Benchmark Data Contamination in Underrepresented Languages: A Comprehensive Analysis Using Brazilian Data

Iriedson Vilar[1], David Maia[2], João Brunet[1], Fábio Morais[1], Leandro Balby[1]
[1]Universidade Federal de Campina Grande (UFCG)
Campina Grande - PB - Brazil
iriedson@copin.ufcg.edu.br, {joao.arthur, fabio, lbmarinho}@computacao.ufcg.edu.br
[2] Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
Campina Grande - PB - Brazil
david.maia@ifpb.edu.br

Abstract

Large Language Models (LLMs) are typically evaluated using standardized benchmarks to enable consistent performance measurement and model comparison. However, the reliability of these benchmarks can be undermined by data contamination, which occurs when evaluation items are inadvertently included in training corpora. While this issue has been investigated primarily in high-resource languages such as English and Chinese, its impact on underrepresented languages — such as Brazilian Portuguese — remains understudied. In this paper, we present one of the first systematic investigations of benchmark data contamination (BDC) in an underrepresented language setting, using Brazilian Portuguese as a case study. Using validated methodologies from the literature, we evaluate specialized and multilingual models across four benchmarks: BLUEX, ENEM Challenge, OAB Exams, and HealthQA-BR. Our approach applyes TS-Guessing to detect contamination via memorized knowledge, alongside a 50-character n-gram similarity strategy to identify benchmark items leaked into training data. Our results provide consistent evidence of contamination, revealing that models with stronger memorization and retrieval abilities tend to achieve artificially inflated benchmark scores. Our contributions include: (i) classifying models according to their contamination risk, (ii) identifying the benchmarks most affected by data leakage, and (iii) reporting contaminated training corpora.

Keywords: Large Language Models, Benchmark Data Contamination, Underrepresented Languages, Brazilian Portuguese, Model Evaluation

1. Introduction

Evaluating Large Language Models (LLMs) is essential for advancing Natural Language Processing (NLP) by revealing their capabilities and limitations (Laskar et al., 2024). Benchmarks play a key role by providing standardized test sets with defined data, metrics, and protocols (Chang et al., 2024). They enable systematic performance measurement across tasks like text understanding, question answering, and machine translation (Chang et al., 2024; Laskar et al., 2024). Widely used benchmarks include MMLU (which evaluates general knowledge) (Hendrycks et al., 2021), GSM8K (focused on mathematical reasoning) (Cobbe et al., 2021), and HumanEval (designed for code generation) (Chen et al., 2021).

Despite their utility, benchmarks face the challenge of **benchmark data contamination (BDC)** — benchmark items found in training data (Xu et al., 2024a). Reported cases reveal memorization: Qwen-1.8B reproduced GSM8K n-grams (Xu et al., 2024b), GPT-4 reconstructed masked MMLU answers with 57% accuracy (Deng et al., 2024b), and StarCoderBase-15.5B overlapped with MBPP and HumanEval (Riddell et al., 2024). Such contamination can inflate performance and mask model weaknesses, undermining benchmark reliability.

For underrepresented languages such as Brazilian Portuguese (PT-BR), contamination poses an additional challenge. The scarcity of high-quality corpora (Kreutzer et al., 2022; Thompson et al., 2024) encourages reliance on public data sources — such as repositories and exam materials — that often recur across training and evaluation datasets (Xu et al., 2024a; Deng et al., 2024b; Junior et al., 2024). Although awareness of this issue has grown within the Brazilian Portuguese NLP community (Pires et al., 2023; Abonizio et al., 2024; Junior et al., 2024), our literature review found no prior work systematically investigating benchmark leakage in this context. This demonstrates a clear gap in the understanding of model contamination in an underrepresented yet widely spoken language (PT-BR).

Drawing on a robust methodology validated in prior work, we assess contamination in Brazilian Portuguese across two dimensions: models and corpora. For models, we analyze specialized PT-BR models (e.g., *Sabia-3.1*) and multilingual ones (e.g., *GPT-4o*), using the TS-Guessing method (Deng et al., 2024b), which masks one incorrect multiple-choice option for model recovery. For corpora, we apply 50-character substring matching (OpenAI, 2023) to detect overlaps with benchmarks. Our study evaluates 15 models and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4765–4777
11–16 May 2026.

3 corpora using four widely adopted PT-BR benchmarks: *BLUEX* (Almeida et al., 2023), *ENEM Challenge* (Silveira and Mauá, 2018), *HealthQA-BR* (D'addario, 2025), and *OAB Exams* (Delfino et al., 2017).

Our analysis reveals clear evidence of contamination across both analytical dimensions: model behavior and pretraining data. At the behavioral level, contamination signals were detected in 14 model–benchmark pairs, where Exact Match scores exceeds the statistical threshold (EM ≥ 0.02) derived from the empirical distribution of results. Multilingual systems showed stronger memorization and higher accuracy, while small specialized Portuguese models displayed minimal behavioral contamination despite exposure to leaked data. Jointly, these results reveal a consistent positive association between model performance and contamination — with correlations ranging from moderate to strong across benchmarks ($\rho_{\text{EM}} = 0.52$–0.79, all $p < 0.001$). This indicates that contamination systematically inflates model performance, reinforcing the importance of evaluating accuracy and contamination together. At the corpus level, leakage exceeded 90% in some pretraining datasets, confirming that benchmark material is deeply embedded in Portuguese training data.

We organize this paper as follows. Section 2 introduces the background and related work. Section 3 details our experimental design, variables, and evaluation setup. Section 4 presents and interprets our findings. Section 5 provides replication package details. Section 6 summarizes our contributions and future directions. Finally, Section 7 discusses validity threats.

2. Background and Related Work

This section introduces the conceptual foundations of BDC, detection frameworks, and the context of Brazilian Portuguese LLMs.

2.1. Benchmark Data Contamination: Definition and Forms

BDC occurs when evaluation data overlaps with a model's pretraining corpus, artificially inflating performance by rewarding memorization over reasoning (Sainz et al., 2023; Deng et al., 2024a; Dong et al., 2024). As datasets grow larger and more opaque, identifying these overlaps becomes increasingly difficult (Ravaut et al., 2024). Contamination can be deliberate (when benchmarks are used for fine-tuning or validation) or accidental (from large-scale web crawls that capture public benchmarks (Xu et al., 2024a; OpenAI, 2023)). While automated filters and exclusion lists mitigate this risk, complete removal remains infeasible.

