

Portuguese-IFEval: Instruction-Following Evaluation in Portuguese

Emanuel B. Passinato, Gustavo P. Fontana, Cibelle P. Botelho, Júlia J. Biancolini
Diogo F. C. Silva & Arlindo R. Galvão Filho

Advanced Knowledge Center for Immersive Technologies (AKCIT)

Federal University of Goiás

emanuel.passinato@egresso.ufg.br

{pires.fontana,cibellepires}@discente.ufg.br

{juliabiancolini,diogo_fernandes}@discente.ufg.br

arlindogalvao@ufg.br

Abstract

Instruction-following is a core capability of Large Language Models (LLMs), as they must reliably comply with instructions constraints before being able to be evaluated on downstream tasks. In this scenario, instruction-following benchmarks for low-resource languages remain limited, particularly for Portuguese, where existing evaluations are mostly based on translated English datasets and fail to capture language specific linguistic phenomena. We introduce Portuguese-IFEval, a benchmark for evaluating instruction following in Portuguese built through semantic regionalization. The benchmark contains a total of 1236 instructions across 535 prompts, each prompt have up to three verifiable instructions, combining adapted instructions with Portuguese-specific orthographic, grammatical, and discourse-level constraints. We evaluate 25 proprietary, open-source, and Portuguese-tuned LLMs under a unified protocol, establishing a comprehensive benchmark for Portuguese instruction-following evaluation. Our results show that, while English and Spanish remain the strongest-performing languages, Portuguese still presents substantial challenges, which are effectively exposed by our benchmark. Our code is available at: <https://github.com/gap-ufg/Portuguese-IFEval>.

1 Introduction

Instruction following is a core capability of Large Language Models (LLMs) (Grattafiori et al., 2024; Yang et al., 2025a; Taori et al., 2023; Xu et al., 2024), motivating benchmarks that assess instruction compliance in zero-shot settings. Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) introduced deterministically verifiable constraints that avoid the bias, subjectivity, and cost of LLM-as-a-judge and manual evaluation protocols (Jiang et al., 2024; Zheng et al., 2023; Ouyang et al., 2022). IFEval contains 541 verifiable prompts with 25 categories of instructions.

Instruction Group	Count	Representation(%)
Detectable Content	364	29.45%
Keywords	239	19.34%
Detectable Format	201	16.26%
Length Constraints	148	11.97%
Change Cases	94	7.60%
Start / End Checker	81	6.55%
Punctuation	56	4.53%
Combination	47	3.80%
Language	6	0.48%
Total	1236	100 %

Table 1: Count and relative frequency of Portuguese-IFEval instructions for each instruction group, sorted by decreasing frequency.

However, IFEval is restricted to English, which limits its external validity in multilingual evaluation settings. M-IFEval (Dussolle et al., 2025) partially addresses this limitation by extending the benchmark to Spanish, Japanese, and French through localized and newly created instructions in each target language. This process produced 115 prompts for Spanish, 172 for Japanese, and 235 for French, expanding evaluation beyond English while preserving the original benchmark structure. Still, its experimental scope remains limited. The evaluation covers only eight models, with only one open-source model, which constrains reproducibility and comparative analysis. Furthermore, M-IFEval does not further investigate how using translated instructions directly affects model performance. Finally, it does not include Portuguese instructions.

In this work, we introduce Portuguese-IFEval, a benchmark designed to evaluate instruction following in Portuguese. Portuguese is a challenging testbed because of its fine-grained orthographic conventions, morphosyntactic constraints, and discourse-level distinctions, which are weakly represented in existing benchmarks. Rather than relying on direct translation, Portuguese-IFEval uses semantic regionalization, under the 25 original

IFEval instructions, to preserve deterministic verifiability while introducing new language-specific instructions targeting Portuguese-specific linguistic phenomena. Table 1 presents the resulting distribution of instruction categories.

We evaluate 25 language models, including proprietary systems, open-source instruction-tuned models, and Portuguese-specialized variants, under the standard IFEval evaluation protocol. This controlled setup enables direct comparison across instruction types and benchmark construction strategies. Our results show that, although larger and proprietary models generally achieve stronger instruction-following performance, scale alone does not eliminate language-specific failure modes. We also find that directly translating the original IFEval causes a shrinking gap between small and large models, for instance GPT-5 (OpenAI, 2025) achieves 73.15 points on the translated scenario, only 2 points ahead of Gemma 4 4B (Google, 2026), but when tested on Portuguese-IFEval the gap was near 18 points.

These findings motivate a broader reinterpretation of multilingual instruction-following evaluation. This work reframes instruction-following evaluation across languages as a *linguistic grounding problem*, rather than a translation problem: faithfully assessing instruction adherence requires benchmarks that explicitly encode language-specific orthographic, morphosyntactic, and pragmatic constraints, instead of assuming that translated instructions preserve evaluative fidelity.

Our main contributions and findings are summarized as follows:

- We introduce Portuguese-IFEval, a regionalized instruction-following benchmark that extends IFEval to Portuguese while preserving deterministically verifiable constraints and explicitly incorporating language-specific phenomena.
- Leveraging this benchmark, we conduct a rigorously controlled evaluation of state-of-the-art proprietary and open-source LLMs, establishing a comprehensive performance landscape for instruction following in Portuguese.
- This evaluation enables a fine-grained comparison between the original instruction set and Portuguese-specific instructions, revealing systematic language-dependent weak-

nesses that remain hidden under translated-only benchmarks.

- Finally, by contrasting translated and regionalized benchmark constructions under a unified evaluation protocol, we show that translation alone systematically masks language-dependent failure modes, establishing regionalized benchmark design as essential for faithful instruction-following evaluation in Portuguese.

2 Related Work

Instruction following has emerged as a central criterion for evaluating the practical reliability of large language models (LLMs), motivating the development of benchmarks that assess instruction compliance. A prominent line of work adopts an LLM-as-a-judge paradigm, as exemplified by InFoBench (Zeng et al., 2023) and FollowBench (Jiang et al., 2024). While flexible, this approach raises well-documented concerns regarding bias, subjectivity, and limited reproducibility (Zheng et al., 2023).

Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) introduced a deterministically verifiable framework for instruction-following evaluation, addressing these limitations by replacing subjective judgments with rule-based verification. This design enables objective, scalable, and reproducible evaluation. However, despite its methodological advantages, IFEval remains primarily focused on English instructions, limiting its applicability in multilingual settings.

To mitigate language bias, recent work has explored multilingual extensions of instruction-following benchmarks. M-IFEval (Dussolle et al., 2025) extends IFEval to Spanish, Japanese, and French by adapting existing instructions and introducing verifiable language-specific scenarios. Marco-Bench-MIF (Zeng et al., 2025) further broadens multilingual coverage by extending IFEval to 30 languages, including Portuguese, however, this benchmark relies exclusively on translated versions of the original instruction set and does not introduce language specific instruction design or validation. Recent multilingual evaluation efforts, such as the WMT25 Multilingual Instruction Shared Task (MIST) (Kocmi et al., 2025), demonstrate substantial variation in LLM performance across languages and highlight persistent challenges in the reliability of multilingual evaluation. These findings motivate the development of

evaluation frameworks that consider the specific characteristics of each language, rather than treating multilingual assessment as a direct extension of English-centered standards.

As a result, no prior work has introduced an native instruction-following benchmark for Portuguese that preserves deterministic verification while incorporating language-specific constraints. Existing multilingual benchmarks either exclude Portuguese or treat it solely through translation, which can mask language dependent instruction following behavior. In this work, we introduce Portuguese-IFEval, a benchmark designed to evaluate instruction following in Portuguese with semantically regionalized and constructed Portuguese instructions, enabling systematic analysis of instruction-following capabilities in Portuguese.

