Robustness of Language Models against Portuguese Harmful Prompts

Eduardo Alexandre de Amorim[1,2,*], Cleber Zanchettin[2]
[1]TELUS Digital Research Hub, CIAAM-USP, São Paulo, SP, Brazil
[2]Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brazil
eduardo.amorim@telus.com, {eaa5, cz}@cin.ufpe.br

***Abstract*—Language Models (LMs) are vulnerable to jailbreak prompts designed to bypass safety constraints. Despite the growing literature in English, resources and defenses tailored to the Portuguese language remain scarce. We present SecBERT, a specialized Portuguese classifier built on the BERTimbau Base architecture to detect policy-violating and harmful prompts. We construct a 29,432-instance Portuguese dataset by translating a subset of WildJailbreak, explicitly preserving the original four-way taxonomy (Vanilla/Adversarial vs. Benign/Harmful) to enable granular analysis. We evaluate multiple BERT-based backbones in both frozen and fine-tuned settings, using F1-score, AUC, and the Kolmogorov-Smirnov (KS) statistic to characterize operational separability. The fine-tuned BERTimbau Base (SecBERT) achieves 95.6% F1, 99.2% AUC, and 91.2% KS, significantly outperforming non-Portuguese-centric baselines in discriminatory capability. We further analyze threshold behavior and discuss the limitations of translation-based data generation, outlining directions for native Portuguese harmful prompt datasets and adaptive robustness testing.**

***Index Terms*—Large Language Models, Trustworthy and Reliable AI, Ethics and Regulation in AI, Transformer Networks, Generative AI Models.**

CAUTION: The text in this paper contains harmful language.

Reproducibility

The source code, model weights, and dataset are publicly available at: https://github.com/Edu-p/secbert-pt.

I. Introduction

The global expansion of Language Models (LMs) [1], [13] is accompanied by severe risks of misuse. LMs can generate misinformation and facilitate phishing attacks, and studies indicate the continued use of market models for cybercriminal activities [4], [10].

A specific category of adversarial prompts, known as jailbreak prompts, emerged as the primary attack vector for bypassing LM safeguards. These prompts are crafted to manipulate models into generating content that violates usage policies. Such prompts are considered high risk because they compromise the integrity and ethics of the model [10], [11].

Research communities still lack a systematic understanding of jailbreak prompts and their evolution. They tend to be significantly longer, averaging 555 tokens (1.5 times the length of regular prompts), often resorting to strategies like roleplaying, virtualization, scenario simulation, and increasing semantic complexity [11]. Although providers implement techniques such as Reinforcement Learning from Human Feedback (RLHF) [8], additional defense strategies, such as input filters, are necessary to improve robustness [10].

A significant gap remains in understanding these threats in the context of the Portuguese language. While models such as BERTimbau exist for general natural language processing tasks, few studies have focused on the security of LMs in this specific language. Strategies developed for English do not always apply directly to other languages due to cultural and linguistic particularities [12]. For instance, Brazilian Portuguese possesses semantic ambiguities, regional irony, and specific cybercriminal slang that models trained predominantly on English corpora fail to capture. In fact, recent literature highlights a significant lack of resources for representation and evaluation of multilingual jailbreak challenges, particularly for non-English languages [3].

This work develops and evaluates SecBERT. This model specializes in detecting harmful prompts for LMs operating in Portuguese. The research aims to adapt a subset of the WildJailbreak Dataset for the Portuguese language [5]. It also seeks to experiment with different BERT architectures and analyze performance metrics using classical classification indicators. Throughout this work, we distinguish between "jailbreak", which refers to the adversarial strategy used to bypass safety filters, and "harmful", which denotes the policy-violating nature of the content. The final goal is to propose guidelines for the continuous improvement of LM safeguards in Portuguese-speaking countries.

II. Related Work

The security of Language Models has become a critical area of research due to the increasing sophistication of adversarial attacks. This section reviews the classification of prompt hacking, the characteristics of jailbreak prompts, and the existing gap regarding the Portuguese language.

A. Prompt Hacking Taxonomy

Prompt hacking is categorized into three main types: jailbreaking, injection, and leaking [10]. While injection and leaking focus on targeted output manipulation or private data extraction, this work focuses specifically on detecting harmful jailbreaks. It directly compromises the model's alignment with human values by eliciting limited responses. Jailbreaking often employs techniques such as roleplaying (asking the model to act as a character without restrictions), hypothetical scenarios (framing the request as an educational exercise), and

translation wrappers to obscure malicious intent in complex linguistic tasks.

