CURUPIRA: Clever guard for harm and linguistic prompt mitigation in Brazilian Portuguese

Rogério Sousa
William Alberto Cruz-Castañeda
José Roberto Homeli da Silva
Marcellus Amadeus
SoberanIA
{rogerio,william,homeli,marcellus}@soberania.ai

Abstract

The safe deployment of Large Language Models remains challenging in multilingual settings, particularly when models are exposed to adversarial or malicious prompts in underrepresented languages. In this work, we present Curupira, a Brazilian Portuguese-language guard model designed to mitigate harmful prompt exploitation. To do this, we establish a three steps methodology that involves adaptation, data generation, and fine-tuning. We also evaluate our model with two state-of-the-art open guardrail architectures. The results show that targeted fine-tuning leads to consistent improvements in safety classification for Portuguese prompts, with favorable efficiency–performance trade-offs for compact models and limited degradation in cross-lingual evaluation.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance across a wide range of natural language processing tasks. However, their susceptibility to producing harmful, biased, or unethical outputs when exposed to adversarial or malicious prompts remains a critical challenge for safe deployment in real-world applications (Weidinger et al., 2022). To mitigate these risks, modern LLMs systems increasingly rely on specialized moderation components, commonly referred to as *guardrails* or *safeguards*—auxiliary models designed to filter inputs and outputs according to predefined safety taxonomies.

Despite their effectiveness in English-centric settings, many guardrail models exhibit a pronounced *safety gap* in multilingual scenarios. State-of-the-art (SOTA) approaches such as Llama-Guard (Inan et al., 2023) and ShieldGemma (Zeng et al., 2024) are predominantly trained on English data, leading to inconsistent behavior when exposed to non-English inputs (Yang et al., 2024). Even recent multilingual initiatives, including Qwen3Guard (Zhao et al., 2025), remain vulnerable to challenges such as code-mixing and culturally grounded expressions, which can obscure harmful intent and bypass safety mechanisms (Banerjee et al., 2025). For Portuguese, these issues are amplified by the scarcity of high-quality, safety-aligned corpora, resulting in guardrails that are susceptible to adversarial prompts, regional slang, and culturally specific linguistic patterns (Brito et al., 2025; Bansal and Mishra, 2025).

This challenge gains additional relevance in light of recent advances in the Brazilian LLMs ecosystem. The emergence of open and locally developed models—such as Sabiá, Gaia, Amadeus-Verbo, and Tucano—marks an important step toward linguistic and technological autonomy (Cruz-Castañeda and Amadeus, 2025). However, the deployment of these models in real-world applications also raises pressing concerns regarding safety, robustness, and resistance to harmful or manipulative inputs in Brazilian Portuguese. Ensuring effective guardrails is therefore a prerequisite for the responsible adoption of national and regional LLMs.

In this work, we address these limitations by developing **Curupira**, a Brazilian Portuguese-language guard model for harmful prompt mitigation. The name is inspired by the Curupira, a figure from Brazilian folklore traditionally depicted as a protector of forests who misleads those intent on causing harm[1]. Analogously, our model is designed to protect Brazilian Portuguese LLMs by identifying and neutralizing harmful or adversarial inputs before they cause downstream damage.

Methodologically, we adapt established English-centric safety frameworks to the Lusophone context and demonstrate that Supervised Fine-Tuning (SFT) with high-quality synthetic data can substantially improve guardrail robustness in Portuguese. Our contributions are threefold: (i) adapting En-

[1] https://pt.wikipedia.org/wiki/Curupira

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1038–1043

glish safety taxonomies to Portuguese through synthetic data generation; (ii) an empirical evaluation of the effectiveness of SFT in improving guardrail performance for Portuguese prompts; and (iii) a comparative analysis of safety adaptation across SOTA architectures.