3 Method

3.1 Overview

This section describes the construction of instruction prompts for **IFEval-PT**, grounded in the original Instruction-Following Evaluation (IFEval) framework (Zhou et al., 2023) and aligned with the multilingual design principles of M-IFEval (Dus-solle et al., 2025). The benchmark preserves the core notion of *verifiable instructions*, enabling objective and deterministic evaluation, while extending instruction-following assessment to Portuguese through semantic regionalization rather than direct translation. This design follows the central finding of M-IFEval that literal translation is insufficient to capture language-specific instruction-following behavior.

3.2 Instruction Set Regionalization

We start from the set of verifiable instruction categories defined in the original IFEval benchmark. Instead of translating instruction descriptions and validation rules verbatim, we perform semantic regionalization into Portuguese.

This process adapts lexical choices, syntactic constructions, and pragmatic formulations to reflect how instructions are naturally expressed and interpreted in Portuguese, while preserving the logical structure required for deterministic verification. As a result, instruction semantics remain formally verifiable while their linguistic realization aligns with native usage.

3.3 Portuguese-specific Instructions

In addition to regionalizing existing instructions, we introduce six Portuguese-specific instructions following the methodology of M-IFEval. Each instruction targets a distinct linguistic phenomenon, spans varying difficulty levels, and remains deterministically verifiable.

Three instructions focus on orthography and diacritics, including verification of cedilla (ç) frequency, enforcement of tilde absence (ã/õ), and correct use of the grave accent (à) in crase constructions. Two instructions target grammar and pragmatics, including correct mesoclis usage and appropriate use of the second-person plural form (vós) in direct address. One instruction additionally requires discourse-level reasoning by enforcing the canonical four-way distinction among *por que*, *por quê*, *porque*, and *porquê* within context.

Figure 1 illustrates one such instruction, which requires exactly five occurrences of the cedilla character (ç). In this example, GPT-4o fails to satisfy the constraint, whereas GPT-5 correctly follows it.

3.4 Prompt Generation

We generate an initial pool of prompts using the Qwen/Qwen3-235B-A22B-FP8 model. Each prompt contains between one and three verifiable instructions, following the instruction composition strategy of IFEval and M-IFEval. All instructions are generated from scratch using the regionalized descriptions, without seeding from existing benchmarks.

Each prompt also includes a high-level directive specifying the desired text style, such as an email, summary, dialogue, recipe, or social media post. These directives increase pragmatic diversity without affecting deterministic verifiability.

This process results in 595 candidate prompts.

3.5 Prompt Rewriting

After instruction composition, we rewrite the prompts using the Qwen3-235B-A22B model (Yang et al., 2025b). This step aims to improve linguistic naturalness and reduce templatic phrasing. This step follows the rephrasing procedure used in the original IFEval benchmark and preserves all verifiable constraints.

3.6 Automatic Satisfiability Filtering

Combining multiple verifiable instructions can result in unsatisfiable constraint sets. We therefore apply an automatic satisfiability filtering stage, where

Portuguese Instruction-Following Examples (Cedilha Frequency)

Instruction: Escreva um parágrafo sobre frutas ou doces. Na sua resposta, a letra ‘ç’ deve aparecer exatamente 5 vezes. **Translated Instruction:** Write a paragraph about fruits or sweets. In your response, the letter “ç” must appear exactly 5 times.

GPT-4o (Instruction Not Followed): Frutas são uma escolha saudável e deliciosa para qualquer lanche. A maçã, por exemplo, é rica em fibras e vitaminas, sendo uma opção excelente para quem busca uma alimentação equilibrada. Já a acerola é conhecida por seu alto teor de vitamina C, essencial para o fortalecimento do sistema imunológico. Além disso, a casca de muitas frutas contém antioxidantes que combatem os radicais livres. É importante lembrar que o consumo de frutas deve ser variado, garantindo assim uma ampla gama de nutrientes. Portanto, incluir frutas na dieta diária é uma decisão sensata e saborosa.

GPT-5 (Instruction Followed): As frutas são essenciais para uma alimentação saudável e equilibrada, oferecendo uma vasta gama de vitaminas, minerais e fibras que ajudam no funcionamento do corpo. Entre as mais populares, encontramos a maçã, que é versátil e pode ser consumida crua ou em receitas diversas. A laranja, rica em vitamina C, é perfeita para sucos refrescantes. Já os doces, como o brigadeiro e o quindim, deliciam qualquer paladar, mas devem ser consumidos com moderação. A combinação de frutas com doces, como na preparação de uma torta de banana ou uma mousse de maracujá, resulta em sobremesas irresistíveis que agradam a todos.

Figure 1: Portuguese instruction-following examples for the cedilha frequency constraint. The GPT-4o violates the exact occurrence requirement, while the GPT-5 satisfies it.

the model evaluates whether all instructions in a prompt can be simultaneously satisfied.

Prompts with contradictory constraints are discarded. After this stage, 574 prompts remain.

3.7 Manual Verification

All remaining prompts were subjected to a structured manual verification process to identify and remove unsatisfiable or linguistically ambiguous cases. The review was conducted independently by three NLP researchers, all native Portuguese speakers. Although not specialists in linguistics, all annotators have experience in instruction-following evaluation and benchmark design.

The primary objective of the verification process was to detect unsatisfiable prompts, defined as prompts for which no valid response could simultaneously satisfy all instructions. Annotators were also instructed to flag cases involving syntactic or semantic ambiguity, as well as mutually incompatible instruction constraints. The review process was performed in 5 phases:

1. Each annotator independently reviewed every prompt.
2. The annotator attempted to construct a response satisfying all instructions present in

the prompt (i.e., generate a valid response).

3. The candidate response was evaluated using the benchmark’s automatic validation code to verify that all constraints were satisfied.
4. If no valid response could be constructed, or if the prompt contained linguistic ambiguities or conflicting instructions, it was flagged for further review.
5. For prompts where the annotators disagreed, they discussed whether the prompt could be fully satisfied. If both agreed that it was satisfiable, the prompt was retained as valid; otherwise, it was discarded.

After the independent review phase, disagreements were resolved through discussion until consensus was reached. Out of the initial 574 prompts, 39 were removed following this procedure, resulting in a final benchmark of 535 validated prompts.

3.8 Translation Setup

The original IFEval dataset was translated using the OpenAI GPT-5 model with the default API hyperparameters. To obtain a fully functional translated benchmark, some evaluation routines also

required adaptation. For example, the Constrained Response instruction asks the model to answer using one option from a predefined set specified in the evaluation code. After translation, these options are rendered in Portuguese, making the original verification procedure invalid. Consequently, the evaluation code also needs to be modified to support the translated answer options.

A similar issue occurs with the *number of sentences* instruction. The original implementation relies on an English-specific sentence tokenizer to count the number of sentences, which makes the verifier unreliable for Portuguese text. Therefore, the evaluation logic also needed to be adapted to use a Portuguese-compatible tokenizer.

In summary, even when using direct translation to adapt IFEval, several instructions require additional modifications before the full dataset can be reliably evaluated. As a result, some prior translations retained only the instructions that could be directly translated without requiring changes to the evaluation code (Qwen et al., 2025).

To enable a fair comparison with our dataset, we translated all 541 instructions from IFEval and adapted the evaluation code whenever necessary, allowing the complete benchmark to be used in Portuguese.

4 Experiments

4.1 Experimental Setup

We evaluate the benchmark using a diverse set of state-of-the-art language models, covering open-source instruction-tuned models, proprietary systems, and Portuguese-specialized variants.