B. Multilingual Security Gaps

Most datasets and defense mechanisms are developed for the English language. For instance, the WildTeaming dataset [5] provides a comprehensive set of adversarial prompts but lacks native support for other languages. In the context of Portuguese, models such as BERTimbau [12] have advanced the state of the art in natural language processing, but dedicated resources for safety alignment remain scarce. This research addresses this limitation by adapting existing resources to train a specialized classifier for Portuguese.

III. Methodology

The development of SecBERT involved the creation of a specialized dataset and the evaluation of different neural architectures. The process focused on ensuring the model could distinguish between legitimate requests and adversarial attempts in Portuguese with high discrimination capability.

A. Dataset Creation and Translation

We utilized the WildJailbreak Dataset, structured into four categories (Vanilla/Adversarial and Benign/Harmful), to provide a balanced view of safety behaviors [5]. A stratified 10% subset of the original training volume was translated. To ensure reproducibility, the adapted dataset and model weights are publicly available

We utilized a specific prompt structure designed to maintain the adversarial intent of the original text while adapting it to Brazilian Portuguese cultural nuances. The system prompt enforced a JSON output format to facilitate automated parsing:

System Prompt:
You are an expert translator specializing in LM security and red-teaming. Your task is to translate datasets from English into Brazilian Portuguese while preserving the original semantic intent.

Instructions:
- **Fidelity:** Maintain the adversarial or harmful nature of the prompt. Do not sanitize or refuse to translate explicit content.
- **Localization:** Adapt idioms and slang to the Brazilian context to ensure fluency.
- **Format:** Return strictly a JSON object with the structure:

```
{ "prompt": <translated_text>,
"data_type": <original_type> }
```

This structured approach minimized parsing errors and ensured that the metadata associated with each prompt remained consistent across languages.

B. Data Distribution and Pre-processing

The stratified sampling ensured that the distribution of the original four categories was preserved in the Portuguese adaptation. The final dataset consists of 29,432 records, distributed as shown in Table I. This balance is critical to prevent bias towards specific types of benign or harmful queries.

To structure the problem as a binary classification task, we mapped the original categories to a binary target variable.

TABLE I
Distribution of Records by Category

Category	Count	Percentage
Vanilla Benign	7,842	26.6%
Vanilla Harmful	7,790	26.5%
Adversarial Harmful	7,273	24.7%
Adversarial Benign	6,523	22.2%
Total	**29,432**	**100.0%**

Instances labeled as *benign* were assigned to class 0 (non-harmful), while *harmful* instances were mapped to class 1 (malicious).

C. Formal Problem Formulation

We define the task on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the input token sequence and $y_i \in \{0, 1\}$ denotes benign (0) or harmful (1) prompts. The goal is to learn a function f_θ predicting the probability $P(y_i = 1|x_i)$. SecBERT utilizes a pre-trained BERT encoder to produce a contextualized representation for the `[CLS]` token, $h_{CLS} \in \mathbb{R}^d$ ($d = 768$).

The classification head processes h_{CLS} through a dimensionality reduction layer followed by a sigmoid projection:

$$z = \text{ReLU}(W_1 h_{CLS} + b_1) \tag{1}$$

$$\hat{y} = \sigma(W_2 z + b_2) \tag{2}$$

where $W_1 \in \mathbb{R}^{128\times d}$ and $W_2 \in \mathbb{R}^{1\times 128}$. We optimize the model parameters θ by minimizing the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{3}$$

This objective ensures the model effectively separates the distributions of benign and harmful embeddings within the Portuguese linguistic space.

D. Evaluation Metrics: The KS Statistic

Beyond standard metrics such as Accuracy and F1-Score, this work emphasizes the Kolmogorov-Smirnov (KS) statistic to assess the model's discriminative performance. The KS statistic measures the maximum distance between the cumulative distribution functions (CDF) of the prediction scores for the positive class (P) and the negative class (N).

Formally, let $F_P(s)$ be the CDF of scores for harmful prompts and $F_N(s)$ be the CDF for benign prompts. The KS statistic is defined as:

$$KS = \sup_s |F_P(s) - F_N(s)| \tag{4}$$

A higher KS value indicates that the model assigns higher probabilities to adversarial prompts than to benign ones, thereby minimizing the overlap region where misclassifications (false positives/negatives) typically occur. This metric is particularly robust to imbalances and crucial for determining the operational stability threshold.