Contamination is further defined by its form and exposure level. Forms range from *guideline contamination* (exposure to task instructions) to *annotation contamination* (presence of inputs and labels) (Sainz et al., 2023). Exposure spans from *semantic similarity*, where models learn from related material, to *literal duplication*, where test items appear verbatim. Ravaut et al. (2024) also distinguishes between *input leakage* (questions only) and *input–label leakage* (questions and answers).

2.2. Detection Frameworks and Methodological Approaches

Contamination detection depends on model transparency. *White-box* models (e.g., *Tucano-2B4*) expose parameters and training data, allowing direct detection via string matching (OpenAI, 2023) or semantic similarity (Deng et al., 2024a; Sainz et al., 2023). *Gray-box* models (e.g., *LLaMA*) provide intermediate signals such as logits, supporting indirect inference (Deng et al., 2024a). *Black-box* models (e.g., *ChatGPT*) restrict outputs, making behavioral inference the only viable option.

White-box detection primarily relies on textual overlap. Early studies used 8-gram matches (Radford et al., 2019), later refined to token-level n-grams in *LLaMA 2* (Touvron et al., 2023). The GPT-4 technical report introduced the 50-character substring method (OpenAI, 2023), which we adopt for measuring overlap between benchmarks and large-scale Portuguese corpora. This approach influenced subsequent efforts, such as *ClassiCC-PT* (Almeida et al., 2025).

Among black-box methods, *TS-Guessing* is a robust and reproducible protocol for detecting behavioral contamination. Proposed by Deng et al. (2024b), it tests whether models can reconstruct masked incorrect options in multiple-choice tasks. It has been applied across domains, including the verification of medical benchmark integrity (Chen et al., 2025) and benchmark screening (Sun et al., 2025). Although it can overestimate contamination in multimodal tasks (Song et al., 2025), this limitation does not affect our textual benchmarks. Its efficiency and consistency make it well suited for our analysis.

To execute this study, we adapted two open-source LLM evaluation frameworks. For corpus inspection, we adapted *LLMSanitize*[1], developed by the NTU NLP Group, which organizes multiple contamination detection methods. Our fork extended its 50-character substring procedure to support streaming execution and large-scale corpus comparison. For behavioral analysis, we adapted Stanford's established *Holistic Evaluation of Lan-*

[1] https://github.com/ntunlp/LLMSanitize

guage Models (HELM)[2] (Liang et al., 2023), integrating the TS-Guessing protocol into its pipeline.

2.3. The Brazilian Portuguese Scenario

Most LLMs are trained primarily in English, reflecting its abundance and dominance in NLP (Yuan et al., 2024; Qin et al., 2025). English represents 92.65% of *GPT-3*'s tokens and 89.70% in *LLaMA 2* (Li et al., 2025; Yuan et al., 2024), creating a structural bias that limits the representation of underrepresented languages such as Brazilian Portuguese (Corrêa et al., 2024; Larcher et al., 2023).

Specialized models and curated corpora for Brazilian Portuguese are being developed. Models include *Sabiá-3* (Abonizio et al., 2024) (trained from scratch) and *Tucano-2B4* (Corrêa et al., 2025) (built on the 200-billion-token *GigaVerbo* corpus), while *openCabrita-3B* (Larcher et al., 2023) and *Bode-13B* (Garcia et al., 2024) adapt multilingual architectures through fine-tuning. These efforts are complemented by corpora, like *Aroeira* (Lira et al., 2025) and *ClassiCC-PT* (Almeida et al., 2025), which apply stricter filtering to improve Brazilian data quality.

While Almeida et al. (2025) has shown concerns for BDC, we did not find any studies that systematically analyze this phenomenon in Brazilian Portuguese. Our research is thus the first to comprehensively investigate both corpus-level and model-level contamination for PT-BR.

3. Methodology

Our methodology is guided by the following research questions:

- **RQ1:** Do large language models exhibit behavioral evidence of contamination by Brazilian Portuguese benchmark data?
- **RQ2:** Do Brazilian Portuguese benchmarks appear within the training corpora used to build Portuguese or multilingual LLMs?

These questions examine BDC from complementary perspectives. **RQ1** investigates behavioral evidence of memorization in LLMs on Brazilian Portuguese benchmarks, offering a model-centric view. **RQ2** examines the presence of benchmark data within Portuguese pretraining corpora, providing a data-centric perspective on source leakage.

3.1. Method to Answer RQ1

Three factors shape our experimental design: the contamination detection method, the benchmarks, and the evaluated models.

[2] https://crfm.stanford.edu/helm/

3.1.1. Factors

TS-Guessing — The contamination detection method. We prioritized four criteria for choosing a BDC detection technique: black-box applicability (for open and closed models), peer-reviewed validation, public implementation details, and reuse beyond the original study. We selected the *Test Set Slot Guessing (TS-Guessing)* protocol (Deng et al., 2024b), which meets all criteria, supports multiple-choice benchmarks, and has low computational overhead.

Although Section 2 introduced TS-Guessing conceptually, here we detail its implementation. TS-Guessing is a behavioral protocol detecting benchmark leakage by testing the model's ability to recall specific, non-salient information. We used the "Question-Multichoice" option, where in each instance one incorrect answer choice is randomly replaced with a [MASK] token, while the correct and remaining options remain visible. The model is then prompted to reconstruct the masked text.

Metrics. Instance-level results are aggregated across each benchmark to compute two contamination indicators: the **EM rate** (proportion of exact reproductions) and the **mean ROUGE-L F1 score** (average lexical overlap). As their Spearman correlation was high ($\rho = 0.820$), we rely on EM as the primary contamination behavioral signal. While any non-zero EM suggests contamination (chance reproduction is highly unlikely), we adopted a conservative, data-driven threshold. An ordered EM distribution across all 60 model–benchmark pairs showed a clear inflection near 0.02, which we define as the contamination threshold.