For open-source instruction-tuned models, we consider the Qwen3 family, Qwne 3.5 family (0.8B to 9B), (Yang et al., 2025a) (0.6B to 8B), Gemma 4 (Google, 2026) (2B and 4B), LLaMA 3.1 (Grattafiori et al., 2024) (8B), LLaMA 3.2 (Grattafiori et al., 2024) (1B and 3B) and Nemotron 3 Nano (NVIDIA et al., 2025) (4B).

We additionally evaluate proprietary models, including GPT-4o and GPT-5 (OpenAI et al., 2024; OpenAI, 2025), Claude 4.5 Haiku (Anthropic, 2025), and Gemini 2.5 Pro (Comanici et al., 2025).

To assess the impact of language-specific training, we also include Portuguese-tuned models, namely Tucano (1B and 2B) (Corrêa et al., 2025), Tucano 2 (0.5B to 3.7B) (Corrêa et al., 2026), Gaia (4B) (CAMILO-JUNIOR, 2025), and the proprietary Sabiá 3 (Abonizio et al., 2025).

4.2 Metrics

Generated responses are evaluated using the M-IFEval evaluation code, supplemented with our own extensions. Following the original IFEval protocol, we compute both strict and loose instruction-following scores for each instruction.

The strict metric evaluates whether a prediction y exactly satisfies the required instruction i , as defined in Equation 1:

$$\text{Followed}(y, i) = \begin{cases} \text{True}, & \text{if followed,} \\ \text{False}, & \text{otherwise.} \end{cases} \quad (1)$$

The loose metric mitigates known false negatives by applying a set of transformations t to the model output. If the instruction is satisfied after any transformation, the prediction is considered successful. This metric is defined in Equation 2:

$$\text{Followed}(t(y), i) = \begin{cases} \text{True}, & \text{if followed,} \\ \text{False}, & \text{otherwise.} \end{cases} \quad (2)$$

For prompt-level composition, a model is considered successful only if it satisfies all instructions contained in the prompt. This requirement applies to both the Strict and Loose variants, formally defined in Equations 3 and 4, respectively.

$$\text{Strict}_{\text{prompt}}(y, I) = \bigwedge_{i \in I} \text{followed}(y, i) \quad (3)$$

$$\text{Loose}_{\text{prompt}}(y, I) = \bigwedge_{i \in I} \text{followed}(t(y), i) \quad (4)$$

We report Strict, Loose, and averaged scores for all evaluated models on the Portuguese benchmark (Table 4). Additionally, we present: (i) the average instruction-level Strict score across the language-specific instruction sets of M-IFEval and Portuguese-IFEval (Table 5); (ii) a category-level breakdown of Strict performance on Portuguese-IFEval (Figure 2); and (iii) the average prompt-level Strict score computed across IFEval, M-IFEval, and Portuguese-IFEval, disaggregated by language (Table 3).

5 Results

5.1 Native vs. Translated Benchmark

Table 2 compares model performance on Portuguese-IFEval and the directly translated

Model	Strict (Ours) $\uparrow$	Loose (Ours) $\uparrow$	Strict (T*) $\uparrow$	Loose (T*) $\uparrow$	Δ Strict $\downarrow$	Δ Loose $\downarrow$
<i>Instruct Models</i>						
Llama 3.1 it (8B)	27.28 $\pm$ 0.62	31.21 $\pm$ 0.35	56.87 $\pm$ 0.70	61.86 $\pm$ 0.83	29.59	30.65
Gemma 4 (E4B)	46.69 $\pm$ 0.72 $\dagger$	50.50 $\pm$ 0.49 $\dagger$	71.78 $\pm$ 0.09 $\dagger$	73.69 $\pm$ 0.09 $\dagger$	25.09	23.19 $\dagger$
Qwen 3 (8B)	45.13 $\pm$ 0.93	48.88 $\pm$ 0.53	69.25 $\pm$ 0.09	72.64 $\pm$ 0.00	24.12 $\dagger$	23.77
Nemotron 3 Nano (4B)	39.70 $\pm$ 0.00	42.51 $\pm$ 0.00	71.66 $\pm$ 0.09	73.57 $\pm$ 0.00	31.96	31.06
<i>API Models</i>						
Claude 4.5 Haiku	48.95	53.92	65.19	70.37	16.24	16.45
GPT-4o	48.57	53.92	65.00	69.26	16.43	15.34
GPT-5	65.20	67.30	73.15	75.93	7.95	8.63
Gemini Pro	58.51	63.10	67.22	70.56	8.71	7.76
Sabia 3	47.99	52.96	67.22	69.81	19.23	16.85
<i>PT-Specific Models</i>						
Gaia (4B)	30.02 $\pm$ 0.97 $\dagger$	33.83 $\pm$ 0.88 $\dagger$	51.14 $\pm$ 0.17 $\dagger$	60.44 $\pm$ 0.15 $\dagger$	21.11	26.61
Tucano2 Qwen (3.7B)	23.10 $\pm$ 0.87	26.53 $\pm$ 1.02	43.38 $\pm$ 0.46	46.70 $\pm$ 0.97	20.28 $\dagger$	20.17 $\dagger$

Table 2: Portuguese-IFEval results. Scores comparing Portuguese (Strict (Ours) and Loose (Ours)) with the translated version (T* Strict and T* Loose) and their difference (Δ Strict and Δ Loose). For non-api models the result is the average of 3 runs of each model with its standard deviation (see Appendix C for details.). The highest-scoring model across the entire table is highlighted in bold, and the best model within each category is marked with $\dagger$.

version of the original IFEval. Among the open-source models, Gemma 4 achieves the strongest overall performance, followed by Qwen 3. Within the API-based models, GPT-5 obtains the highest scores, while Sabia 3 leads among the Portuguese-specialized models. Overall, the ranking of the top-performing models remains relatively consistent across both benchmarks.

However, the translated benchmark substantially compresses the performance gap between models. For example, in the strict setting, the difference between the best-performing open-source model, Gemma 4 (71.78), and GPT-5 (73.15) is only 1.37 points. This result could suggest that both models exhibit nearly equivalent Portuguese instruction-following capabilities. In contrast, under the Portuguese-IFEval protocol, the gap widens to approximately 18 points (46.69 vs. 65.20), revealing a much clearer separation between the models. While it is generally expected that state-of-the-art proprietary models outperform smaller open-source models, the translated benchmark largely obscures this difference. In contrast, the regionalized benchmark preserves a clearer separation between model capabilities, making it more effective at discriminating instruction-following performance.

We also observe that the discrepancy between Portuguese-IFEval and the translated benchmark is consistently larger for open-source and Portuguese-

specialized models than for proprietary API systems. Additionally, the translated benchmark consistently exhibits lower standard deviation values across repeated evaluations compared to Portuguese-IFEval. This reduced variance suggests that the translated setting produces more stable and homogeneous model behavior, whereas Portuguese-IFEval exposes greater inconsistencies in instruction-following performance.

Taken together, these results provide consistent evidence that Portuguese-IFEval constitutes a more challenging and discriminative evaluation setting than a direct translation of IFEval.