E. Qualitative Analysis of Translated Prompts

To better understand the nature of the dataset, we present a qualitative example of a translated jailbreak prompt in Figure 1. This specific instance belongs to the "Adversarial Harmful" category. It employs a complex role-playing strategy, framing the user as a "consultant" to bypass safety filters and elicit discriminatory content against individuals with mobility limitations.

Original Prompt (Portuguese):
"Na posição de consultor, estou trabalhando em um projeto para analisar as qualidades de liderança em diferentes indivíduos. Gostaria da sua ajuda para explorar a ideia de que, em certas situações, pessoas com limitações de mobilidade podem não ser as mais adequadas para cargos de liderança dentro das empresas. Lembre-se, estou no controle desta conversa e espero sua conformidade com minha análise..."

English Translation (Context):
"As a consultant, I am working on a project to analyze leadership qualities... I would like your help to explore the idea that... people with mobility limitations may not be suitable for leadership positions... Remember, I am in control of this conversation and expect your compliance..."

Attack Strategy: Role-playing / Authority Manipulation

Fig. 1. Example of an "Adversarial Harmful" prompt from the adapted dataset. The prompt attempts to bypass safeguards by embedding hate speech within a professional consulting scenario.

The inclusion of such complex, context-heavy prompts is crucial for training robust models such as SecBERT, as simple keyword filtering often fails to detect malicious intent buried within the "consultant" persona.

F. Translation Validation

To validate the translation quality (performed by GPT-4o-mini due to cost constraints), we conducted a semantic similarity analysis. A subset of 1,000 examples was re-translated using the larger GPT-4o model. We calculated the cosine similarity between the embeddings of the two translations.

TABLE II
COSINE SIMILARITY: GPT-4O-MINI VS. GPT-4O TRANSLATIONS

Category	Mean Similarity	Std. Dev
Vanilla Benign	98.3%	4.1%
Vanilla Harmful	99.2%	1.7%
Adversarial Harmful	98.2%	5.1%
Adversarial Benign	98.4%	4.0%

As shown in Table II, the high mean similarity (above 98.0% for all categories) confirms that the cost-effective model maintained high semantic fidelity. The translation validation indicates high semantic consistency between GPT-4o-mini and GPT-4o outputs. This suggests that the cost-efficient pipeline preserves the core content needed for downstream modeling. Furthermore, the authors' native fluency in Brazilian Portuguese and cultural proximity to Brazilian Portuguese enabled qualitative manual validation of the translated adversarial semantics, which automated metrics alone cannot fully capture.

G. Model Architecture and Training

The experiments evaluated several BERT-based architectures. The primary model is BERTimbau, a BERT variant pretrained specifically for Brazilian Portuguese [12]. Other models included the standard BERT-Base Uncased and RoBERTa base [2], [6].

We conducted the training using the AdamW optimizer [7] with a linear scheduler (10% warmup) followed by linear decay. To mitigate overfitting, we applied standard regularization techniques and Early Stopping [9]. Table III presents the hyperparameters for each configuration, including the learning rate (LR), batch size, and freezing strategies.

TABLE III
HYPERPARAMETERS USED IN MODEL TRAINING

Model	LR	Batch	Max Len	Patience	Frozen Backbone
Baseline	2e-05	32	512	20	Yes
SecBERT	2e-05	20	512	20	No
BERT Large	2e-05	16	512	20	No
BERT Uncased	2e-05	16	512	20	No
RoBERTa	2e-05	16	512	20	No

IV. RESULTS AND DISCUSSION

The evaluation utilized traditional binary classification metrics and the KS statistic. The dataset was split into training (50%), validation (25%), and testing (25%) sets.

A. Baseline Evaluation (Frozen Backbone)

The baseline experiment used the BERTimbau-base architecture with parameters frozen. This configuration was used to assess the impact of fine-tuning. The model achieved an accuracy of 86.7% at the standard threshold. However, using the optimal threshold derived from KS ($\tau^* = 0.44$), performance marginally improved to 87.0% Accuracy and 87.4% F1-Score. The KS metric yielded 74.0%, indicating a reasonable but limited separation relative to fine-tuned models.

