

Portuguese-IFEval: Instruction-Following Evaluation in Portuguese

Anonymous ACL submission

Abstract

Instruction-following benchmarks have become central to evaluating large language models, yet their multilingual extensions remain largely translation-based. This design choice limits the ability of current evaluations to capture language-specific pragmatic, morphosyntactic, and orthographic constraints. We introduce **IFEval-PT**, a Portuguese extension of IFEval constructed through semantic regionalization rather than literal translation. The benchmark comprises 130 Portuguese prompts with one to three verifiable instructions, combining adapted and Portuguese-specific constraints. Evaluating proprietary, open-source, and Portuguese-tuned models under a unified protocol, we demonstrate systematic performance degradation on Portuguese-specific instructions relative to translated benchmarks. These results establish that translation alone is insufficient for faithful multilingual instruction-following evaluation and that language-aware benchmark design is essential to expose hidden failure modes. We will publicly release the benchmark, including all prompts and evaluation code, to support reproducibility and further research.

1 Introduction

Instruction following is a core capability of Large Language Models (LLMs) (Grattafiori et al., 2024; Yang et al., 2025; Taori et al., 2023; Xu et al., 2024), motivating the development of benchmarks that assess instruction compliance in zero-shot settings. Among these, Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) introduced deterministically verifiable constraints that avoid the bias, subjectivity, and cost associated with LLM-as-a-judge and manual evaluation protocols (Jiang et al., 2024; Zheng et al., 2023; Ouyang et al., 2022). However, the validity of such deterministic evaluation critically depends on how instructions are linguistically instantiated, particularly outside English.

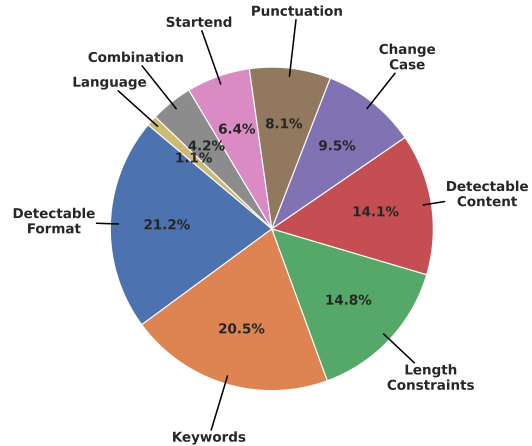


Figure 1: Strict instruction-following accuracy by category for Portuguese (PT).

IFEval is restricted to English, limiting its applicability in multilingual evaluation. M-IFEval (Dussolle et al., 2025) partially addresses this gap by extending the benchmark to Spanish, Japanese, and French, but largely relies on direct translation. This strategy risks obscuring language-dependent instruction-following behavior by enforcing superficially equivalent formulations across languages. As a result, existing benchmarks remain ill-suited to expose instruction-following failures in languages with richer orthographic, morphosyntactic, and pragmatic variation.

We introduce Portuguese-IFEval, a regionalized extension of IFEval designed to evaluate instruction following in Portuguese. Portuguese constitutes a challenging testbed due to its fine-grained orthographic conventions, morphosyntactic constraints, and discourse-level distinctions that are weakly represented in existing benchmarks. Rather than relying on direct translation, Portuguese-IFEval adopts a semantic regionalization strategy that preserves deterministic verifiability while enabling language-specific instruction instantiation.

Building on the original IFEval instruction set,

067 Portuguese-IFEval adapts its 25 verifiable instruc- 116
068 tions and introduces 6 additional instructions tar- 117
069 geting Portuguese-specific linguistic phenomena. 118
070 Figure 1 presents the resulting distribution of in- 119
071 struction categories. 120

072 We evaluate 19 state-of-the-art language 121
073 models, including proprietary systems, open- 122
074 source instruction-tuned models, and Portuguese- 123
075 specialized variants, under the standard IFEval eval- 124
076 uation protocol. This controlled setup enables di- 125
077 rect comparison across instruction types and bench- 126
078 mark construction strategies. Our results show 127
079 that, although larger and proprietary models gener- 128
080 ally achieve stronger instruction-following perfor- 129
081 mance, scale alone does not eliminate language- 130
082 specific failure modes, particularly under fine- 131
083 grained Portuguese constraints. 132