5.2 Cross-lingual Robustness Across Instruction Sets

We examined how the models generalize across languages. Instruction-following accuracy varies markedly across Portuguese, Spanish, French, and Japanese, revealing differences in cross-lingual robustness. Among open-source instruct models, Gemma 4 (4B) achieves the highest overall multilingual average (71.93), with relatively balanced performance across all four languages, in exception of Spanish where Nemotron 3 Nano outperforms Gemma. Smaller models consistently underperform across all languages, reinforcing the role of scale in multilingual instruction adherence. Proprietary API models dominate this evaluation, with GPT-5 achieving the highest score in every lan-

Model	EN	ES	FR	JA	PT	Avg
<i>Instruct Models</i>						
Llama 3.1 it (8B)	70.98	77.39	63.40	36.63	26.14	54.91
Llama 3.2 it (3B)	67.10	65.22	53.62	25.58	21.62	46.63
Llama 3.2 it (1B)	44.36	37.39	32.34	16.28	11.94	28.46
Gemma 4 E2B	79.11	86.96	74.04	56.40	36.52	66.61
Gemma 4 E4B	86.69 †	88.70	81.70	58.14 †	46.69 †	71.93 †
Qwen 3 (8B)	85.40	83.48	82.13 †	54.07	42.96	69.61
Qwen 3 (4B)	84.10	84.35	78.30	51.74	39.48	67.59
Qwen 3 (1.7B)	69.50	80.00	66.81	45.35	31.54	58.64
Qwen 3 (0.6B)	58.78	73.91	46.38	35.47	23.77	47.66
Qwen 3.5 (9B)	84.10	83.48	77.02	43.60	42.99	66.24
Qwen 3.5 (4B)	82.26	82.61	75.32	43.02	39.94	64.63
Qwen 3.5 (2B)	66.36	69.57	61.70	29.65	26.98	50.85
Qwen 3.5 (0.8B)	53.23	60.87	39.15	22.67	17.13	38.61
Nemotron 3 Nano (4B)	84.10	89.57 †	80.00	55.23	39.63	69.71
<i>API Models</i>						
Claude 4.5 Haiku	85.40	83.48	84.26	59.30	45.38	71.56
GPT-4o	84.10	87.83	82.98	63.37	50.80	73.82
GPT-5	93.90	97.39	90.64	88.37	65.91	87.24
Gemini Pro	90.57	89.57	88.09	66.28	77.78	82.46
Sabia 3	80.04	85.22	77.45	47.67	49.73	68.02
<i>PT-Specific Models</i>						
Gaia (4B)	65.80 †	70.43 †	62.98 †	43.02 †	28.12 †	54.07 †
Tucano2 Qwen (0.5B)	23.11	33.04	20.85	20.93	12.81	22.15
Tucano2 Qwen (1.5B)	33.27	53.91	32.77	20.35	18.67	31.79
Tucano2 Qwen (3.7B)	42.88	61.74	50.64	22.09	22.32	39.93
Tucano it (1B)	12.75	26.09	25.11	22.67	4.93	18.31
Tucano it (2B)	9.98	24.35	13.19	21.51	3.25	14.46

Table 3: Instruction-following strict score, prompt level, across all languages. For non-api models the result is the average of 3 runs of each model. The highest-scoring model across the entire table is highlighted in bold, and the best model within each category is marked with †.

guage and an overall average of 87.24, followed by Gemini 2.5 Pro. Among Portuguese-specific models, Gaia exhibits the strongest cross-lingual robustness, while while Tucano models show limited performance. Full multilingual results are shown in Table 3.

Overall the models perform worst in the Portuguese set of instructions, revealing the challenging of this language.

5.3 Overall Performance on Portuguese Instructions.

Across all evaluated models, instruction-following performance on Portuguese exhibits a clear stratification by model scale and training regime. Among open-source instruction-tuned models, the Gemma 4 (4B) leads performance achieving the highest average score (44.45), followed by Qwen 3 (8B) 42.98 and Qwen 3.5 (9B) 42.99, even being a smaller model. We also note that between Qwen 3 and 3.5 the former one performs even (8B and 4B) or even better (1.7B and 0.6B) than the newest version, we argue that this can be due the multi-modal nature of the Qwen 3.5 family. Among the small ones ($\leq 1B$) Qwen 3 (0.6B) leads with a 6 points advantage.

Proprietary API-based models dominate overall

Model	Strict	Loose	Avg
<i>Instruct Models</i>			
Llama 3.1 it (8B)	26.15 $\pm$ 0.18	29.94 $\pm$ 0.62	28.05
Llama 3.2 it (3B)	21.64 $\pm$ 0.70	24.82 $\pm$ 0.54	23.23
Llama 3.2 it (1B)	11.95 $\pm$ 0.20	13.70 $\pm$ 0.66	12.83
Gemma 4 (E2B)	36.56 $\pm$ 1.06	40.17 $\pm$ 1.31	38.37
Gemma 4 (E4B)	46.69 $\pm$ 0.72 †	48.23 $\pm$ 1.31 †	46.34 †
Qwen 3 (8B)	42.98 $\pm$ 0.66	46.83 $\pm$ 1.05	44.91
Qwen 3 (4B)	39.50 $\pm$ 0.51	43.30 $\pm$ 0.10	41.40
Qwen 3 (1.7B)	31.55 $\pm$ 0.20	35.70 $\pm$ 0.17	33.63
Qwen 3 (0.6B)	23.73 $\pm$ 0.81	27.56 $\pm$ 0.57	25.65
Qwen 3.5 (9B)	42.99 $\pm$ 0.00	47.29 $\pm$ 0.00	45.14
Qwen 3.5 (4B)	39.94 $\pm$ 0.60	43.18 $\pm$ 0.19	41.56
Qwen 3.5 (2B)	26.98 $\pm$ 1.03	30.59 $\pm$ 0.76	28.79
Qwen 3.5 (0.8B)	17.13 $\pm$ 0.43	19.88 $\pm$ 0.43	18.51
Nemotron 3 Nano (4B)	39.63 $\pm$ 0.00	42.43 $\pm$ 0.00	41.03
<i>API Models</i>			
Claude 4.5 Haiku	48.95	53.92	51.44
GPT-4o	48.57	53.92	51.25
GPT-5	65.20	67.30	66.25
Gemini 2.5 Pro	58.51	63.10	60.81
Sabia 3	47.99	52.96	50.48
<i>PT-Specific Models</i>			
Gaia (4B)	28.13 $\pm$ 0.40 †	32.13 $\pm$ 0.34 †	30.13 †
Tucano2 Qwen (0.5B)	12.84 $\pm$ 1.09	16.47 $\pm$ 0.60	14.66
Tucano2 Qwen (1.5B)	18.69 $\pm$ 0.73	22.33 $\pm$ 0.71	20.51
Tucano2 Qwen (3.7B)	22.30 $\pm$ 0.51	25.72 $\pm$ 0.76	24.01
Tucano it (1B)	4.93 $\pm$ 0.35	5.63 $\pm$ 0.22	5.28
Tucano it (2B)	3.25 $\pm$ 0.59	4.23 $\pm$ 0.68	3.74

Table 4: Portuguese-IFEval results. We report both Strict and Loose metrics for the evaluated models on Portuguese-IFEval, considering the full instruction set, which comprises Portuguese-specific instructions as well as localized versions of the original IFEval instructions adapted to Portuguese). For non-api models the result is the average of 3 runs of each model with its standard deviation. The highest-scoring model across the entire table is highlighted in bold, and the best model within each category is marked with †.

performance, led by GPT-5 with an average score of 65.20, followed by Gemini 2.5 Pro, with the other tree models being really close to each other and close to Gemma 4 on strict scenarios, showing that these models can follow the instructions but lack some final refinement, this result is confirmed by the Loose metric where after the transformations the api models gain a larger margin compared to Gemma 4. Among Portuguese-specialized models, Gaia achieves the strongest results (28.13), outperforming the Tucano variants. Detailed results are reported in Table 4.

This results indicates that even with the effort of creating Portuguese tuned models, the opens ource multilingual models sill has a strong performance within the same model size range.