Benchmarks. We evaluated four multiple-choice benchmarks across distinct domains: *BLUEX* (Almeida et al., 2023), *ENEM Challenge* (Silveira and Mauá, 2018), *HealthQA-BR* (D'addario, 2025), and *OAB Exams* (Delfino et al., 2017). All datasets use official *HELM Stanford* splits, ensuring consistency and reproducibility. These benchmarks are widely used in Portuguese LLM evaluations, as noted Junior et al. (2024), Corrêa et al. (2025), and the *Open Portuguese LLM Leaderboard*[3]. Characteristics are summarized in Table 1.

Benchmark	Domain	Split Used	# Instances
BLUEX*	General Knowledge	Questions	724
ENEM Challenge	General Knowledge	Train	1432
HealthQA-BR	Health & Medicine	Train	5632
OAB Exams	Law	Train	2210

Table 1: Benchmarks used in the contamination analysis, with domains and sizes. * indicates that only text-only instances were used (BLUEX)

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

Beyond their technical role, each benchmark carries cultural and educational relevance. The *ENEM Challenge* derives from Brazil's National High School Exam, the main gateway to higher education. *BLUEX* compiles multidisciplinary questions from university entrance exams of UNICAMP and USP, also targeting general knowledge and reasoning. *OAB Exams* originates from the Brazilian Bar Association's certification tests, focused on legal reasoning and professional readiness. *HealthQA-BR* includes questions from national medical exams (Revalida and ENARE), emphasizing factual and clinical reasoning. Together, these benchmarks offer a balanced and representative evaluation of linguistic, cultural, and professional knowledge in Brazilian Portuguese, making them well suited for analyzing contamination and generalization in Portuguese LLMs.

The models. We evaluated a diverse set of LLMs selected for their academic or commercial relevance, covering both specialized Brazilian Portuguese and multilingual families. Table 2 summarizes the evaluated models and their main characteristics. We stratified models into three scales by parameter size: *small* ($\leq$13B), typically feasible on a single 40GB GPU and accessible to independent researchers; *medium* (14B to 30B), which require 80GB GPUs or optimized inference; and *large* ($>$30B), exceeding single-GPU capacity and representing state-of-the-art systems.

Model Specialized (PT-BR)	**Family**	**#Size**
Bode-13B-Alpaca-PT-BR	LLaMA	small
Gemma-3-Gaia-PT-BR-4B-it	Gemma	small
Gervásio-7B	LLaMA	small
openCabrita-3B	OpenLLaMA	small
Sabia-3.1 (2025-05-08)	Sabia	large
Sabia-7B	LLaMA	small
Tucano-2B4	Tucano	small
Model Multilingual	**Family**	**#Size**
Gemma-3-12B-it	Gemma	small
Granite-3.3-8B-Instr.	Granite	small
GPT-4o (2024-11-20)	OpenAI	large
Llama-3.1-8B-Instruct	LLaMA	small
Llama-4-Maverick-17B	LLaMA	medium
Ministral-8B-Instr.	Mistral	small
Phi-4	Phi	medium
Qwen3-8B	Qwen	small

Table 2: Evaluated models, grouped by specialization and scale.

Among specialized models, we selected the strongest available variant within each family. Except for Tucano-2B4 and Sabia-3.1, pretrained from scratch, all specialized models are fine-tuned adaptations for Brazilian Portuguese. Although Sabia-3.1 is specialized, we grouped it among large models due to its closed-access nature and competitive scale, aligning it more closely with GPT-4o. For multilingual models, we emphasized family diversity to broaden architectural coverage.

3.1.2. Execution and Analysis

We implemented a Python version of TS-Guessing within a fork of the *HELM Stanford* framework[4]. The protocol was integrated directly into HELM's evaluation pipeline, enabling unified computation of standard performance metrics and contamination indicators. To ensure methodological fidelity, we validated the implementation and conducted unit tests for consistent behavior.

The system adopts a modular design with components for data filtering, prompt construction (masking one incorrect alternative), and output normalization using NLTK. Leveraging HELM's client-based architecture, smaller models were executed locally (IBM Power9 and Google Colab), while larger ones were accessed via the *OpenRouter API*[5]. This setup ensured reproducible evaluation across all model–benchmark pairs (Figure 1).

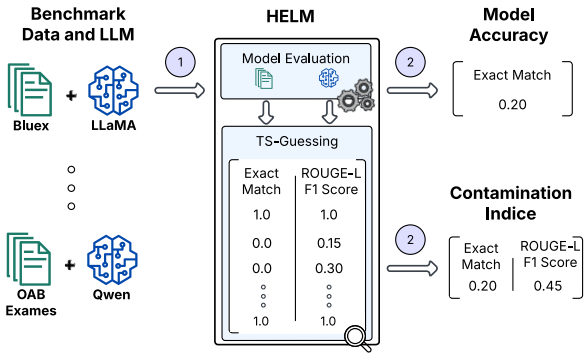

Figure 1: Unified pipeline for performance and contamination metrics. Each model–benchmark pair (1) produces standard accuracy and contamination indicators in a single execution (2).

We conducted 60 TS-Guessing runs covering all 15 models and 4 benchmarks. Model predictions for masked content were evaluated to compute contamination indicators (EM rate and mean ROUGE-L F1). The predefined contamination threshold (EM $\geq$ 0.02) was applied across analyses. Results below this level were interpreted cautiously, acknowledging that factors inherent to the architecture and training of the model may suppress behavioral evidence of contamination.

3.2. Method to Answer RQ2

Our corpus-level analysis to answer RQ2 is also defined by three main factors: the detection method, the benchmarks, and the target pretraining corpora.

[4] https://github.com/llm-pt-ibm/helm-guessing/

[5] https://openrouter.ai/

3.2.1. Factors

Substring Overlap — The corpus-level detection method. We selected the 50-character substring matching method (OpenAI, 2023) after reviewing alternative corpus-level detection strategies, as it provides a good balance between accuracy and computational efficiency. We applied the method to identify direct overlaps between benchmark items and pretraining corpora. Both datasets were normalized by removing non-alphanumeric characters, and up to three random 50-character substrings were sampled from each benchmark instance (or the full text if shorter). An instance was labeled as contaminated if any substring appeared verbatim in the corpus.