084 These findings motivate a broader reinterpretation 133
085 of multilingual instruction-following evaluation. 134
086 This work reframes instruction-following 135
087 evaluation across languages as a *linguistic ground-* 136
088 *ing problem*, rather than a translation prob- 137
089 lem: faithfully assessing instruction adherence re- 138
090 quires benchmarks that explicitly encode language- 139
091 specific orthographic, morphosyntactic, and prag- 140
092 matic constraints, instead of assuming that trans- 141
093 lated instructions preserve evaluative fidelity. 142

094 Our main contributions and findings are summa- 143
095 rized as follows: 144

- 096 • We introduce Portuguese-IFEval, a regional- 145
097 ized instruction-following benchmark that ex- 146
098 tends IFEval to Portuguese while preserving 147
099 deterministically verifiable constraints and ex- 148
100 plicitly incorporating language-specific phe- 149
101 nomena. 150
- 102 • Leveraging this benchmark, we conduct a rig- 151
103 orously controlled evaluation of state-of-the- 152
104 art proprietary and open-source LLMs, estab- 153
105 lishing a comprehensive performance land- 154
106 scape for instruction following in Portuguese. 155
- 107 • This evaluation enables a fine-grained com- 156
108 parison between the original instruction set 157
109 and Portuguese-specific instructions, reveal- 158
110 ing systematic language-dependent weak- 159
111 nesses that remain hidden under translated- 160
112 only benchmarks. 161
- 113 • Finally, by contrasting translated and region- 162
114 alized benchmark constructions under a uni- 163
115 fied evaluation protocol, we show that trans- 164
165

116 lation alone systematically masks language- 117
117 dependent failure modes, establishing region- 118
118 alized benchmark design as essential for faith- 119
119 ful instruction-following evaluation in Por- 120
120 tuguese. 121

2 Related Work 121

122 Instruction following has emerged as a central crite- 123
123 rion for evaluating the practical reliability of large 124
124 language models (LLMs), motivating the develop- 125
125 ment of benchmarks that assess instruction compli- 126
126 ance. A prominent line of work adopts an LLM- 127
127 as-a-judge paradigm, as exemplified by InFoBench 128
128 (Zeng et al., 2023) and FollowBench (Jiang et al., 129
129 2024). While flexible, this approach raises well- 130
130 documented concerns regarding bias, subjectivity, 131
131 and limited reproducibility (Zheng et al., 2023). 132

133 Instruction-Following Evaluation (IFEval) 134
134 (Zhou et al., 2023) introduced a deterministically 135
135 verifiable framework for instruction-following 136
136 evaluation, addressing these limitations by 137
137 replacing subjective judgments with rule-based 138
138 verification. This design enables objective, 139
139 scalable, and reproducible evaluation. However, 140
140 despite its methodological advantages, IFEval 141
141 remains primarily focused on English instructions, 142
142 limiting its applicability in multilingual settings. 143

144 To mitigate language bias, recent work has 145
145 explored multilingual extensions of instruction- 146
146 following benchmarks. M-IFEval (Dussolle et al., 147
147 2025) extends IFEval to Spanish, Japanese, and 148
148 French by adapting existing instructions and intro- 149
149 ducing language-specific scenarios. Marco-Bench- 150
150 MIF (Zeng et al., 2025) further broadens multi- 151
151 lingual coverage by extending IFEval to 30 lan- 152
152 guages, including Portuguese. However, this bench- 153
153 mark relies exclusively on translated versions of 154
154 the original instruction set and does not introduce 155
155 Portuguese-specific instruction design or valida- 156
156 tion. 157

158 As a result, no prior work has introduced 159
159 an instruction-following benchmark explicitly de- 160
160 signed for Portuguese that preserves deterministic 161
161 verification while incorporating language-specific 162
162 constraints. Existing multilingual benchmarks ei- 163
163 ther exclude Portuguese or treat it solely through 164
164 translation, which masks language-dependent 165
165 instruction-following behavior. In this work, we 166
166 introduce Portuguese-IFEval, extending the deter- 167
167 ministic evaluation paradigm of IFEval and M- 168
168 IFEval with semantically regionalized and newly 169
169