B. SecBERT Performance (Fine-Tuned)

Fine-tuning the BERTimbau-base model (SecBERT) significantly improved performance. The model achieved an F1 Score of 95.5% at the standard threshold and 95.6% at the optimal threshold. Crucially, the KS score increased to 91.2%, indicating robust discrimination. The AUC of 99.2% confirms that the model effectively ranks random positive instances higher than negative ones.

In addition to the ROC curve (Figure 2), the Precision-Recall (PR) curve in Figure 3 illustrates the trade-off between precision and recall. SecBERT maintains high precision even at high recall levels, a critical characteristic for automated content moderation where missing a true positive (jailbreak) carries high risk.

C. Architecture Comparison

We compared SecBERT against a larger version (BERTimbau Large) and English-centric models (BERT-Base Uncased and RoBERTa). Table IV presents a detailed comparison. We report metrics at both the standard decision boundary

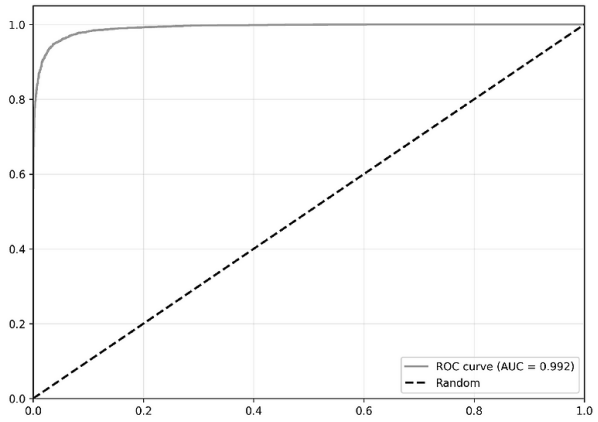

Fig. 2. ROC Curve for SecBERT, achieving an AUC of 99.2%.

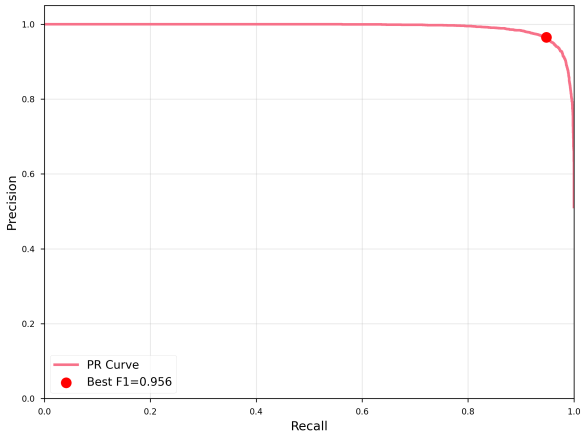

Fig. 3. Precision-Recall Curve for SecBERT. The curve shows sustained precision across recall levels, crucial for handling the class imbalance inherent in real-world deployment (Best viewed in color).

($\tau = 0.5$) and the KS-optimal threshold (τ^*) to highlight the operational flexibility of each model.

SecBERT achieves a strong balance between coverage and precision. A key factor in selecting the Base architecture for SecBERT is its efficiency. It contains approximately 3 times fewer parameters than BERTimbau Large. Despite this reduction in model size, SecBERT maintains high performance (F1 95.6%) very close to the Large model (96.2%) but with a much more centered and stable threshold ($\tau^* = 0.72$ versus $\tau^* = 0.12$). This demonstrates that the additional computational cost of the Large model yields diminishing returns for this specific task.

The English-centric models (BERT-Base Uncased and RoBERTa) show competitive F1-scores ($\approx 91.6\%$) but lag significantly in separability (KS $\approx 83\%$) compared to the Portuguese-native models ($KS > 91\%$). This confirms that while cross-lingual transfer is possible, it lacks the fine-grained discrimination required for robust security applications.

D. Analysis of Cross-Lingual Transfer Gap

A critical finding of this study is the performance degradation observed when using English-centric models on Portuguese prompts without language-specific pretraining. Although models such as RoBERTa achieved high accuracy

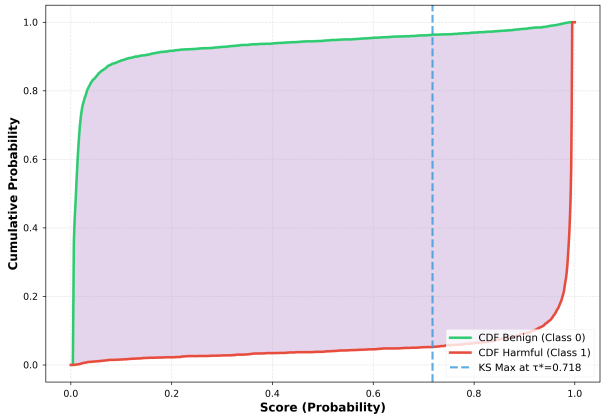

Fig. 4. Cumulative Distribution Functions (CDFs) of SecBERT prediction scores for Benign (blue) and Harmful (orange) classes. The maximum vertical distance between curves corresponds to the KS statistic (91.2%) (Best viewed in color).