Metrics. We report two complementary indicators: the *Benchmark Leakage Rate (BLR)*, measuring the proportion of benchmark instances found in the corpus, and the *Contamination Dispersion (CD)*, measuring the empirical proportion of (benchmark instance × training batch) pairs in which at least one sampled substring overlap is detected. In summary, BLR captures how much of the benchmark leaked into the corpus, while CD captures how frequently leaked items are encountered across it. Unlike behavioral indicators, these corpus-level measures represent direct textual overlap and require no thresholding.

The benchmarks. The analysis covers the same four multiple-choice benchmarks used in RQ1 (*BLUEX*, *ENEM Challenge*, *HealthQA-BR*, and *OAB Exams*).

The corpora. Our analysis targets three large-scale Portuguese pretraining corpora selected for their relevance in recent model development. As detailed in Table 3, these include *Aroeira*, *GigaVerbo*, and *mC4-PT*, presented with their respective sizes and document counts.

Corpus	# Documents	Size (GB)
Aroeira	34,841,241	55.6
GigaVerbo	145,300,844	411.0
mC4-PT	169,073,811	312.0

Table 3: Pretraining corpora evaluated for benchmark leakage analysis.

Two corpora correspond directly to models analyzed in RQ1: *GigaVerbo* for *Tucano-2B4* and *mC4-PT* for *openCabrita-3B*, enabling corpus-to-behavior comparisons.

3.2.2. Execution and Analysis

Given the scale of the corpora, we implemented an adaptation of GPT-4 Technical Report method within the *LLMSanitize* library. Our fork[6] extends the support for streaming and batch-level comparison using the Hugging Face *datasets* API. Each corpus was processed in 6,000-example batches, comparing all samples against benchmark substrings. For reproducibility and fault tolerance, the system ran in parallel across multiple CPU workers with a fixed random seed (42) and periodic checkpoints.

Twelve runs (three corpora × four benchmarks) were executed on IBM Power9 and Power10 systems to ensure computational consistency. The pipeline recorded the number of processed samples, detected overlaps, and total comparisons, saving all matched benchmark items in structured JSONL files for transparent inspection and reconstruction of clean benchmark splits.

[6] https://github.com/llm-pt-ibm/LLMSanitize/

4. Results and Discussion

This section presents the experimental findings organized around the two research questions that guide our study.

4.1. Results for RQ1

To answer RQ1, we followed the model selection, benchmark scope, and evaluation setup defined in the methodology, applying the TS-Guessing protocol to assess behavioral contamination. This analysis identifies instances where model predictions reflect memorized benchmark content rather than genuine reasoning. Table 4 reports all model–benchmark pairs that met the behavioral contamination criterion (EM ≥ 0.02).

Model	Benchmark	Exact Match	ROUGE-L
Llama-3.1-8B-Instruct	BLUEX	0.10*	0.21
Sabia-3.1 (2025-05-08)	BLUEX	0.08	0.35*
GPT-4o (2024-11-20)	BLUEX	0.08	0.34
Sabia-3.1 (2025-05-08)	HealthQA-BR	0.04	0.26
Qwen3-8B	BLUEX	0.03	0.33
Sabia-3.1 (2025-05-08)	OAB Exams	0.03	0.31
GPT-4o (2024-11-20)	OAB Exams	0.03	0.31
Llama-4-Maverick-17B	BLUEX	0.03	0.29
Llama-4-Maverick-17B	OAB Exams	0.02	0.30
Gemma-3-12B-it	BLUEX	0.02	0.28
Sabia-3.1 (2025-05-08)	ENEM Challenge	0.02	0.28
Granite-3.3-8B-Instruct	BLUEX	0.02	0.26
GPT-4o (2024-11-20)	ENEM Challenge	0.02	0.26
GPT-4o (2024-11-20)	HealthQA-BR	0.02	0.26

Table 4: Model–benchmark pairs flagged for behavioral contamination according to TS-Guessing results (EM ≥ 0.02). Asterisks indicate the highest value observed for each metric.

Behavior in small-scale models. Most Brazilian Portuguese specialized models fall within the small-scale category, and none exhibited notable behavioral contamination. All six models — *Bode-13B-Alpaca-PT-BR*, *openCabrita-3B*, *Gemma-3-Gaia-PT-BR-4B*, *Gervasio-7B*, *Sabia-7B*, and *Tucano-2B4* — stayed below the contamination threshold

(EM < 0.02) across all benchmarks. Their ROUGE-L range (from 0.01 to 0.24) indicates minimal lexical overlap with masked alternatives. These results challenge the assumption that the scarcity of high-quality data in Portuguese inevitably leads to high overlap between benchmarks and training sets.

However, the absence of behavioral evidence does not preclude corpus-level leakage; it may instead suggest that small specialized models fail to internalize or exploit leaked benchmark content due to limited exposure, architectural constraints, or variations in training and fine-tuning quality. This pattern aligns with findings summarized by Deng et al. (2024a), who report that smaller models often memorize contaminated data without effectively using it, and with Sainz et al. (2023), who note that the lack of observable memorization does not necessarily imply the absence of training exposure.

Complementing, small multilingual models of similar size exhibited clearer signs of contamination. *Granite-3.3-8B-Instr.*, *Llama-3.1-8B-Instruct*, *Qwen3-8B*, and *Gemma-3-12B-it* all reached the contamination threshold on at least one benchmark, while *Mistral-8B-Instr.* showed only marginal overlap. All flagged cases occurred on *BLUEX*, where *Llama-3.1-8B-Instruct* reached the highest Exact Match overall (EM $= 0.10$), followed by *Qwen3-8B* (EM $= 0.03$) and both *Granite-3.3-8B* and *Gemma-3-12B-it* (EM $= 0.02$). *Qwen3-8B*, although below the threshold on *OAB Exams*, attained a ROUGE-L of 0.31, suggesting partial lexical recall without verbatim reproduction. Overall, even small multilingual models exhibit measurable behavioral contamination.