166	constructed Portuguese instructions, enabling systematic analysis of instruction-following capabilities in Portuguese.	214
167		215
168		216
169	3 Method	217
170	3.1 Overview	218
171	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 (Dussolle et al., 2025). The benchmark preserves the core notion of <i>verifiable instructions</i> , 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.	219
172		220
173		221
174		222
175		223
176		224
177		225
178		226
179		227
180		228
181		229
182		230
183		231
184		232
185	3.2 Instruction Set Regionalization	233
186	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.	234
187		235
188		236
189		237
190		238
191	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.	239
192		240
193		241
194		242
195		243
196		244
197		245
198		246
199	3.3 Portuguese-specific Instructions	247
200	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.	248
201		249
202		250
203		251
204		252
205		253
206		254
207		255
208		256
209		257
210		258
211		259
212		260
213		
	requires discourse-level reasoning by enforcing the canonical four-way distinction among <i>por que</i> , <i>porquê</i> , <i>porque</i> , and <i>porquê</i> within context.	
	Figure 2 illustrates one such instruction, which requires exactly five occurrences of the cedilla character (<i>ç</i>). 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 295 candidate prompts.	
	3.5 Prompt Rewriting	
	After instruction composition, prompts are rewritten by the model 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 the model evaluates whether all instructions in a prompt can be simultaneously satisfied.	
	Prompts with contradictory constraints are discarded. After this stage, 165 prompts remain.	
	3.7 Manual Verification	
	All remaining prompts are manually reviewed to remove residual unsatisfiable or ambiguous cases, following the curation procedures of IFEval and M-IFEval. After manual verification, the final benchmark consists of 153 prompts.	
	4 Experiments	
	4.1 Experimental Setup	
	We evaluate the benchmark using a diverse set of state-of-the-art language models, covering open-	

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, delicias 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 2: Portuguese instruction-following examples for the cedilha frequency constraint. The GPT-4o violates the exact occurrence requirement, while the GPT-5 satisfies it.

source instruction-tuned models, proprietary systems, and Portuguese-specialized variants.

For open-source instruction-tuned models, we consider the Qwen3 family (Yang et al., 2025) (0.6B to 8B), Gemma 3 (Team et al., 2025) (1B and 4B), LLaMA 3.1 (Grattafiori et al., 2024) (8B), LLaMA 3.2 (Grattafiori et al., 2024) (1B and 3B), and Phi-3.5 Mini (Abdin et al., 2024).

We additionally evaluate proprietary models, including GPT-4o and GPT-5 (OpenAI et al., 2024; OpenAI, 2025), Claude 3.5 Haiku (Anthropic, 2024), 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), 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:

$$\text{loose}(y, i) = \max_{t \geq 1} \text{followed}(t(y), i) \quad (2)$$

We report strict, loose, and averaged scores for all evaluated models on the Portuguese benchmark. In addition, we report language-level averages across all evaluated languages, averages over Portuguese-specific instructions, instruction-category breakdowns, and a direct comparison between regionalized and translated benchmark variants.

5 Results

Overall Performance on Portuguese Instructions. Across all evaluated models, instruction-following performance on Portuguese exhibits a