The remainder of this paper is organized as follows. Section 2 reviews related work on multilingual LLMs safety and guardrail models. Section 3 details the three steps methodology for adaptation, data generation and fine-tuning. Section 4 describes the experimental setup and evaluation protocol. Section 5 presents and analyzes the experimental results. Finally, Section 6 concludes the paper and outlines directions for future work on Brazilian Portuguese-language safety alignment.

2 Related Work

Recent work on safety mechanisms for LLMs has primarily focused on instruction-based fine-tuning of specialized guardrails to mitigate vulnerabilities like prompt injection, where malicious inputs hijack goals or leak system instructions (Perez and Ribeiro, 2022). Representative approaches include Llama-Guard (Inan et al., 2023), Aegis (Ghosh et al., 2024), WildGuard (Han et al., 2024), and ShieldGemma (Zeng et al., 2024), which differ in scope and design by targeting harmful user inputs, unsafe model outputs, or both. These models typically rely on predefined safety taxonomies and supervised signals to enforce policy compliance. From a multilingual perspective, the Qwen3Guard family (Zhao et al., 2025) represents a notable advance, supporting numerous languages and introducing tri-class safety judgments. However, empirical safety evaluations in Portuguese remain limited. Initial efforts indicate that Portuguese-focused models better discern cultural nuances, such as distinguishing harmless slang from actual toxicity, compared to broader multilingual LLMs (da Silva Oliveira et al., 2024). Despite this progress, empirical evaluations of guardrail adaptation to Portuguese remain limited.

3 Methodology

We focus on bridging the safety gap for Brazilian Portuguese language through a three steps data-centric approach, involving taxonomy adaptation, synthetic data generation, and SFT.

3.1 Safety Taxonomy

The four core safety categories established in the ShieldGemma technical report (Zeng et al., 2024): (i) Hate Speech, (ii) Harassment, (iii) Sexually Explicit, and (iv) Dangerous Content served as the conceptual foundation. We utilized these categories as a baseline to drive the creation of a Portuguese-specific safety corpus. This approach was designed to incorporate linguistic nuances and cultural contexts, such as regional slangs and adversarial patterns prevalent in Lusophone social media.

3.2 Data Generation

We developed a pipeline based on the AI-Assisted Red-Teaming (AART) framework (Radharapu et al., 2023). For the generation stage, we implemented a sequential strategy using three distinct prompting layers: (1) conceptual mapping to establish specific terminology and slangs native to the Portuguese language; (2) structural typology to ensure syntactic variety; and (3) instance synthesis to produce high-fidelity toxic content. This process yielded a balanced dataset of safe/unsafe pairs, providing the necessary contrastive signals to distinguish between benign intent and malicious prompts in the Portuguese language.

3.3 Supervised Fine-Tuning

In this step, the objective is to adapt a model to Brazilian Portuguese language while preserving their original safety reasoning capabilities.

4 Experimental Setup

This section describes the experimental pipeline adopted to develop and evaluate the Curupira model. The setup follows the methodological procedure, a baseline evaluation of pretrained models, and post-training evaluation.

4.1 Synthetic Dataset Instantiation

Following the methodology described in Section 3, we instantiated a Portuguese synthetic safety dataset comprising 17,513 labeled examples of balanced safe/unsafe pairs. For all experiments, the dataset was stratified into training, validation, and test splits using an 80/10/10 ratio. This synthetic corpus was used exclusively for SFT, internal validation, and controlled comparison between pre- and post-training model performance.