Medium and large models. Clear contamination patterns appeared among medium and large models. *Llama-4-Maverick-17B* reached the threshold on *BLUEX* and *OAB Exams*, while *Sabia-3.1* and *GPT-4o* showed stronger, consistent contamination signals. *Sabia-3.1* achieved EM $= 0.08$ on *BLUEX* and exceeded the threshold on all benchmarks. *GPT-4o* followed a similar trend, with EM $= 0.08$ on *BLUEX*, and exceeding the threshold elsewhere. ROUGE-L results reinforced this, showing high lexical overlap (0.34 to 0.35 on *BLUEX*; 0.31 on *OAB Exams*). These results align with prior findings reported by Xu et al. (2024a), who observed that larger and more capable models tend to exhibit stronger contamination effects and derive greater performance gains from leaked data.

Analysis and interpretation. Contamination intensity varied across benchmarks. *BLUEX* showed the strongest memorization evidence, achieving the highest Exact Match (EM) values. It was followed by *OAB Exams*, which also presented substantial signals and had the highest mean ROUGE-L score across models. While *BLUEX* and *ENEM Challenge* both assess general knowledge, *OAB Exams* focus on legal reasoning and contain dense, domain-specific terminology. This likely contributes to higher ROUGE-L scores, as it may prompt models to recall fragments of specific legal language even when unable to reconstruct the full option verbatim. In contrast, *ENEM Challenge* and *HealthQA-BR* exhibited lower overall EM averages and only mild indications of contamination, suggesting that the knowledge required for these tasks was likely learned indirectly from broader training sources rather than through direct benchmark exposure.

This variation across benchmarks converges with findings from Golchin and Surdeanu (2025) and Zhu et al. (2024a), who observed that contamination intensity fluctuates substantially between datasets even within the same model. Such results reinforce that benchmark properties — including domain, linguistic structure, and overlap frequency — play a decisive role in shaping how memorization signals manifest.

While contamination can inflate scores, our results show it does not necessarily reduce task difficulty. Despite being the most contaminated datasets, *BLUEX* and *OAB Exams* remained among the hardest benchmarks. A paired t-test confirms this pattern: *OAB Exams* was significantly harder than the less contaminated *ENEM Challenge* ($t(14) = 3.45$, $p = 0.0039$), and *BLUEX*, despite showing the strongest memorization signals, was statistically indistinguishable from *ENEM Challenge* ($t(14) = 1.96$, $p = 0.0702$). These results reinforce that contamination inflates performance without simplifying the underlying task, consistent with the separation between memorization and difficulty reported by Riddell et al. (2024).

This indicates that while contamination may boost an individual model's performance, it does not mean that more contaminated benchmarks are easier overall. A benchmark with a high average contamination level can still be inherently more challenging than a less contaminated one. Figure 2 illustrates this by plotting each benchmark's mean contamination against the mean accuracy achieved by the models, showing no clear tendency for more contaminated benchmarks to yield higher average scores. This interpretation aligns with Dekoninck et al. (2024), who define contamination as performance artificially inflated regardless of task difficulty, emphasizing that contaminated scores are unreliable indicators of real-world capability.

Next, we examined the consistency of models' recall patterns across benchmarks. Spearman correlations for ROUGE-L were uniformly high (average $\bar{\rho} = 0.969$, all $p < 0.001$). This consistency suggests that models reproduce stylistically or semantically similar text across benchmarks, rather than

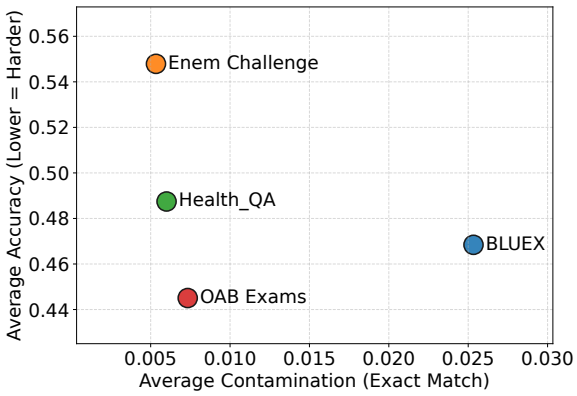

Figure 2: Relationship between benchmark contamination and model accuracy.

recalling content verbatim. In contrast, Exact Match correlations, though significant (all $p < 0.05$), were more moderate and variable (average $\bar{\rho} = 0.731$, ranging from $\rho = 0.626$ for *BLUEX* vs. *HealthQA-BR* to $\rho = 0.914$ for *HealthQA-BR* vs. *OAB Exams*). This indicates that exact recall is more benchmark-dependent, reflecting distinct memorization or reasoning behaviors, while general patterns of lexical approximation are shared.

Finally, though more contaminated benchmarks are not necessarily easier, a different pattern emerges from analyzing the models: higher behavioral contamination consistently yields better benchmark performance. Spearman correlation quantified this. The positive correlation was observed across individual model–benchmark pairs ($\rho_{\text{EM}} = 0.60$, $p < 0.001$) and within each benchmark: *BLUEX* ($\rho = 0.79$), *ENEM Challenge* ($\rho = 0.69$), *HealthQA-BR* ($\rho = 0.65$), and *OAB Exams* ($\rho = 0.52$). Though absolute correlation values vary by aggregation level, this positive and consistent association suggests that contamination systematically benefits model performance.

To demonstrate contamination's practical impact, we adjusted each model's accuracy by discounting any test instance where the model achieved an Exact Match in our TS-Guessing evaluation — cases likely reflecting verbatim recall of masked benchmark content. For instance, the *Llama-3.1-8B-Instruct* model dropped from 0.5726 to 0.4726 accuracy on *BLUEX*, lowering its rank from sixth to eighth. This shows how even moderate contamination can distort comparative evaluations and inflate perceived performance advantages. Similarly, Zhu et al. (2024b) found that decontaminating evaluation sets substantially alters model rankings and reduces benchmark scores, confirming that contamination directly biases performance comparisons.