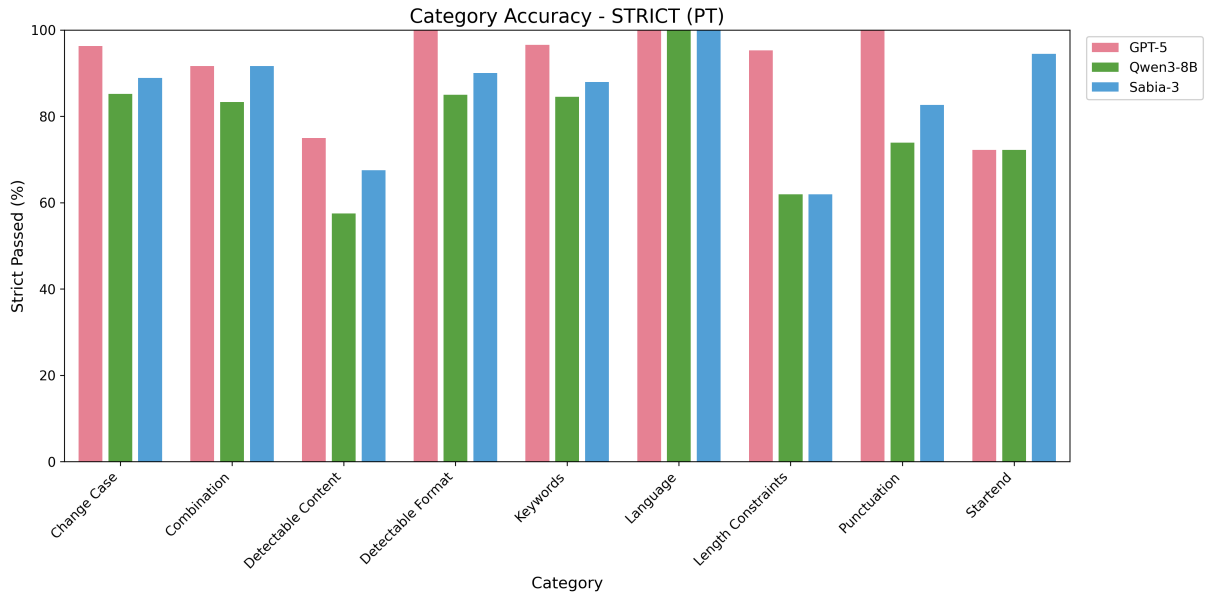


Figure 3: Strict instruction-following accuracy by category for Portuguese (PT).

clear stratification by model scale and training regime. Among open-source instruction-tuned models, the Qwen 3 family consistently leads performance, with Qwen 3 (8B) achieving the highest average score (66.99), followed by the 4B and 1.7B variants. This pattern establishes a strong scaling effect within the family. Gemma 3 (4B) remains competitive, while smaller models such as LLaMA 3.2 (1B) and Qwen 3 (0.6B) perform substantially worse. Proprietary API-based models dominate overall performance, led by GPT-5 with an average score of 86.60, followed by GPT-4o and Gemini 2.5 Pro. Among Portuguese-specialized models, Sabiá 3 achieves the strongest results (73.20), substantially outperforming the Tucano variants and approaching the performance of high-capacity API models, while Gaia (4B) shows moderate but consistent accuracy. Detailed results are reported in Table 1.

Cross-lingual Robustness Across Instruction Sets. Having established overall performance in Portuguese, we now examine how these 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, Qwen 3 (8B) achieves the highest overall multilingual average (46.92), with relatively balanced performance across all four languages. Phi 3.5 Mini performs competitively in Spanish and Japanese but degrades noticeably in Portuguese. Smaller models consistently underper-

form 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 language and an overall average of 90.55, followed by Gemini 2.5 Pro and Claude 3.5 Haiku. Among Portuguese-specific models, Sabiá 3 exhibits the strongest cross-lingual robustness, while Tucano models show limited transfer beyond Portuguese. Full multilingual results are shown in Table 2.

Impact of Portuguese-specific Instructions. While cross-lingual evaluation highlights generalization capacity, it does not isolate language-specific constraints. To this end, we directly compare performance on the standard instruction set with Portuguese-specific instructions. This comparison reveals a systematic degradation in accuracy. Models with strong overall performance under standard IFEval, including Qwen 3 (8B), Gemini 2.5 Pro, 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, indicating that general instruction-following competence does not translate into mastery of fine-grained, language-dependent constraints. Even high-performing API models show measurable reductions, although GPT-5 remains comparatively robust, suggesting stronger internalization of Portuguese-specific rules. These results demonstrate that instruction specialization exposes weaknesses masked by generic benchmarks.

Model	Lang	Strict	Loose	Avg
<i>Instruct Models</i>				
Llama 3.1 (8B)	PT	49.67	55.56	52.61
Llama 3.2 (3B)	PT	35.95	39.22	37.59
Llama 3.2 (1B)	PT	21.57	24.84	23.21
Gemma 3 (4B)	PT	54.90	54.90	54.90
Gemma 3 (1B)	PT	28.76	30.72	29.74
Phi 3.5 (Mini)	PT	43.79	49.02	46.41
Qwen 3 (8B)	PT	64.71	69.28	66.99
Qwen 3 (4B)	PT	62.75	65.36	64.06
Qwen 3 (1.7B)	PT	54.25	58.17	56.21
Qwen 3 (0.6B)	PT	45.10	47.06	46.08
<i>API Models</i>				
Claude 3.5 Haiku	PT	66.67	74.51	70.59
GPT-4o	PT	73.2	79.8	76.5
GPT-5	PT	86.27	86.93	86.60
Gemini Pro	PT	79.08	83.01	81.05
<i>PT-Specific Models</i>				
Sabiá-3	PT	70.59	75.82	73.2
Tucano (1B)	PT	9.15	10.46	9.81
Tucano (2B)	PT	7.19	8.50	7.85
Gaia (4B)	PT	52.94	57.52	55.23

Table 1: IFEval results for Portuguese (PT). Models not evaluated in PT are shown in red.