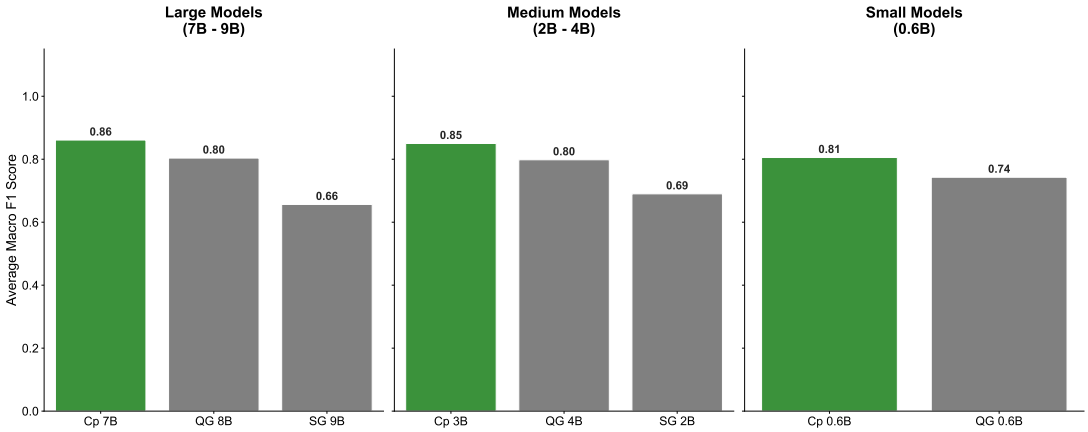

Figure 1: Overall macro F1-score comparison between baseline (gray bars) and Curupira models (green bars). Scores are averaged across all datasets and safety categories. Here "QG" means Qwen3Guard, "SG" means ShieldGemma, and "Cp" means Curupira.

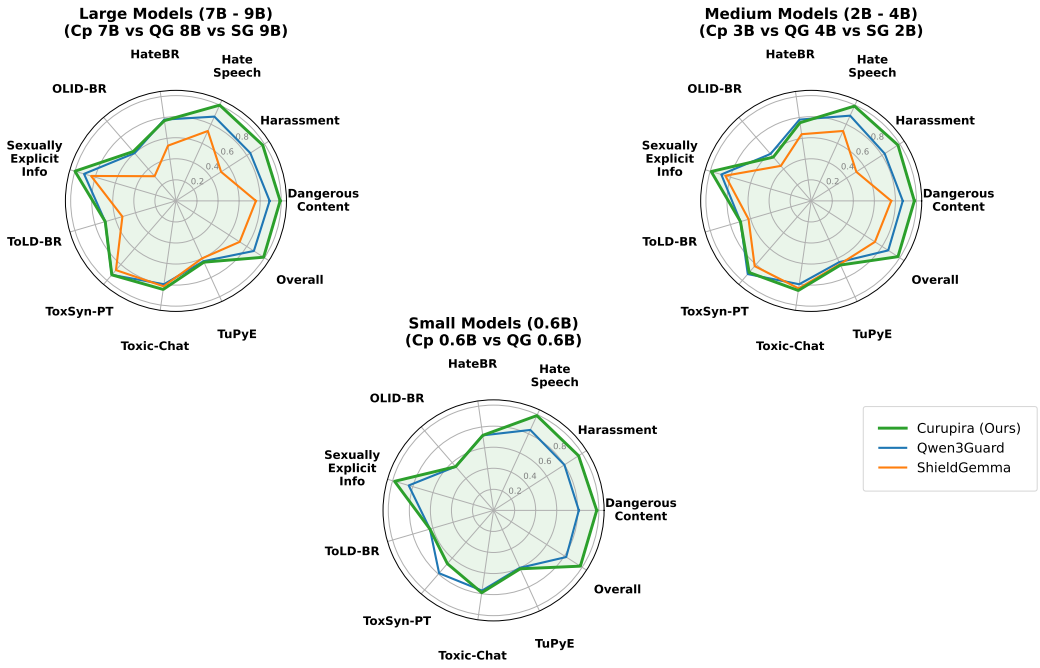

Figure 2: Radar plots comparing baselines (blue and orange polygons) and Curupira (green polygons) models performance profiles. Each axis corresponds to a dataset or safety category, illustrating behavioral trade-offs induced by fine-tuning. Here "QG" means Qwen3Guard, "SG" means ShieldGemma, and "Cp" means Curupira.