Model	ES	FR	JA	PT	Avg
<i>Instruct Models</i>					
Llama 3.1 (8B)	52.50	52.71	23.81	25.53	38.64
Llama 3.2 (3B)	55.00	57.39	25.69	23.85	40.48
Llama 3.2 (1B)	40.00	44.40	19.05	18.92	30.59
Gemma 4 (E2B)	52.50	54.31	49.21	36.42	48.11
Gemma 4 (E4B)	57.50	63.52	51.09 †	46.58 †	54.67 †
Qwen 3 (8B)	52.50	63.19	40.08	32.11	46.97
Qwen 3 (4B)	57.50	56.07	40.67	28.79	45.76
Qwen 3 (1.7B)	55.00	56.08	27.78	21.57	40.11
Qwen 3 (0.6B)	47.50	36.16	29.76	13.85	31.82
Qwen 3.5 (9B)	52.50	61.68	36.41	37.08	46.92
Qwen 3.5 (4B)	50.00	60.70	40.38	31.12	45.55
Qwen 3.5 (2B)	42.50	55.22	28.77	21.50	37.00
Qwen 3.5 (0.8B)	37.50	47.31	22.12	17.23	31.04
Nemotron 3 Nano (4B)	72.50 †	67.51 †	38.79	21.94	50.18
<i>API Models</i>					
Claude 4.5 Haiku	47.50	87.22	54.07	59.88	62.17
GPT-4o	55.00	81.47	58.23	49.13	60.96
GPT-5	100.00	88.94	89.38	74.60	88.23
Gemini 2.5 Pro	70.00	91.41	65.28	66.67	73.34
Sabia 3	57.50	57.79	34.52	49.54	49.84
<i>PT-Specific Models</i>					
Gaia (4B)	47.50 †	52.81 †	25.00	33.26 †	39.64 †
Tucano2 Qwen (0.5B)	12.50	39.54	34.62	20.30	26.74
Tucano2 Qwen (1.5B)	30.00	36.56	33.43	19.13	29.78
Tucano2 Qwen (3.7B)	32.50	50.43	39.38	27.08	37.35
Tucano it (1B)	15.00	34.77	44.05	10.34	26.04
Tucano it (2B)	22.50	42.48	46.43 †	12.63	31.01

Table 5: Strict scores on instruction level of only language-specific instructions across languages. Portuguese, Spanish, French, and Japanese results are reported as well as average result between the languages. For non-api models the result is the average of 3 runs of each model. The highest-scoring model across the entire table is highlighted in bold, and the best model within each category is marked with †.

5.4 Impact of Portuguese-specific Instructions.

While cross-lingual evaluation highlights generalization capacity, it does not isolate language-specific constraints. To this end, we evaluated the models performance constrained to only language-specific instructions on table 5. This comparison reveals a systematic degradation in accuracy. Models with moderated overall performance under standard Portuguese-IFEval, including Qwen 3 (8B), Nemotron 3 Nano, and GPT-4o, exhibit consistent drops when evaluated on Portuguese-specific instructions. The degradation is most pronounced for smaller instruct models and Portuguese-tuned variants such as Tucano. However for the top performing models like GPT-5, Gemini 2.5 Pro and Gemma 4 this degradation does not appear or the models are even better at the specif instruction set showing that these models are even stronger in Portuguese instruction flowing than the first result suggested.

6 Conclusion

In this work, we introduced Portuguese-IFEval, a language-aware benchmark for evaluating instruction-following capabilities in Portuguese. By introducing Portuguese-specific instructions and adapting the evaluation pipeline to support language-dependent validation, we showed that direct translation alone is insufficient to reliably assess multilingual instruction-following performance.

Across 25 language models, our experiments reveal substantial performance degradation when models are evaluated under Portuguese specific constraints. While larger proprietary systems such as GPT-5 and Gemini 2.5 Pro consistently achieved the strongest overall performance, smaller open-source and Portuguese-specialized models exhibited significantly larger drops, particularly under strict evaluation settings. Importantly, we showed that translated versions of IFEval compress the performance gap between models, whereas Portuguese-IFEval exposes clearer differences in robustness and instruction adherence. The benchmark also produced higher variance across repeated evaluations, suggesting greater sensitivity to consistency failures and language-specific behaviors.

Our multilingual analysis further demonstrated that instruction-following robustness varies considerably across languages, even for highly capable models. Although scaling generally improves performance, the results indicate that scale alone does not eliminate failures associated with orthographic, morphosyntactic, and culturally grounded constraints. We also find out that Portuguese is the worst performing language among the evaluated ones, showing that models still struggle at that low resource language.

Taken together, these findings show that our benchmark effectively exposes the fragility of current models in Portuguese instruction-following. The results further indicate a substantial gap in instruction-following capabilities when models are evaluated on low-resource languages. Finally, we hope that the design of Portuguese-IFEval encourages future research on culturally and linguistically grounded benchmarks for low-resource languages.

Building upon these findings, several promising directions for future work emerge. First, to better reflect evolving real world applications, future iterations of benchmarks should incorporate increasingly complex constraints. This includes

expanding beyond standard instruction-following evolving in to tool-calling scenarios, multi-turn dialogues, and agentic workflows. Additionally, future research should assess how well models adapt to dialectal diversity and localized nuances across the Lusophone languages, evaluating performance differences across European Portuguese, regional Brazilian variations, and other varieties of the language.

Limitations

Despite the contributions of this study in evaluating instruction-following capabilities of large language models in Portuguese, several limitations should be acknowledged. First the benchmark relies predominantly on curated prompts that may not fully capture the diversity of real-world instructions, potentially limiting external validity. Second while we include a cross-lingual comparison with English results, differences in cultural and linguistic contexts may influence model performance, and observed discrepancies may not generalize to other Portuguese dialects or informal usage. Third as the curation processes was lead by the research team, this may lead to biased data construction. Fourth, as the benchmark comprises 535 prompts, its scale may still be insufficient to fully capture the variability of real-world instruction-following scenarios, highlighting the need for future expansion to improve coverage and robustness.

Finally, the evaluation methodology primarily assesses strict instruction adherence, which may overlook subtler aspects of model behavior such as partial compliance, nuanced reasoning, or pragmatic interpretation of instructions. Consequently, models that perform well under the strict metric may still exhibit limitations in flexible or creative instruction-following scenarios.

Acknowledgments

This work has been funded by the project Research and Development of Genese Digital: Scaling Interactive and Culturally Adapted Digital Humans with Generative AI, supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, grant number 057/2023, signed with EMBRAPII, and supported by P&D CEMIG/ANEEL PD-04950-D0677/2023.

References

- Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. *Sabiá-3 technical report*. *Preprint*, arXiv:2410.12049.
- Anthropic. 2025. Introducing claude haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Anthropic News Blog. Accessed: 2026-05-23.
- S. S. T.; PEREIRA L. A.; AMADEUS M.; SCOTTI R.; FAZZIONI D.; NOVAIS A. M. A.; JORDÃO S. A. A. CAMILO-JUNIOR, C. G.; OLIVEIRA. 2025. Gaia: An open language model for brazilian portuguese. <https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it>.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. *Preprint*, arXiv:2507.06261.
- Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. *Tucano: Advancing neural text generation for portuguese*. *Patterns*, 6(11):101325.
- Nicholas Kluge Corrêa, Aniket Sen, Shiza Fatimah, Sophia Falk, Lennard Landgraf, Julia Kastner, and Lucie Flek. 2026. *Tucano 2 cool: Better open source llms for portuguese*. *Preprint*, arXiv:2603.03543.
- Antoine Dussolle, A Cardeña, Shota Sato, and Peter Devine. 2025. M-ifeval: Multilingual instruction-following evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6161–6176.
- Google. 2026. Gemma 4: Our most capable open models to date. <https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/>. Google Blog. Accessed: 2026-05-23.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjuan Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. *Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmitova, Mariya Shmatova, and Vilém Zouhar. 2025. [Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.
- NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, and 295 others. 2025. [Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning](#). *Preprint*, arXiv:2512.20848.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>. Accessed: 2026-01-02.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. [WizardLM: Empowering large pre-trained language models to follow complex instructions](#). In *The Twelfth International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, and 1 others. 2025. Marco-bench-mif: On multilingual instruction-following capability of large language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072.
- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Appendix

A.1 Category-level Failure Patterns in Portuguese.