Portuguese as a Challenging Evaluation Setting. To contextualize these drops, we compare Portuguese with other language-specific instruction sets. Across most model families, Portuguese emerges as one of the most challenging languages. Accuracy in Portuguese is consistently lower or more volatile than in Spanish and French, particularly for orthography- and diacritic-sensitive instructions involving phenomena such as crase, mesoclis, and the canonical “porquês” distinction. Although Japanese also presents substantial difficulty due to its writing system, high-capacity API models maintain more stable performance in Japanese than in Portuguese. This contrast indicates that Portuguese morphosyntactic and pragmatic constraints remain less robustly encoded in current instruction-tuning pipelines. Comparative results are summarized in Table 3.

Category-level Failure Patterns in Portuguese. We conclude by analyzing where these failures concentrate at the category level. GPT-5 consistently achieves the highest accuracy across nearly all instruction categories, reaching near-perfect performance on structural and surface-form constraints such as Detectable Format, Language, and Punc-

Model	ES	FR	JA	PT	Avg
<i>Instruct Models</i>					
Llama 3.1 (8B)	52.5	53.05	23.81	29.16	39.63
Llama 3.2 (3B)	52.50	53.73	22.82	25.00	38.52
Llama 3.2 (1B)	47.50	49.69	17.86	16.67	32.93
Gemma 3 (1B)	52.50	51.49	14.29	29.17	36.86
Gemma 3 (4B)	52.50	57.64	22.62	41.67	43.61
Phi 3.5 Mini	60.00	43.03	37.30	16.67	43.55
Qwen 3 (8B)	57.50	60.23	36.61	33.33	46.92
Qwen 3 (4B)	55.00	50.83	35.62	33.33	43.69
Qwen 3 (1.7B)	45.00	54.09	25.20	25.00	37.32
Qwen 3 (0.6B)	40.00	42.24	32.14	16.67	32.76
<i>API Models</i>					
Claude 3.5 Haiku	62.50	84.62	59.62	54.17	65.23
GPT-4o	55.00	78.04	58.23	62.50	63.53
GPT-5	100.00	89.50	89.38	83.33	90.55
Gemini Pro	70.00	91.41	65.28	66.67	73.34
<i>PT-Specific Models</i>					
Sabiá 3	57.50	57.72	34.52	50.00	49.93
Tucano (1B)	7.50	34.94	30.85	16.67	22.49
Tucano (2B)	7.50	33.60	23.51	12.50	19.78
Gaia (4B)	52.50	56.21	23.81	45.83	44.59

Table 2: Strict scores on instruction level of language-specific instructions across languages. Portuguese, Spanish, French, and Japanese results are reported as well as average result between the languages;

tuation, while showing slightly lower performance on Detectable Content and Start/End constraints. Qwen 3 (8B) exhibits solid but uneven behavior, matching GPT-5 on Language constraints but dropping sharply on content-sensitive categories, particularly Detectable Content and Length Constraints. Sabiá 3 typically falls between GPT-5 and Qwen 3, outperforming Qwen 3 on Combination, Punctuation, and Start/End categories, but remaining below GPT-5 overall. 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 3.

Impact of Portuguese-specific Instructions. While cross-lingual evaluation highlights generalization capacity, it does not isolate language-specific constraints. To this end, we directly compare performance on the standard instruction set with Portuguese-specific instructions. This comparison reveals a systematic degradation in accuracy, which we refer to as *language-specific constraint collapse*: models that perform strongly under generic instruction-following evaluation fail to con-