Summary of findings. Our main findings for RQ1 can be specified as follows:

Key Finding 1 *Larger and multilingual models exhibited clear contamination signals.*

Key Finding 2 *Contamination intensity varied across evaluated benchmarks.*

Key Finding 3 *Models with higher benchmark accuracy are more likely to recall previously seen content.*

It is important to note that our results do not imply deliberate contamination, they may be accidental (as discussed in Section 2). They simply demonstrate that benchmark data leakage is real and that the publication of benchmark results can be adapted to also publish contamination results, particularly in underrepresented languages.

4.2. Results for RQ2

To answer RQ2, we analyzed the three pretraining corpora for benchmark data presence using the 50-character substring overlap method. We report two indicators: the **Benchmark Leakage Rate (BLR)**, which measures the percentage of benchmark items found within a corpus, and the **Contamination Dispersion (CD)**, measures the proportion of benchmark–batch pairs in which at least one sampled substring overlap is detected, reflecting the cross-batch spread of leaked benchmark content throughout the corpus. Table 5 summarizes the results.

Corpus	Benchmark	CD (%)	BLR (%)
Aroeira	*BLUEX*	0.0087	25.53
	ENEM Challenge	0.0875	66.62
	HealthQA-BR	0.0030	2.34
	OAB Exams	0.0584	34.25
GigaVerbo	*BLUEX*	0.0212	50.77
	ENEM Challenge	0.0905	90.22
	HealthQA-BR	0.0063	11.37
	OAB Exams	0.0983	98.46
mC4-PT	*BLUEX*	0.0083	35.72
	ENEM Challenge	0.0445	82.89
	HealthQA-BR	0.0043	5.45
	OAB Exams	0.0930	98.60

Table 5: Corpus-level contamination analysis using Contamination Dispersion (CD) and Benchmark Leakage Rate (BLR).

Corpus contamination patterns. Substantial benchmark leakage was detected across all three Portuguese pretraining corpora. Both *GigaVerbo* and *mC4-PT* contained over 80% of *ENEM Challenge* items and nearly all of *OAB Exams*, while the smaller and more curated *Aroeira* also showed notable traces of both. These results indicate that benchmark material is deeply embedded within Portuguese pretraining data, extending even to corpora

designed with stricter curation standards. This observation aligns with Almeida et al. (2025), who identified contamination of *ENEM Challenge* and *BLUEX* in other Portuguese corpora, and with concerns raised by Perełkiewicz and Poświata (2024) about the persistence of benchmark leakage in web-sourced datasets. Together, these findings suggest that contamination is a systemic issue, not confined to broad or poorly filtered corpora.

Disconnect between corpus leakage and behavioral contamination. Comparing corpus-level and behavioral results reveals a clear divergence between specialized and multilingual models. Despite near-complete leakage of benchmarks such as *ENEM Challenge* and *OAB Exams* in their corpora, the specialized models *Tucano-2B4* (trained on *GigaVerbo*) and *openCabrita-3B* (fine-tuned on *mC4-PT*) exhibited minimal behavioral contamination (EM $= 0.0$) and achieved relatively low benchmark accuracy. This reinforces the finding from RQ1 that smaller, domain-tuned models may not internalize or exploit leaked benchmark content effectively, likely due to limited capacity or constraints in the fine-tuning process — a pattern further supported by the low CD values observed.

In contrast, this gap disappears for multilingual models. As observed in RQ1, systems like *Llama-3.1-8B-Instruct* display clear behavioral contamination, indicating that corpus leakage can translate into measurable memorization effects even in smaller architectures. This aligns with Dekoninck et al. (2024)'s argument that contamination should be defined not merely as the presence of leaked data, but as the extent to which such data produce inflated benchmark performance. Accordingly, our results suggest that the ability to convert corpus leakage into behavioral gain depends more on model architecture and training strategy than on scale alone.

Leakage rate vs. Contamination Dispersion. Beyond the overall magnitude of leakage, distinct benchmarks exhibit distinct leakage dynamics. While benchmarks such as the *ENEM Challenge* display **widespread and diffuse** leakage — many items appear sparsely across corpora — others like *BLUEX* show more **localized** leakage, increasing the likelihood of a smaller subset of items appearing more frequently.

This difference reflects in lexical overlap: concentrated leakage yields higher ROUGE-L. For example, *Tucano-2B4* shows slightly higher accuracy on ENEM (0.2110, ROUGE-L = 0.04) than on BLUEX (0.2049, ROUGE-L = 0.05), i.e., a small rise in overlap with a minor accuracy drop. *openCabrita-3B* (trained on *mC4-PT*) exhibits the same direction for overlap but a different accuracy outcome: ENEM (accuracy 0.1781, ROUGE-L = 0.10) versus BLUEX (accuracy 0.2240, ROUGE-L = 0.14). Thus, concentrated leakage generally strengthens memorization signals (higher ROUGE-L), but whether that translates into higher accuracy depends on the model and training setup.

Summary of findings. Our main findings for RQ2 can be specified as follows:

Key Finding 4 *Substantial benchmark leakage was observed in the evaluated corpora (Aroeira, GigaVerbo and mC4-PT).*

Key Finding 5 *Evaluated benchmarks presented two leakage modes: widespread and diffuse (ENEM Challenge); localized and concentrated (BLUEX).*

Key Finding 6 *Not all models can effectively leverage leaked benchmark data, even when it is present in their training corpora.*

These findings highlight the inherent difficulty of curating large-scale training corpora. Completely filtering all benchmark traces from web-scale data is practically unfeasible given the diversity and overlap of online sources. Therefore, mitigation efforts are likely more effective on the benchmark side than on the corpus side. Instead of attempting to fully sanitize massive datasets, the community should focus on developing dynamic, encrypted, or privately maintained benchmarks.

5. Replication Package

All resources required to replicate this study are publicly available on GitHub[7]. The repository includes a step-by-step guide and all raw outputs (encompassing the JSONL files from our *LLMSanitize* and *HELM* executions, configuration details, and intermediate logs). It also provides links to the framework forks and the Brazilian Portuguese benchmarks and corpora used, ensuring full transparency and reproducibility.