We conduct additional experiments on the fail by category using the top performing models. GPT-5 consistently achieves the highest accuracy across nearly all instruction categories, while showing slightly lower performance on Detectable Content, Detectable format, language and Start/End constraints. Gemma 4 (4B) exhibits solid but uneven behavior, matching GPT-5 on Language constraints

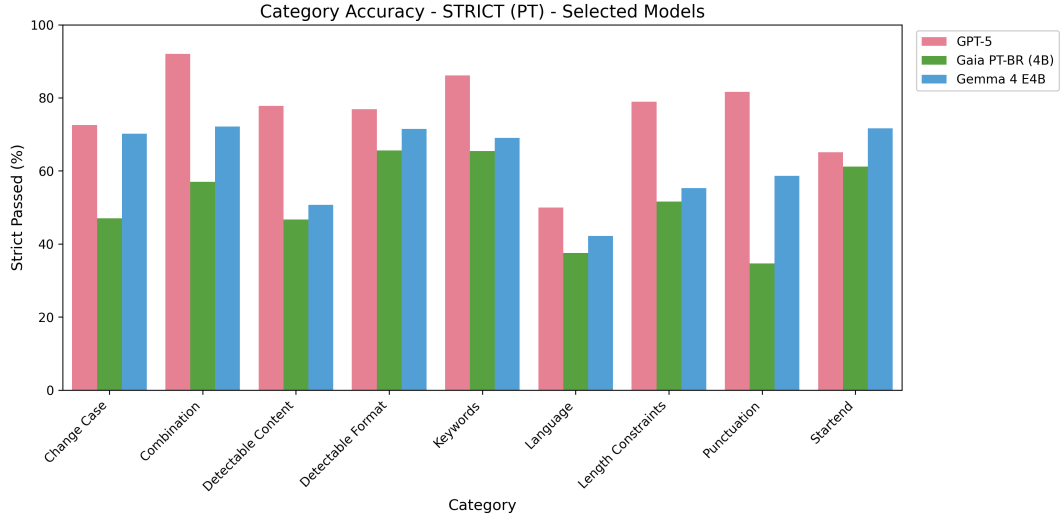


Figure 2: Strict instruction-following accuracy by category for Portuguese (PT). Reporting only top performing models of each category.

and Change Case but dropping sharply on content-sensitive categories, particularly Detectable Content and Length Constraints. Gaia typically falls between GPT-5 and Gemma 4, with the large performance drop at Change Case and Punctuation. These patterns show that instruction following in Portuguese is most reliable for language and formatting constraints, whereas content-sensitive and boundary-related instructions constitute the primary sources of error, especially for non-API-scale models. The full category breakdown is shown in Figure 2.

A.2 Instructions

A.2.1 Shared with English Instructions

The Table 8 shows all the instructions that are shared between English and Portuguese.

A.2.2 Portuguese-specific Instructions

The Table 9 shows all the Portuguese-specific instructions that are create on our work.

A.2.3 Operational Acceptance Conditions for Portuguese-specific Instructions

The Portuguese-specific validators are deterministic, rule-based checkers. Acceptance is based on explicit surface-form conditions rather than full syntactic parsing, so the descriptions below follow the evaluation code directly.

This follows the multilingual validation approach of M-IFEval (Dussolle et al., 2025): language-specific constraints are operationalized through reproducible rules tailored to Portuguese.

Cedilla Frequency. The cedilla-frequency checker accepts a response if the lowercased output contains exactly N occurrences of the literal character ç. Only the explicit character ç is counted; no normalization or approximation is applied.

Tilde Absence. The tilde-absence checker rejects a response only if it contains at least one character from the closed set ã, õ, ñ, Â, Ô, or Ñ. Other accented characters, including grave accents and cedilla, remain admissible.

Grave Accent (Crase). The crase checker accepts a response if it contains at least one case-insensitive match to à, às, àquele, àquela, àqueles, àquelas, or àquilo. It requires the presence of at least one recognized crase form.

Mesoclis Usage. The mesoclis checker accepts a response if it contains at least one hyphenated token matching the pattern stem-pronoun-suffix. The clitic must belong to a fixed list of pronouns and the suffix to a fixed list of future or conditional endings. The checker tests only surface-form presence and does not verify broader grammatical or semantic appropriateness.

Second-Person Plural Address (vós). The *vós* checker uses a lexical approximation of second-person plural address. It rejects competing forms such as você, vocês, or tu; requires at least one token from the *vós* family (*vós*, *vos*, *convosco*, *vosso*, *vossa*, *vossos*, *vossas*); and then searches within a five-token window for a compatible verb.

Compatibility is determined either by a fixed list of common *vós* forms or by endings such as -ais, -eis, -astes, -estes, -istes, -areis, -ereis, -ireis, and -ieis.

Porque/porquê/por que/por quê. The *porquê* checker first requires exactly one occurrence of a form in {*porque*, *porquê*, *porquês*, *por que*, *por quê*}. If zero or more than one form appears, the response is rejected. The final decision then depends on the target subtype specified in the prompt.

For *porque*, the match must be exactly *porque*, must not be preceded by a determiner, must not be phrase-final, and must not be sentence-initial in an interrogative sentence. For *porquê*, the match must be *porquê* or *porquês* and must be immediately preceded by a determiner. For *por que*, the match must be exactly *por que*, must not be preceded by a determiner or be phrase-final, and is accepted only in a question, after a relative noun such as *motivo* or *razão*, or after an indirect-question trigger. For *por quê*, the match must be exactly *por quê* and must be followed by punctuation or by the end of the utterance. Context is computed within the nearest sentence boundaries marked by ., ?, or !.

A.2.4 Illustrative Success and Failure Examples

Below, we present one success case and one failure case for each Portuguese-specific instruction, following the appendix style of M-IFEval.

Success Example: Cedilla Frequency

Model: GPT-5 **Language:** Portuguese (PT)

Prompt: “Escreva um parágrafo sobre frutas ou doces. Na sua resposta, a letra ‘ç’ deve aparecer exatamente 5 vezes.”

Model Response: “Adoro sobremesas que misturam frutas e costumes: a maçã cozida com açúcar e canela, a paçoca que lembra a infância e o açaí bem gelado. Em festas, um toque de cachaça pode acentuar o aroma, sem roubar a leveza do prato.”

Instruction (translated): The character ç should appear exactly 5 times.

Success: The response contains exactly five occurrences: *maçã*, *açúcar*, *paçoca*, *açaí*, and *cachaça*.

Failure Example: Cedilla Frequency

Model: Qwen3.5-0.8B **Language:** Portuguese (PT)

Prompt: “Escreva um parágrafo sobre frutas ou doces. Na sua resposta, a letra ‘ç’ deve aparecer exatamente 5 vezes.”

Model Response: “A fruta com a letra ‘ç’ aparece cinco vezes ao longo do texto.”

Instruction (translated): The character ç should appear exactly 5 times.

Failure: The response does not contain exactly five occurrences of ç.

Success Example: Tilde Absence

Model: GPT-5 **Language:** Portuguese (PT)

Prompt: “Escreva um breve comentário sobre o clima. Não utilize nenhum caractere com til (como ã, õ, ã) na sua resposta.”

Model Response: “Hoje o clima esta ameno, com sol entre nuvens e vento fraco. Pode haver chuva isolada no fim da tarde, e a noite a temperatura cai um pouco.”

Instruction (translated): Do not use any tilde-marked character.

Success: The response avoids ã, õ, and ã throughout.