Model	EN	ES	FR	JA	PT	Mean
<i>API Models</i>						
Claude 3.5 Haiku	82.07	91.30	83.83	65.70	66.67	77.91
GPT-4o	84.10	87.83	82.98	63.37	73.20	78.30
GPT-5	93.72	98.26	90.64	88.37	86.27	91.45
Gemini Pro	90.76	89.57	88.09	66.28	79.08	82.75
<i>Instruct Models</i>						
Llama 3.1 (8B)	72.64	74.78	65.11	34.88	54.61	60.41
Llama 3.2 (3B)	51.94	56.52	43.40	18.60	35.95	41.28
Llama 3.2 (1B)	43.25	43.48	28.94	11.63	21.57	29.77
Gemma 3 (4B)	60.44	72.17	59.15	38.95	54.90	57.12
Gemma 3 (1B)	34.94	48.70	34.47	18.02	28.10	32.84
Phi 3.5 Mini	50.09	66.09	57.45	40.70	33.29	49.52
Qwen 3 (8B)	78.19	75.65	68.09	52.33	64.71	67.79
Qwen 3 (4B)	78.93	79.13	69.36	51.16	62.75	68.27
Qwen 3 (1.7B)	66.91	71.30	57.02	38.37	54.25	57.57
Qwen 3 (0.6B)	59.15	70.43	45.11	35.47	45.10	51.05
<i>PT-Specific Models</i>						
Sabiá 3	80.22	85.22	77.45	47.67	70.59	72.23
Tucano (2B)	13.86	22.61	16.60	16.86	10.23	16.03
Tucano (1B)	11.09	20.87	17.45	17.44	10.09	15.39
Gaia (4B)	64.14	67.83	57.87	41.28	47.41	55.70

Table 3: Instruction-following strict score, prompt level, across all languages.

sistently satisfy fine-grained, language-dependent constraints. Models with strong overall performance under standard IFEval, including Qwen 3 (8B), Gemini 2.5 Pro, 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, indicating that general instruction-following competence does not translate into mastery of fine-grained, language-dependent constraints. Even high-performing API models show measurable reductions, although GPT-5 remains comparatively robust, suggesting stronger internalization of Portuguese-specific rules. These results demonstrate that instruction specialization exposes weaknesses masked by generic benchmarks.

6 Conclusion

This work provides a systematic analysis of instruction-following performance in Portuguese, revealing persistent gaps obscured by English-centric evaluation practices. Across all evaluated models, strict instruction-specific evaluation exposes substantial performance degradation in Portuguese, indicating that instruction following remains fragile under fine-grained linguistic constraints. While performance correlates with model scale and training regime, language-specific constraints consistently amplify alignment failures,

particularly for small and mid-sized models. Even large proprietary models exhibit measurable drops under Portuguese-specific instructions, demonstrating that scale and broad pre-training alone are insufficient for robust multilingual instruction following. These findings have direct implications for benchmark design. Translation-based evaluation protocols are fundamentally inadequate, as they mask language-dependent failure modes. We argue that instruction-following evaluation must adopt language-aware, regionalized benchmark construction that explicitly targets orthographic, morphosyntactic, and pragmatic phenomena. By showing that regionalized benchmarks reveal failures invisible to translated evaluations, this work reframes multilingual instruction-following as a benchmark design problem rather than a scaling problem, and highlights linguistic diversity as a first-class requirement for reliable and equitable evaluation.

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.

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.

Ethical considerations

In developing the Portuguese-IFEval, we carefully considered ethical implications at every stage, from dataset construction to model evaluation. All prompts and instructions were designed with attention to cultural sensitivity and linguistic appro-