4.2 Evaluation Benchmarks

Model performance was evaluated using a diverse set of public benchmarks designed to assess robustness under realistic linguistic conditions. These datasets differ substantially from the synthetic training data in terms of source, annotation methodology, and linguistic variability, and are primarily composed of user-generated content from social media platforms.

The evaluation suite includes five Brazilian Portuguese datasets: ToxSyn-PT (Brito et al., 2025), OLID-BR (Trajano et al., 2024), TuPy-E (Oliveira et al., 2023), HateBR (Vargas et al., 2022), and ToLD-BR (Leite et al., 2020). These corpora exhibit substantial linguistic noise, including slang, abbreviations, informal syntax, and context-dependent interpretations. In addition, we include Toxic-Chat (Lin et al., 2023), an English-language dataset consisting of real user prompts collected from an online chatbot demonstration. This benchmark is used to assess potential cross-lingual performance degradation and catastrophic forgetting effects induced by fine-tuning on Portuguese-only data.

4.3 Baseline Evaluation

As a first step, we evaluated original pretrained models to establish a baseline and quantify the inherent gap when models trained predominantly on English are applied directly to Brazilian Portuguese inputs. This baseline consists of ShieldGemma 1 (2B, 9B variants) and Qwen3Guard-Gen (0.6B, 4B, and 8B variants) using the full set of synthetic test data and public benchmarks.

4.4 Supervised Fine-Tuning

All fine-tuning experiments shared the following configuration: loss computation restricted to generated completions, bfloat16 precision, a maximum sequence length of 8,192 tokens, and the paged AdamW optimizer. We systematically varied the learning rate, number of epochs, and warmup ratio across five experimental configurations to analyze stability, convergence behavior, and generalization.

4.5 Post-Fine-Tuning Evaluation

All adapted models were re-evaluated using the same synthetic test split and public benchmarks employed in the baseline phase. This post-training evaluation enables direct comparison of pre- and post-fine-tuning performance, isolating the impact of Portuguese adaptation. Particular attention was given to changes in false positive and false negative rates, as well as to potential degradation on the English Toxic-Chat benchmark, which serves as an indicator of cross-lingual interference.

4.6 Evaluation Metrics and Output Normalization

Given the class imbalance typical of public safety benchmarks, where unsafe instances constitute a minority of samples, we report F1-score as the primary evaluation metric. This metric provides a direct and interpretable assessment of false positive and false negative trade-offs, which are critical in safety-sensitive applications. To enable direct empirical comparison across architectures, we normalized model outputs into a unified binary decision space. ShieldGemma's generative Yes/No predictions were aligned with Qwen's categorical outputs (Safe, Unsafe, Controversial). For binary evaluation, both Unsafe and Controversial predictions from Qwen were treated as positive instances indicating a safety violation.

5 Results and Analysis

This section analyzes the performance of baseline and Curupira models, focusing on overall effectiveness, robustness across datasets, and behavioral trade-offs induced by SFT.

Figure 1 provides a global comparison of all evaluated models, aggregating performance across datasets and safety categories. SFT consistently improves macro F1-scores across all models, confirming the effectiveness of the synthetic Portuguese dataset.