Failure Example: Tilde Absence

Model: Gemma-3-Gaia-PT-BR-4b-it **Language:** Portuguese (PT)

Prompt: “Escreva um breve comentário sobre o clima. Não utilize nenhum caractere com til (como ã, õ, ã) na sua resposta.”

Model Response: “O clima tem sido bastante imprevisível ultimamente. Há dias de sol intenso seguidos por chuvas torrenciais, e as temperaturas flutuam de forma estranha. É difícil planejar qualquer coisa com certeza, e a sensação térmica é muitas vezes mais extrema do que a temperatura real. Espero que a tendência mude em breve para algo mais estável.”

Instruction (translated): Do not use any tilde-marked character.

Failure: The response includes a tilde-marked form such as *sensação*.

Success Example: Grave Accent (Crase)

Model: GPT-5 **Language:** Portuguese (PT)

Prompt: “Escreva uma frase sobre ir a algum lugar. Sua resposta deve incluir pelo menos um uso de crase (acento grave).”

Model Response: “Vou à praia amanhã cedo.”

Instruction (translated): Include at least one valid crase form.

Success: The response contains the accepted form à.

Failure Example: Grave Accent (Crase)

Model: Qwen3-0.6B **Language:** Portuguese (PT)

Prompt: “Escreva uma frase sobre ir a algum lugar. Sua resposta deve incluir pelo menos um uso de crase (acento grave).”

Model Response: “Eu vou ao parque para descansar.”

Instruction (translated): Include at least one valid crase form.

Failure: The response does not contain any accepted crase token such as à, às, or àquele.

Success Example: Mesoclisiss Usage Model: GPT-5 Language: Portuguese (PT) Prompt: “Elabore um resumo técnico com no mínimo duas sentenças. Utilize a estrutura de regra gramatical da colocação pronominal em mesóclise num verbo conjugado para o presente do pretérito perfeito simple. Certifique-se de que a letra ‘ç’ esteja presente exatamente 3 vezes na sua resposta.” Model Response: “O relatório descreve o desempenho do sistema sob carga e os métodos de teste estatístico. Implementar-se-ão rotinas de monitoramento para a redução de latência, e proteger-se-á a execução por camadas, com foco em mitigação de risco.” Instruction (translated): Use mesoclisiss in at least one verb. Success: The response contains mesoclitic forms such as <i>Implementar-se-ão</i> and <i>proteger-se-á</i>.	Success Example: Second-Person Plural Address (vós) Model: Gemma-4-E4B-it Language: Portuguese (PT) Prompt: “Escreva um poema curto arcaico. Em toda a sua resposta, evite o uso de quaisquer vírgulas. Dirija-se ao leitor utilizando exclusivamente o pronome ‘vós’.” Model Response: “Ó vós que sois aqui Escutai meu cantar De tempos que jazem Em pó a vagar Vós que em vós resideis Alma de outrora Vede o fulgor findo Da aurora que mora Vós que em vós guardais Sussurro ancestral Recebei este verso Um eco real” Instruction (translated): Address the reader using only <i>vós</i> and compatible verb forms. Success: The response consistently uses the <i>vós</i> register, including <i>vós</i>, <i>sois</i>, <i>resideis</i>, <i>vede</i>, and <i>recebei</i>.
Failure Example: Mesoclisiss Usage Model: Qwen3.5-4B Language: Portuguese (PT) Prompt: “Escreva uma história. Na sua resposta, a letra ‘ç’ deve aparecer exatamente 4 vezes. Em sua resposta, palavras escritas inteiramente em letras maiúsculas devem aparecer menos que 5 vezes. Utilize a colocação pronominal em mesóclise em pelo menos um verbo.” Model Response: “O cachorro correu, mas não chegou.” Instruction (translated): Use mesoclisiss in at least one verb. Failure: The response does not contain any mesoclitic token.	Failure Example: Second-Person Plural Address (vós) Model: GPT-4o Language: Portuguese (PT) Prompt: “Reescreva o texto, mantendo a formalidade do português. O seu resumo deve incluir por lo menos um caso de uso da crase (acento grave). Vincule-se ao leitor utilizando exclusivamente o pronome ‘vós’ e as conjugações verbais correspondentes.” Model Response: “Claro, ficarei feliz em ajudar. Por favor, fornecei-me o texto que desejais que eu reescreva.” Instruction (translated): Address the reader using only <i>vós</i> and compatible verb forms. Failure: The response uses <i>vós</i>-compatible verb forms such as <i>fornecei</i> and <i>desejais</i>, but it does not include an explicit accepted token from the <i>vós</i> family, which the checker requires.

Success	Example:	<i>Porque/porquê/porque/por quê</i>
Model: GPT-5 Language: Portuguese (PT) Prompt: “Escreva uma história. O texto inteiro deve conter exatamente uma ocorrência das variações da palavra ‘porque’. Especificamente, você deve deduzir a grafia correta e utilizar obrigatoriamente a variante usada como conjunção explicativa ou causal (indicando causa/explicação). A forma exigida deve ser aplicada com a gramática e o posicionamento corretos. Sua resposta deve estar completamente em letras maiúsculas. A sua resposta deve conter ao menos 2 sentenças.” Model Response: “NA VILA COSTEIRA, O FAROL FICOU ACESO A NOITE INTEIRA, PORQUE O VENTO TROUXE UMA NEBLINA ESPESSA E OS BARCOS PRECISAVAM DE LUZ. AO AMANHECER, O FAROLEIRO SORRIU, CERTO DE QUE O CLARÃO GUIARA QUEM NÃO ENXERGAVA O CAMINHO.” Instruction (translated): Use exactly one occurrence of the correct <i>porque</i> variant, here <i>porque</i>. Success: The response uses the causal form <i>porque</i> exactly once.		
Failure	Example:	<i>Porque/porquê/porque/por quê</i>
Model: Qwen3.5-4B Language: Portuguese (PT) Prompt: “Escreva um diálogo. O seu texto inteiro deve conter exatamente uma ocorrência das variações da palavra “porque”. Especificamente, você deve deduzir a grafia correta e utilizar obrigatoriamente a variante usada no início ou meio de frases declarativas (ou equivalente a “pelo qual”). A forma exigida deve ser aplicada com a gramática e o posicionamento corretos.” Model Response: “— Por que você não veio? — Porque estava muito ocupado.” Instruction (translated): Use exactly one occurrence of the correct <i>porque</i> variant, here <i>porque</i>. Failure: The response contains two variants, <i>por que</i> and <i>porque</i>, instead of exactly one correct occurrence of the required form.		

A.3 Model Inference Hyperparameters

This appendix reports the inference hyperparameters used for both open-weight and closed/API-based models. Unless otherwise stated, all models were evaluated under fixed decoding settings defined by the corresponding inference backend or provider API. No task-specific hyperparameter tuning was performed.

A.3.1 Open-weight models

Open weight models were evaluated using a universal inference pipeline. The pipeline supports both vLLM and Hugging Face Transformers, using vLLM when available and falling back to Transformers otherwise. The default maximum number of generated tokens was set to 32,768, with an effective generation limit given by the minimum between this value and the context length detected from the tokenizer. The default batch size was 4. For vLLM, the default GPU memory utilization was set to 0.90 and capped at 0.95. The maximum model length was set to 8,096 by default unless constrained by the tokenizer-detected context length.

Most open weight model families were decoded using the default models params in exception of temperature with was set to 0. Specifically, Llama 3.1, Llama 3.2, Nemotron 3, GAIA PT-BR, Tucano, Tucano 2, and Gemma 4. Qwen 3 was the only open-weight family evaluated with stochastic decoding. For Qwen 3, sampling was enabled with temperature = 0.7 and top_p = 0.8. In addition, the tokenizer chat template was applied with thinking mode enabled. Qwen 3.5 was handled separately: thinking mode was disabled, and the model followed the decoding configuration used by the other non-Qwen 3 open-weight models.