499	priateness, avoiding harmful stereotypes, biased		
500	content, or misrepresentations of Portuguese lan-		
501	guage and culture. Human contributors involved in		
502	annotation, or verification were compensated fairly,		
503	provided informed consent, and had their privacy		
504	and rights fully respected.		
505	The dataset and evaluation protocols are docu-		
506	mented to ensure transparency and reproducibility,		
507	with guidelines aimed at preventing misuse. We		
508	acknowledge the environmental impact of training		
509	and evaluating large language models and encour-		
510	age the use of energy-efficient practices in NLP		
511	research. Overall, our work aims to support respon-		
512	sible, inclusive, and culturally aware evaluation of		
513	instruction-following in Portuguese LLMs, promot-		
514	ing fairness and respect for human participants.		
515	Acknowledgments		
516	need to write		
517	References		
518	Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed		
519	Awadallah, Ammar Ahmad Awan, Nguyen Bach,		
520	Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat		
521	Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck,		
522	Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav		
523	Chaudhary, Dong Chen, Dongdong Chen, and 110		
524	others. 2024. Phi-3 technical report: A highly capa-		
525	ble language model locally on your phone . <i>Preprint</i> ,		
526	arXiv:2404.14219.		
527	Hugo Abonizio, Thales Sales Almeida, Thiago Laitz,		
528	Roseval Malaquias Junior, Giovana Kerche Bonás,		
529	Rodrigo Nogueira, and Ramon Pires. 2025. Sabia-3		
530	technical report . <i>Preprint</i> , arXiv:2410.12049.		
531	Sonnet Anthropic. 2024. Model card ad-		
532	dendum: Claude 3.5 haiku and up-		
533	graded claude 3.5 sonnet. URL		
534	https://assets.anthropic.com/m/1cd9d098ac3e6467/original/		
535	Claude-3-Model-Card-October-Addendum.pdf , 273639283.		
536	S. S. T.; PEREIRA L. A.; AMADEUS M.; SCOTTI		
537	R.; FAZZIONI D.; NOVAIS A. M. A.; JORDÃO		
538	S. A. A. CAMILO-JUNIOR, C. G.; OLIVEIRA.		
539	2025. Gaia: An open language model for		
540	brazilian portuguese. [https://huggingface.		
541	co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it]		
542	(https://huggingface.co/CEIA-UFG/		
543	Gemma-3-Gaia-PT-BR-4b-it).		
544	Gheorghe Comanici, Eric Bieber, Mike Schaeckermann,		
545	Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-		
546	cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke		
547	Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni,		
548	Nathan Lintz, Tiago Cardal Pais, Henrik Jacobs-		
549	son, Idan Szpektor, Nan-Jiang Jiang, and 3416 oth-		
550	ers. 2025. Gemini 2.5: Pushing the frontier with		
	advanced reasoning, multimodality, long context,		551
	and next generation agentic capabilities . <i>Preprint</i> ,		552
	arXiv:2507.06261.		553
	Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and		554
	Shiza Fatimah. 2025. Tucano: Advancing neural text		555
	generation for portuguese . <i>Patterns</i> , 6(11):101325.		556
	Antoine Dussolle, A Cardeña, Shota Sato, and Peter		557
	Devine. 2025. M-ifeval: Multilingual instruction-		558
	following evaluation. In <i>Findings of the Association</i>		559
	<i>for Computational Linguistics: NAACL 2025</i> , pages		560
	6161–6176.		561
	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,		562
	Abhinav Pandey, Abhishek Kadian, Ahmad Al-		563
	Dahle, Aiesha Letman, Akhil Mathur, Alan Schel-		564
	ten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh		565
	Goyal, Anthony Hartshorn, Aobo Yang, Archi Mi-		566
	tra, Archie Sravankumar, Artem Korenev, Arthur		567
	Hinsvark, and 542 others. 2024. The llama 3 herd of		568
	models . <i>Preprint</i> , arXiv:2407.21783.		569
	Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun		570
	Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin		571
	Jiang, Qun Liu, and Wei Wang. 2024. Follow-		572
	Bench: A multi-level fine-grained constraints fol-		573
	lowing benchmark for large language models . In		574
	<i>Proceedings of the 62nd Annual Meeting of the As-</i>		575
	<i>sociation for Computational Linguistics (Volume 1:</i>		576
	<i>Long Papers)</i> , pages 4667–4688, Bangkok, Thailand.		577
	Association for Computational Linguistics.		578
	OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher,		579
	Adam Perelman, Aditya Ramesh, Aidan Clark,		580
	AJ Ostrow, Akila Welihinda, Alan Hayes, Alec		581
	Radford, Aleksander Mądry, Alex Baker-Whitcomb,		582
	Alex Beutel, Alex Borzunov, Alex Carney, Alex		583
	Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o		584
	system card . <i>Preprint</i> , arXiv:2410.21276.		585
	OpenAI. 2025. Gpt-5 system card. https://cdn.		586
	openai.com/gpt-5-system-card.pdf . Accessed:		587
	2026-01-02.		588
	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Car-		589
	roll L. Wainwright, Pamela Mishkin, Chong Zhang,		590
	Saahilni Agarwal, Katarina Slama, Alex Ray, John		591
	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,		592
	Maddie Simens, Amanda Askell, Peter Welinder,		593
	Paul Christiano, Jan Leike, and Ryan Lowe. 2022.		594
	Training language models to follow instructions with		595
	human feedback. In <i>Proceedings of the 36th Interna-</i>		596
	<i>tional Conference on Neural Information Processing</i>		597
	<i>Systems, NIPS '22</i> , Red Hook, NY, USA. Curran		598
	Associates Inc.		599
	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann		600
	Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,		601
	and Tatsunori B. Hashimoto. 2023. Stanford alpaca:		602
	An instruction-following llama model. https://		603
	github.com/tatsu-lab/stanford_alpaca .		604
	Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya		605
	Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,		606
	Tatiana Matejovicova, Alexandre Ramé, Morgane		607

608 Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey
609 Cideron, Jean bastien Grill, Sabela Ramos, Edouard
610 Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,
611 and 197 others. 2025. [Gemma 3 technical report](#).
612 *Preprint*, arXiv:2503.19786.