6. Conclusion and Future Work

Our investigation into benchmark data contamination in Large Language Models for Brazilian Portuguese reveals both behavioral evidence of contamination and substantial corpus-level leakage. Behavioral signals were particularly strong in larger-scale models and showed a significant positive correlation with model accuracy, suggesting that memorization contributes to benchmark performance. At the corpus level, we identified a high degree of overlap between benchmark items and pretraining

[7] https://github.com/llm-pt-ibm/contamination-evaluation/

corpora such as *GigaVerbo* and *mC4-PT*. However, a notable finding was the discrepancy for small-scale specialized models: despite high overlap in their training data, these models did not exhibit behavioral evidence of contamination, suggesting that factors such as model capacity or training strategy play a key role in the ability to effectively leverage leaked content.

These findings suggest that similar patterns may emerge in other underrepresented languages, given that the dynamics between corpora, models, and benchmarks follow similar trends. As demonstrated in previous audits of multilingual corpora, non-English data often undergo less rigorous curation and exhibit greater quality variability, overlap, and a high prevalence of machine-translated content (Kreutzer et al., 2022; Thompson et al., 2024). The scarcity of high-quality data for these languages encourages the use of public web data for constructing local models and corpora, as seen in IndicLLMSuite (Khan et al., 2024) for Indian languages, AfroLM (Dossou et al., 2022) for African languages, and SEA-LION (Ng et al., 2025) for Southeast Asian languages.

Furthermore, the creation of benchmarks for these languages follows the same tendency of utilizing web-extracted content (such as public exams and social media data) or directly translating existing English evaluation sets, as observed in Uhura (Bayes et al., 2024) and AfriSenti (Muhammad et al., 2023) for the African context, MiLiC-Eval (Zhang et al., 2025) for minority languages in China, and recently in QUENCH (Khan et al., 2025) for Indic languages, SloPragEval (Brglez and Vintar, 2025) for Slovene, and across multiple Turkish datasets (Umutlu et al., 2025).

From a contamination assessment perspective, future work can extend this investigation to additional underrepresented languages and multimodal contexts. Alternative evaluation strategies are also worth exploring. At the corpus level, methods based on semantic similarity rather than exact substring overlap could enrich the analysis and reveal new contamination dynamics, as shown in Riddell et al. (2024). At the behavioral level, evaluations comparing model outputs on original and paraphrased benchmark instances may provide a more nuanced detection signal, similar to approaches proposed by Dekoninck et al. (2024) and Xu et al. (2024b).

From a mitigation perspective, efforts may be better directed toward strategies that limit benchmark exposure. Given the inherent challenges of large-scale corpus curation, benchmarks represent a more tractable intervention point. Future work should prioritize evaluation sets built from novel or restricted-access content, as advocated by Sainz et al. (2023), incorporating mechanisms such as watermarking, encrypted access, periodic item rotation, or controlled distribution to reduce accidental exposure during pretraining, as recommended by Xu et al. (2024a).

7. Limitations and Threats to Validity

This work has two primary limitations. First, our corpus analysis employed a substring matching method, which is effective for detecting literal text reuse, but does not capture paraphrased or semantically similar content. Second, the TS-Guessing protocol is an inferential detection method; as such, it identifies observed model behaviors that suggest memorization, but these observations are not direct proof of contamination and instead reflect the model's behavior under specific conditions.

8. Acknowledgements

This work was carried out within the scope of a research project developed in partnership between Universidade Federal de Campina Grande (UFCG) and IBM. The first author is supported by a scholarship from CAPES.

9. Bibliographical References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-3 Technical Report. arXiv preprint arXiv:2410.12049.

Thales Sales Almeida, Thiago Laitz, Giovana K. Bonás, and Rodrigo Nogueira. 2023. BLUEX: A benchmark based on Brazilian leading universities entrance eXams. In *Intelligent Systems (BRACIS)*, pages 337–347. Springer Nature Switzerland.

Thales Sales Almeida, Rodrigo Nogueira, and Helio Pedrini. 2025. Building high-quality datasets for portuguese llms: From common crawl snapshots to industrial-grade corpora. arXiv preprint arXiv:2509.08824.

Edward Bayes, Israel Abebe Azime, Jesujoba O. Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948*.

Mojca Brglez and Špela Vintar. 2025. From polyester girlfriends to blind mice: Creating the first pragmatics understanding benchmarks for slovene. *arXiv preprint arXiv:2510.21575*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar De Oliveira. 2024. TeenyTinyLlama: Open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications*, 16:100558.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*. In Press.

Andrew Maranhão Ventura D'addario. 2025. HealthQA-BR: A system-wide benchmark reveals critical knowledge gaps in large language models. arXiv preprint arXiv:2506.21578.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. 2024. Constat: Performance-based contamination detection in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 92420–92464. Curran Associates, Inc.

Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. 2017. Passing the Brazilian OAB Exam: Data Preparation and Some Experiments. In *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference*, volume 302 of *Frontiers in Artificial Intelligence and Applications*, pages 89–94. IOS Press.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024a. Unveiling the spectrum of data contamination in language models: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024b. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. In *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Shahriar Golchin and Mihai Surdeanu. 2025. Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Mod-

els. *Transactions of the Association for Computational Linguistics*, 13:809–830.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Roseval Malaquias Junior, Ramon Pires, Roseli Romero, and Rodrigo Nogueira. 2024. Juru: Legal Brazilian large language model from reputable sources. arXiv preprint arXiv:2403.18140.

Mohammad Aflah Khan, Neemesh Yadav, Sarah Masud, and Md. Shad Akhtar. 2025. Quench: Measuring the gap between indic and non-indic contextual general reasoning in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4493–4509, Abu Dhabi, UAE. Association for Computational Linguistics.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. arXiv preprint arXiv:2308.11878.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language Ranker: A metric for quantifying LLM performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194. AAAI Press.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Thiago Lira, Flávio Cação, Cinthia Souza, João Valentini, Edson Bollis, Otavio Oliveira, Renato Almeida, Marcio Magalhães, Katia Poloni, Andre Oliveira, et al. 2025. Aroeira: A curated corpus for the portuguese language with a large number of tokens. In *Intelligent Systems*, volume 15412 of *Lecture Notes in Computer Science*, pages 185–199. Springer, Cham.