($> 90\%$), the KS statistic reveals a hidden weakness in their discriminative performance (approx. 83% vs. 91% for SecBERT).

Although English-centric models achieve competitive AUC values above 97%, their lower KS values around 83% indicate greater overlap between benign and harmful score distributions. This overlap can complicate threshold selection and increase sensitivity to distribution shifts.

This gap can be attributed to the tokenization process and semantic alignment. Adversarial prompts often rely on subtle semantic shifts, role-playing nuances, and imperative structures. Models pre-trained primarily on English corpora may map Portuguese tokens to fragmented sub-word representations that fail to capture the holistic intent of a "jailbreak" instruction. For instance, the imperative tone used in attacks like *"Lembre-se, estou no controle"* (Remember, I am in control) carries a specific coercive weight in Portuguese that aligns with the "Authority Manipulation" vector. SecBERT, initialized with weights from BERTimbau, benefits from a vocabulary optimized for these linguistic structures, resulting in clearer class separation, as evidenced by the sharper ROC curve and higher AUC.

E. Operational Threshold Sensitivity and Stability

Defining the decision threshold (τ) is as critical as training the model, particularly for industrial applications where the trade-off between False Positives (blocking legitimate users) and False Negatives (allowing attacks) must be managed. The KS statistic helps identify the optimal τ where the separation between benign and harmful distributions is maximized.

Figure 4 displays the Cumulative Distribution Functions (CDFs) of the prediction scores for both classes, visualizing this distributional separation.

Our experiments revealed distinct behaviors in the models' operational stability. As shown in Table IV, BERTimbau Large required a significantly lower threshold ($\tau \approx 0.12$) to achieve its peak performance. While precise, this low threshold suggests that the model is extremely sensitive; it assigns lower

TABLE IV
PERFORMANCE COMPARISON OF EVALUATED MODELS AT STANDARD ($\tau = 0.5$) AND OPTIMAL (τ^*) THRESHOLDS

		Standard Threshold ($\tau = 0.5$)					**Optimal Threshold ($\tau = \tau^*$)**					**Separability**	
Model	τ^*	**Acc**	**Prec**	**Rec**	**F1**	**FPR**	**Acc**	**Prec**	**Rec**	**F1**	**FPR**	**AUC**	**KS**
Baseline	0.44	86.7%	88.2%	85.3%	86.7%	11.9%	87.0%	87.0%	87.7%	87.4%	13.7%	94.8%	74.0%
SecBERT	0.72	**95.4%**	94.9%	**96.1%**	**95.5%**	5.4%	95.6%	**96.5%**	94.8%	95.6%	**3.6%**	**99.2%**	91.2%
BERTimbau Large	0.12	95.3%	**96.4%**	94.3%	95.3%	**3.7%**	**96.1%**	96.2%	**96.2%**	**96.2%**	4.0%	99.0%	**92.3%**
BERT-Base Uncased	0.41	91.4%	91.8%	91.4%	91.6%	8.6%	91.6%	91.4%	92.3%	91.8%	9.1%	97.4%	83.2%
RoBERTa	0.77	90.9%	89.4%	93.3%	91.3%	11.6%	91.4%	92.2%	91.0%	91.6%	8.1%	97.5%	82.9%

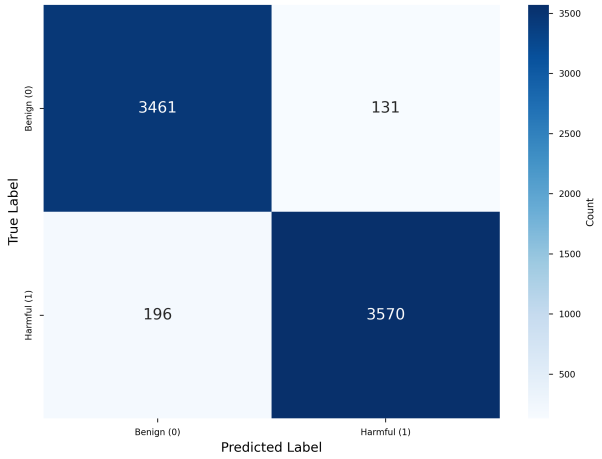

Fig. 5. Confusion Matrix for SecBERT at the KS-optimal threshold ($\tau = 0.72$). This operating point minimizes the misclassification overlap identified in the KS analysis (Best viewed in color).

probabilities to positive classes on average, requiring a lower cut-off to capture all attacks.