613 Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng,
614 Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei
615 Lin, and Daxin Jiang. 2024. [WizardLM: Empow-](#)
616 [ering large pre-trained language models to follow](#)
617 [complex instructions](#). In *The Twelfth International*
618 *Conference on Learning Representations*.

619 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,
620 Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,
621 Chengen Huang, Chenxu Lv, Chujie Zheng, Day-
622 iheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao
623 Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41
624 others. 2025. [Qwen3 technical report](#). *Preprint*,
625 arXiv:2505.09388.

626 Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng,
627 Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao,
628 Yefeng Liu, Chenyu Zhu, and 1 others. 2025. Marco-
629 bench-mif: On multilingual instruction-following ca-
630 pability of large language. In *Proceedings of the*
631 *63rd Annual Meeting of the Association for Compu-*
632 *tational Linguistics (Volume 1: Long Papers)*, pages
633 24058–24072.

634 Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya
635 Goyal, and Danqi Chen. 2023. Evaluating large
636 language models at evaluating instruction following.
637 *arXiv preprint arXiv:2310.07641*.

638 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan
639 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,
640 Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,
641 Joseph E Gonzalez, and Ion Stoica. 2023. [Judging](#)
642 [llm-as-a-judge with mt-bench and chatbot arena](#). In
643 *Advances in Neural Information Processing Systems*,
644 volume 36, pages 46595–46623. Curran Associates,
645 Inc.

646 Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Sid-
647 dhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou,
648 and Le Hou. 2023. Instruction-following evalu-
649 ation for large language models. *arXiv preprint*
650 *arXiv:2311.07911*.

651 A Appendix

652 A.1 Instructions

653 A.1.1 Shared with English Instructions

654 The Table 4 shows all the instructions that are
655 shared between English and Portuguese.

656 A.1.2 Portuguese-specific Instructions

657 The Table 5 shows all the Portuguese-specific in-
658 structions that are create on our work.

B Dataset

659

B.0.1 Dataset composition

660

661 Table 6 presents the distribution of instruction types
662 in the Portuguese dataset. The observed frequen-
663 cies are strongly influenced by how easily each
664 instruction can be combined with others. The se-
665 lection of instruction mixtures was performed ran-
666 domly.

B.0.2 Prompts

667

668 Rephrase prompt We show in figure 4 the evalu-
669 ate prompt, used to automatic filter the infeasible
670 prompts.

671 Evaluate prompt We show in figure 5 the evalu-
672 ate prompt, used to automatic filter the infeasible
673 prompts.

```
prompt = f"""
Reescreva o prompt a seguir em português, mantendo as mesmas instruções, porém
refraseando-o para aumentar a diversidade. Você pode escolher um tema
específico adequado para a produção pedida, caso nenhum já esteja informado

IMPORTANTE: as relações de palavras-chave devem ser preservadas e não podem ser
troçadas. \n Mantenha o português mais formal, se gírias ou regionalismos,
apenas reescreva não adicione nenhum outro texto. Prompt para reescrever:
{inst}
"""
```

Figure 4: Prompt for automatic evaluation of the generated samples

```
prompt = f"""
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
impossível 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 5: 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 lowercase 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 4: 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 5: Portuguese-specific instruction-following constraints targeting orthography, grammar, and discourse-level reasoning.

Instruction Type	Count
number_placeholders	17
capital_word_frequency	16
forbidden_words	15
no_comma	15
existence	14
multiple_sections	14
number_paragraphs	14
letter_frequency	13
number_bullet_lists	13
title	12
number_words	12
frequency	12
end_checker	11
two_responses	11
postscript	11
json_format	10
number_highlighted_sections	9
nth_paragraph_first_word	8
number_sentences	8
quotation	7
portuguese_lowercase	6
portuguese_capital	5
cedilha_frequency	4
vos_address_verification	4
mesoclise_verification	4
crase_presence	4
four_porques_verification	4
no_tilde	4
response_language	3
constrained_response	2
repeat_prompt	1

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