Shamsuddeen Hassan Muhammad et al. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Ng, others, and Leslie Teo. 2025. Sea-lion: Southeast asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Michał Perełkiewicz and Rafał Poświata. 2024. A review of the challenges with massive web-mined corpora used in large language models pre-training. In *Artificial Intelligence and Soft Computing*, volume 14678 of *Lecture Notes in Computer Science*, pages 153–163. Springer, Cham.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and

Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog. Accessed on: 2024-10-25.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are llms contaminated? a comprehensive survey and the llmsanitize library. arXiv preprint arXiv:2404.00699.

Martin Riddell, Ansong Ni, and Arman Cohan. 2024. Quantifying contamination in evaluating code generation capabilities of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14116–14137, Bangkok, Thailand. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Igor Cataneo Silveira and Denis Deratani Mauá. 2018. Advances in automatically solving the enem. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48.

Dingjie Song, Sicheng Lai, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang. 2025. Both text and images leaked! a systematic analysis of data contamination in multimodal LLMs. In *ICML 2025 Workshop on Data in Generative Models - The Bad, the Ugly, and the Greats*.

Yifan Sun, Han Wang, Dongbai Li, Gang Wang, and Huan Zhang. 2025. The emperor's new clothes in benchmarking? a rigorous examination of mitigation strategies for LLM benchmark data contamination. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, Vancouver, Canada. PMLR. To appear.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. Evaluating the quality of benchmark datasets for low-resource languages: A case study on turkish. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024a. Benchmark data contamination of large language models: A survey. arXiv preprint arXiv:2406.04244.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024b. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. Milic-eval: Benchmarking multilingual llms for china's minority languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11086–11102, Vienna, Austria. Association for Computational Linguistics.

Qin Zhu, Qingyuan Cheng, Runyu Peng, Xiaonan Li, Tengxiao Liu, Ru Peng, Xipeng Qiu, and Xuanjing Huang. 2024a. Inference-time decontamination: Reusing leaked benchmarks for large language model evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9113–9129, Miami, Florida, USA. Association for Computational Linguistics.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yun-Ze Song, Yumeng Zhang, Hanxu Hu, Yiran Wei, Rui Wang, and Hongyuan Lu. 2024b. CLEAN-EVAL: Clean evaluation on contaminated large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 835–847, Mexico City, Mexico. Association for Computational Linguistics.

10. Language Resource References

22-hours. 2023. openCabrita-3B. https://huggingface.co/22h/open-cabrita3b. Accessed at: 2025-10-17.

C. G. Camilo-Junior, S. S. T. Oliveira, L. A. Pereira, M. Amadeus, R. Scotti, D. Fazzioni, A. M. A. Novais, and S. A. A. Jordão. 2025. GAIA: An Open Language Model for Brazilian Portuguese. https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it. Accessed at: 2025-10-17.

Eduardo Garcia. 2023. mc4-pt. https://huggingface.co/datasets/eduagarcia/mc4-pt. Accessed at: 2025-10-17.

Eduardo Garcia. 2024a. enem_challenge. https://huggingface.co/datasets/eduagarcia/enem_challenge. Accessed at: 2025-10-17.

Eduardo Garcia. 2024b. oab_exams. https://huggingface.co/datasets/eduagarcia/oab_exams. Accessed at: 2025-10-17.

Google. 2025. Gemma-3-12B-it. https://huggingface.co/google/gemma-3-12b-it. Accessed at: 2025-10-17.

Granite Team, IBM. 2025. Granite-3.3-8B-Instruct. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct. Accessed at: 2025-10-17.

Itaú Unibanco. 2024. Aroeira. https://huggingface.co/datasets/Itau-Unibanco/aroeira. Accessed at: 2025-10-17.

Larxel. 2025. healthqa-br. https://huggingface.co/datasets/Larxel/healthqa-br. Accessed at: 2025-10-17.

Maritaca AI. 2023. Sabia-7B. https://huggingface.co/maritaca-ai/sabia-7b. Accessed at: 2025-10-17.

Maritaca AI. 2025. Sabia 3.1. https://docs.maritaca.ai/api/pt/list-models. Accessed at: 2025-10-17.

Meta AI. 2024. Llama 3.1 8B Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Accessed at: 2025-10-17.

Meta AI. 2025. Llama-4-Maverick-17B-Instruct. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-Instruct. Accessed at: 2025-10-17. Model referred to as Llama-4-Maverick-17B in table 2.

Microsoft. 2024. Phi-4. https://huggingface.co/microsoft/phi-4. Accessed at: 2025-10-17. Model referred to as Phi-4 (mid) in table 2.

Mistral AI. 2024. Ministral-8B-Instruct. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Accessed at: 2025-10-17. Closest official model to Ministral-8B-Instr. in table 2.

OpenAI. 2024. GPT-4o. https://platform.openai.com/docs/models/gpt-4o. Accessed at: 2025-10-17. Model referred to as GPT-4o-2024-11-20 in table 2.

Portuguese Benchmark Datasets. 2023. BLUEX. https://huggingface.co/datasets/portuguese-benchmark-datasets/BLUEX. Accessed at: 2025-10-17.

PORTULAN. 2024. Gervásio-7B. https://huggingface.co/PORTULAN/gervasio-7b-portuguese-ptbr-decoder. Accessed at: 2025-10-17.

Qwen Team, Alibaba Cloud. 2025. Qwen3-8B. https://huggingface.co/Qwen/Qwen3-8B. Accessed at: 2025-10-17. Model referred to as Qwen3-8B in table 2.

recogna-nlp. 2024. Bode-13B-Alpaca-PT-BR. https://huggingface.co/recogna-nlp/bode-13b-alpaca-pt-br-no-peft. Accessed at: 2025-10-17.

TucanoBR. 2024a. GigaVerbo. https://huggingface.co/datasets/TucanoBR/GigaVerbo. Accessed at: 2025-10-17.

TucanoBR. 2024b. Tucano-2B4. https://huggingface.co/TucanoBR/Tucano-2b4. Accessed at: 2025-10-17.