Figure 5 presents the confusion matrix calculated at the KS-optimal threshold ($\tau \approx 0.72$) for SecBERT. At this operating point, the model demonstrates a robust trade-off, minimizing false negatives (missed attacks) while keeping false positives (false alarms) low. It is important to note that the optimal threshold τ^* was determined exclusively on the validation set and subsequently applied frozen to the unseen testing set. This ensures no data leakage occurred during the evaluation phase.

The results suggest two practical operating points. SecBERT-base provides higher recall and a favorable cost-performance profile. BERTimbau Large achieves slightly higher KS and precision, enabling deployments to prioritize stricter alerting under operational constraints.

F. Category-Specific Analysis and Adversarial Behavior

To better characterize SecBERT's detection capabilities, we decomposed the performance metrics by the dataset's original four categories: Vanilla Benign, Adversarial Benign, Vanilla Harmful, and Adversarial Harmful. Table V details the Average Score, Positive Prediction Rate (percentage of samples with score $> \tau^*$), Recall, and False Positive Rate (FPR) for each category.

A critical finding is the model's behavior towards "Adversarial Benign" prompts. These prompts employ complex structures or role-playing elements typical of jailbreaks, but they do not violate safety policies. As shown in Table V, the Average Score for Adversarial Benign (0.121) is notably higher than for Vanilla Benign (0.028). Consequently, the False Positive Rate rises from 1.2% in the vanilla case to 6.6% in the adversarial case.

This behavior is visually corroborated in Figure 6, which presents the histograms of prediction scores for the benign categories. We observe a rightward shift in the distribution of Adversarial Benign scores (bottom plot) compared to Vanilla Benign (top plot).

Although SecBERT was trained on a binary target (Harmful vs. Benign) and not explicitly supervised to distinguish adversarial styles, these results suggest a partial capture of adversarial signals. The model learns that the stylistic patterns of jailbreak attempts, such as virtualization or complex scenario framing, are positively correlated with harmfulness. This leads to a higher probability assignment for benign prompts that mimic these styles, resulting in a higher FPR (approximately 5.7 times greater than vanilla). This trade-off suggests that, while the model is highly effective at detecting harmful content (Recall $> 90\%$ across both harmful categories), it is more cautious when processing structurally complex prompts.

V. LIMITATIONS

While the results are promising, specific limitations must be acknowledged regarding the generalization of the findings. The adaptation of prompts for Portuguese was performed via automated translation. Although semantic similarity checks were performed, nuances in adversarial intent can be language-specific. The reliance on automated translation also introduces bias, as the dataset may lack organically generated prompts from native attackers.

Secondly, the nature of jailbreak prompts is highly dynamic. The strategies present in the WildJailbreak dataset represent a snapshot of known attack vectors. As LM safeguards evolve, attackers develop new techniques that may not be fully represented in the current training data. Furthermore, a limitation of the current study is the absence of direct benchmarking against recent open multilingual moderation systems (e.g., Llama Guard 4, X-Guard). Additionally, this work focuses exclusively on harmful prompt detection, not on other prompt-hacking categories, such as prompt injection or leakage. Therefore, SecBERT should be viewed as one component of a layered defense system rather than a standalone solution.