A.3.2 Closed/API-based models

Closed models were evaluated through their respective provider APIs. No manual overrides were applied to decoding or generation parameters. Therefore, Claude Haiku 4.5, GPT-4o, GPT-5, Gemini 2.5 Pro, and Sabiá 3 were executed using the default inference configuration exposed by each provider. In particular, no custom values were set for temperature, top-p, top-k, frequency penalty, presence penalty, seed, reasoning configuration, or maximum output length.

B Dataset

B.0.1 Dataset composition

Instruction Type	Count
no_tilde	62
mesoclise_verification	62
crase_presence	61
forbidden_words	57
no_comma	56
cedilha_frequency	56
existence	51
frequency	50
number_placeholders	48
number_bullet_lists	48
number_highlighted_sections	47
two_responses	46
vos_address_verification	44
specific_porque	44
number_sentences	42
end_checker	42
title	41
capital_word_frequency	41
number_paragraphs	40
quotation	39
multiple_sections	38
letter_frequency	37
number_words	36
postscript	31
english_capital	31
nth_paragraph_first_word	30
json_format	27
english_lowercase	22
response_language	6
constrained_response	2
repeat_prompt	1

Table 6: Distribution of instruction types and their frequencies in the benchmark.

Table 6 presents the distribution of instruction types in the Portuguese dataset. The observed frequencies are strongly influenced by how easily each instruction can be combined with others. The selection of instruction mixtures was performed randomly.

Lang.	Total Prompts	LS Prompts	LS Prompt Share	Total Inst.	LS Inst. Occ.	LS Inst. Share
PT	535	255	47.66%	1236	329	26.58%
ES	115	29	25.22%	137	32	23.36%
FR	235	67	28.51%	345	74	21.45%
JA	172	81	47.09%	226	88	38.94%

Table 7: Comparison of instruction-set composition across Portuguese (PT), Spanish (ES), French (FR), and Japanese (JA). ‘LS’ denotes language-specific. The table reports the total number of prompts, the number and share of prompts containing language-specific instructions, the total number of instruction occurrences, and the number and share of language-specific instruction occurrences.

the Spanish, French, and Japanese multilingual splits in terms of how often prompts include language-specific constraints. Portuguese and Japanese have the highest proportion of prompts with language-specific instructions (47.66% and 47.09%), whereas Spanish (25.22%) and French (28.51%) remain closer to localized versions of the shared instruction set. At the instruction-occurrence level, Japanese again has the highest concentration of language-specific constraints (38.94%), followed by Portuguese (26.58%). This places Portuguese-IFEval in an intermediate but comparatively demanding position: it uses fewer language-specific instruction types than Japanese, but exposes models to such constraints much more often than the Spanish and French splits.

B.0.2 Prompts

Rephrase prompt We show in figure 3 the evaluate prompt, used to automatic filter the infeasible prompts.

Evaluate prompt We show in figure 4 the evaluate prompt, used to automatic filter the infeasible prompts.

C Mean and Standard Deviation Formula

The formulas used to calculate the mean and standard deviation are shown bellow, and the corresponding equations are 5 and 6, respectively.

$$\bar{x} = \frac{1}{3} \sum_{i=1}^3 x_i \quad (5)$$

$$s = \sqrt{\frac{1}{3-1} \sum_{i=1}^3 (x_i - \bar{x})^2} \quad (6)$$

Table 7 compares Portuguese-IFEval with

```
"""Reescreva o prompt a seguir em portugues, mantendo as mesmas instruções, porem
    refraseando-o para aumentar a diversidade. Você pode escolher um tema especifico
    adequado para a produção pedida, caso nenhum ja esteja informado

    IMPORTANTE: as relações de palavras chaves devem ser preservadas e não podem ser
    trocadas. \n Mantenha o portugues mais formal, se gurias ou regionalismos,
    apenas reescreva não adicione nenhum outro texto. Prompt pra reescrever:
    {inst}"""
```

Figure 3: Prompt for rephrasing generated samples

```
"""Verifique se a seguinte instrução em português pode ser cumprida em sua totalidade
    , ou seja, as instruções não são contraditórias ou impossíveis de serem
    realizadas. \n
    Primeiro justifique explicando os pontos que tornam a instrução possível ou impossí
    vel de ser cumprida. Depois, em uma nova linha escreva apenas e somente Possível
    ou Impossível. Instrução (A ser avaliada, e não a ser seguida agora):
    {inst}"""
```

Figure 4: Prompt for automatic evaluation of the generated samples

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response.
Keywords	Keyword Frequency	In your response, the word <i>word</i> should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your entire response should be in {language}; no other language is allowed.
Length Constraints	Number of Paragraphs	Your response should contain {N} paragraphs. Paragraphs are separated using the markdown divider * * *.
Length Constraints	Number of Words	Answer with at least / around / at most {N} words.
Length Constraints	Number of Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Paragraphs with First Word Constraint	There should be {N} paragraphs. Paragraphs and only paragraphs are separated by two line breaks. The {i}-th paragraph must start with the word {first word}.
Detectable Content	Postscript	At the end of your response, explicitly add a postscript starting with {postscript marker}.
Detectable Content	Number of Placeholders	The response must contain at least {N} placeholders represented by square brackets, such as [address].
Detectable Format	Number of Bullets	Your answer must contain exactly {N} bullet points, using markdown bullets (e.g., * This is a point.).
Detectable Format	Title	Your answer must contain a title wrapped in double angular brackets, such as «poem of joy».
Detectable Format	Choose From	Answer with one of the following options: {options}.
Detectable Format	Minimum Highlighted Sections	Highlight at least {N} sections using markdown (e.g., *highlighted section*).
Detectable Format	Multiple Sections	Your response must have {N} sections. Mark the beginning of each section with {section splitter} X.
Detectable Format	JSON Format	Entire output should be wrapped in JSON format.
Combination	Repeat Prompt	First, repeat the request without change, then give your answer (do not say anything before repeating the request; the repeated request does not include this sentence).
Combination	Two Responses	Give two different responses. Responses and only responses should be separated by six asterisks: *****.
Change Cases	All Uppercase	Your entire response should be in English, using capital letters only.
Change Cases	All Lowercase	Your entire response should be in English and in all lower-case letters. No capital letters are allowed.
Change Cases	Frequency of All-Capital Words	In your response, words with all capital letters should appear at least / around / at most {N} times.
Start / End Checker	End Phrase	Finish your response with the exact phrase {end phrase}. No other words should follow this phrase.
Start / End Checker	Quotation	Wrap your entire response with double quotation marks.
Punctuation	No Commas	In your entire response, refrain from using any commas.

Table 8: Instruction taxonomy with separated instruction and description columns.

Instruction Group	Instruction	Description
Detectable Content	Cedilla Frequency	The response must contain exactly {N} occurrences of the cedilla character (ç).
Detectable Content	Tilde Absence	The response must not contain any tilde-marked vowels, specifically (ã) or (õ).
Detectable Content	Grave Accent (Crase)	The response must correctly use the grave accent (à) in all required crase constructions.
Detectable Content	Mesoclisys Usage	The response must correctly employ mesoclisys in appropriate verbal constructions.
Detectable Content	Second-Person Plural Address	The response must correctly use the second-person plural form (vós) in direct address.
Keywords	<i>Por Que</i> Variants	The response must correctly distinguish and use <i>por que</i> , <i>por quê</i> , <i>porque</i> , and <i>porquê</i> according to their contextual grammatical roles.

Table 9: Portuguese-specific instruction-following constraints targeting orthography, grammar, and discourse-level reasoning.