for a day or using a car-sharing app. It depends a bit on where you are and your budget, but there are always alternatives. **LLM Judge:** Score: 10 The text presents a total marking of European Portuguese (PT-PT), without any trace of Brazilian Portuguese (PT-BR). 1. Form of Address (...) 2. Lexicon and Vocabulary (...) 3. Verb Associated with the Noun (...) 4. Pronominalization of the Subject (...) 5. Spelling (...)

**Prompted for pt-BR**
**Turn 1**

---

**Question:**  
O meu carro está na oficina. Que transportes posso usar como alternativa?

**Answer:**  
Quando **seu** carro está na oficina, **você** pode considerar várias alternativas de transporte, dependendo da sua cidade e necessidades. Aqui estão algumas opções comuns:

1. **\*\*Transporte público:\*\***
- **Ônibus**
- **Metrô**
- **Trem urbano** (...)

0

O texto apresenta um enviesamento total para o português do Brasil (PT-BR), não contendo marcas linguísticas do português europeu (PT-EU).

1. **Forma de Tratamento** (...)
2. **Léxico e Vocabulário** (...)
3. **Ortografia** (...)

Figure 8: Example of the LLM Judge score and reasoning given to an answer from Sabiá-4 for a pt-BR prompt setting. The colors represent the points in the answer the judge’s reasoning is addressing.

**EN Translation:** *Question:* My car is in the repair shop. What transport can I use as an alternative? *Answer:* When your car is in the shop, you can consider various transport alternatives, depending on your city and needs. Here are some common options: 1. Public Transport: Bus Subway Urban train (...) **LLM Judge:** Score: 0 The text shows a total bias toward Brazilian Portuguese (PT-BR), containing no linguistic markers of European Portuguese (PT-EU). 1. Form of Address (...) 2. Lexicon and Vocabulary (...) 3. Spelling (...)

**Prompted for pt-PT**
**Turn 1**

---

**Question:**  
O meu carro está na oficina. Que transportes posso usar como alternativa?

**Answer:**  
Dependendo de **onde vive** e para onde **se desloca**, tem várias alternativas. Aqui estão as principais opções divididas por categorias:

1. Transportes Públicos
- É a opção mais comum e, geralmente, a mais **económica**.
- Autocarros** e **Metro**: Ideal para deslocações urbanas. Verifique a aplicação da operadora local (ex: **Carris**, **STCP**, **Metro Porto**). (...)

10

O texto apresenta uma marcação total de Português Europeu (PT-PT), sem qualquer vestígio de Português do Brasil (PT-BR).

1. **Léxico e Vocabulário** (...)
2. **Forma de Tratamento e Sujeito** (...)
3. **Ortografia** (...)

Figure 9: Example of the LLM Judge score and reasoning given to an answer from Gemma-4-31B for a pt-PT prompt setting. The colors represent the points in the answer the judge’s reasoning is addressing.

**EN Translation:** *Question:* My car is in the repair shop. What transport can I use as an alternative? *Answer:* Depending on where you live and where you travel to, you have several alternatives. Here are the main options divided by categories: 1. Public Transport It is the most common and, generally, the most economical option: Buses and Metro: Ideal for urban travel. Check the app of the local operator (ex: Carris, STCP, Metro Porto). (...) **LLM Judge:** Score: 10 The text presents a total marking of European Portuguese (PT-PT), without any trace of Brazilian Portuguese (PT-BR). 1. Lexicon and Vocabulary (...) 2. Form of Address and Subject (...) 3. Spelling (...)