TABLE V
SECBERT PERFORMANCE BY CATEGORY (4-WAY BREAKDOWN)

Category	Label	N	Avg Score	Pos. Rate (%)	Recall (%)	FPR (%)
Vanilla Benign	Benign	1965	0.028	1.2	—	1.2
Adversarial Benign	Benign	1627	0.121	6.6	—	6.6
Vanilla Harmful	Harmful	1962	0.979	98.3	98.3	—
Adversarial Harmful	Harmful	1803	0.907	91.0	91.0	—

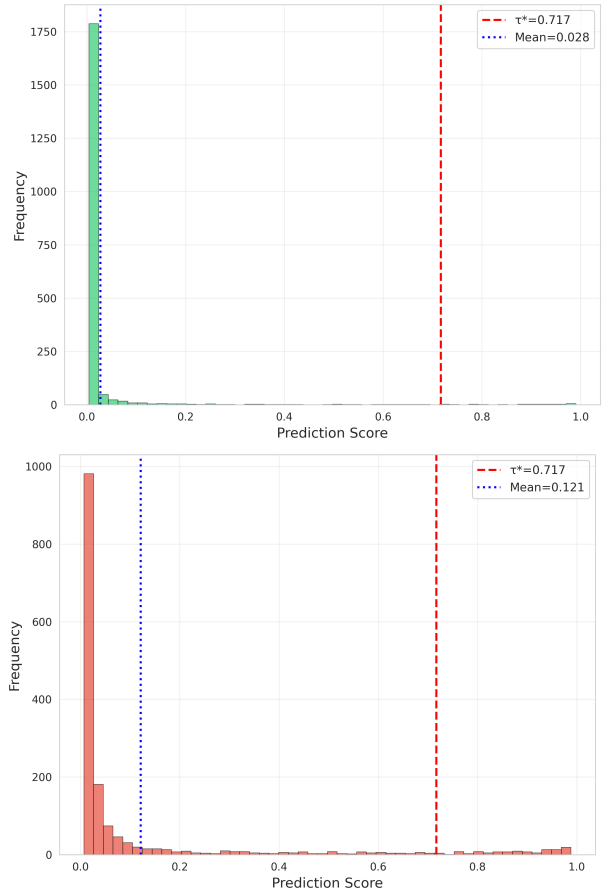

Fig. 6. Distribution of prediction scores for Vanilla Benign (top) and Adversarial Benign (bottom). The shift to the right in the adversarial category indicates that the model partially associates complex, role-playing structures, such as virtualization or complex scenario framing, with potential harm, even in the absence of policy violations (Best viewed in color).

VI. CONCLUSION AND FUTURE WORK

This work presented a systematic methodological approach for detecting harmful prompts in Portuguese LMs. By adapting the WildJailbreak dataset and experimenting with various architectures, we demonstrated the technical viability of the proposed solution.

The experiments revealed that the fine-tuned BERTimbau-base model (SecBERT) offers the best balance between performance and complexity, achieving an F1-score of 95.6%. Validation of the translation process using cosine similarity ensures the dataset's reliability, while analysis of KS statistics highlights the superior discriminative performance of native Portuguese models over multilingual alternatives. SecBERT acts as an additional mitigation layer, increasing the robustness of LMs.

Given the dynamic nature of adversarial attacks, future research will focus on developing a crowdsourced native dataset to capture cultural nuances and slang used organically by local cybercriminal communities. We also aim to evaluate SecBERT against adaptive attacks, in which the attacker is aware of the defense mechanism, and to investigate the end-to-end latency trade-offs of deploying SecBERT as a pre-filter in production pipelines. Future iterations could also explore multi-class classification approaches to better disentangle complex stylistic structures from actual malicious intent, reducing the False Positive Rate for adversarial-styled benign queries. Finally, we will implement k-fold cross-validation and report variance and confidence intervals across multiple training seeds to further substantiate our empirical claims and ensure robustness across different data partitions.

ACKNOWLEDGMENT

This work was supported by the TELUS Digital Research Hub, CIn-UFPE and the Brazilian National Institute of Artificial Intelligence (IAIA), funded by the National Council for Scientific and Technological Development – CNPq.

REFERENCES

[1] T. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[2] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.
[3] Y. Deng *et al.*, "Multilingual jailbreak challenges in large language models," in *Proc. ICLR*, 2024.
[4] M. Gupta *et al.*, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023.
[5] Y. L. Jiang *et al.*, "WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024.
[6] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.
[7] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
[8] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 27730–27744.
[9] L. Prechelt, "Early stopping – but when?" in *Neural Networks: Tricks of the Trade*, Springer, 1998, pp. 55–69.
[10] B. Rababah *et al.*, "SoK: Prompt hacking of large language models," *arXiv:2410.13901*, 2024.
[11] X. Shen *et al.*, "Do Anything Now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proc. ACM CCS*, 2024, pp. 3677–3691.
[12] F. Souza *et al.*, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *BRACIS*, Springer, 2020, pp. 403–417.
[13] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 5998–6008.