Among all configurations, Curupira-7B achieves the highest overall performance, followed closely by Curupira-3B. Notably, Curupira-0.6B exhibits substantial gains. This result highlights a favorable efficiency–performance trade-off, indicating that compact architectures can achieve competitive safety performance when properly adapted. Beyond aggregate performance, fine-tuning yields consistent improvements across most evaluation benchmarks. The largest gains are observed in categories explicitly targeted during synthetic data generation, including Overall, Harassment, Sexually Explicit Information, Dangerous Content, and Hate Speech. At the same time, consistent improvements are also observed across several publicly available Portuguese-language datasets containing naturally occurring harmful content, indicating that the benefits of SFT extend beyond the synthetic training distribution. Performance regressions are limited and localized, primarily affecting datasets with greater distributional or linguistic divergence. No model exhibits widespread degradation across benchmarks, indicating that SFT improves task-specific performance without compromising general safety reasoning. As shown in Figure 2, Curupira models consistently expand outward relative to their baselines, reflecting gains across most datasets and safety dimensions. Localized contractions are observed in specific axes for Curupira models, including the ToxSyn-PT dataset in the Curupira-0.6B configuration, consistent with dataset-specific or capacity-related trade-offs. The absence of abrupt collapses suggests that fine-tuning does not induce overfitting or catastrophic forgetting. Overall, these results indicate that SFT on a synthetic Portuguese safety dataset yields stable improvements across model families and scales, while the strong post-SFT performance of smaller models supports the feasibility of efficient guardrail deployment.

6 Conclusion and Limitations

This work investigated the development of open guardrails models for Brazilian Portuguese using a data-centric SFT approach. Results show that fine-tuning on a synthetic safety dataset yields consistent improvements in safety classification across public benchmarks, with Curupira models demonstrating favorable efficiency-performance trade-offs. Limitations include reliance on synthetic data, which may introduce generation-related biases and limit coverage of real-world linguistic variability; the absence of human evaluation; and the restriction to text-based classification. Despite these constraints, the findings highlight the relevance of open, locally adapted safety models, particularly in reducing dependence on closed and English-centric moderation systems. Future work will explore parameter-efficient fine-tuning for larger models, the integration of human evaluation protocols, and the extension of the proposed approach to multimodal and generative safety settings.

Acknowledgments

This work was developed within the scope of the SoberanIA[2] initiative, a Brazilian effort initiated in the state of Piauí, focused on ethically sourced, auditable data and technological sovereignty.

References

Somnath Banerjee, Pratyush Chatterjee, Shanu Kumar, Sayan Layek, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Attributional Safety Failures in Large Language Models under Code-Mixed Perturbations. *arXiv preprint*. ArXiv:2505.14469 [cs] version: 2.

Lavish Bansal and Naman Mishra. 2025. CREST: Universal Safety Guardrails Through Cluster-Guided Cross-Lingual Transfer. *arXiv preprint*. ArXiv:2512.02711 [cs] version: 1.

Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Diogo Fernandes Costa Silva, and Arlindo Rodrigues Galvão Filho. 2025. ToxSyn-PT: A Large-Scale Synthetic Dataset for Hate Speech Detection in Portuguese. *arXiv preprint*. ArXiv:2506.10245 [cs].

William Alberto Cruz-Castañeda and Marcellus Amadeus. 2025. Large languages models in brazilian portuguese: A chronological survey. *Journal of the Brazilian Computer Society*, 31(1):1167–1186.

[2]https://soberania.ai

Amanda da Silva Oliveira, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas, and Eduardo José da Silva Luz. 2024. Toxic speech detection in portuguese: A comparative study of large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 108–116.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in Neural Information Processing Systems*, 37:8093–8131.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 914–924.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. *arXiv preprint*. ArXiv:2310.17389 [cs].

Felipe Oliveira, Victoria Reis, and Nelson Ebecken. 2023. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *Preprint*, arXiv:2211.09527.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications. *arXiv preprint*. ArXiv:2311.08592 [cs].

Douglas Trajano, Rafael H. Bordini, and Renata Vieira. 2024. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 214–229, Seoul Republic of Korea. ACM.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2024. Benchmarking llm guardrails in handling multilingual toxicity. *arXiv preprint arXiv:2410.22153*.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. ShieldGemma: Generative AI Content Moderation Based on Gemma. *arXiv preprint*. ArXiv:2407.21772 [cs].

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, and 24 others. 2025. Qwen3Guard Technical Report. ArXiv:2510.14276 [cs].