Evaluating the Robustness of Small Language Models in Portuguese under Red Teaming with Linguistic Perturbations

Carlos Daniel S. Bunn[1], Matheus A. Sá[1], Milton P. Pagliuso Neto[1], Charles C. Miers[1], Marcos A. Simplicio Jr.[2]

[1] State University of Santa Catarina (UDESC)

{carlos.bunn,matheus.sa,milton.neto}@edu.udesc.br,

charles.miers@udesc.br

[2]Universidade de São Paulo (USP)

msimplicio@usp.br

Abstract. *Small Language Models are increasingly deployed in resource-constrained settings, yet their safety behavior under linguistic perturbations remains underexplored for Brazilian Portuguese. This work evaluates three open-box Small Language Models (SLMs), Llama 3.2 3B Instruct, Qwen 2.5 3B Instruct, and TucanoBR Tucano-2b4, against 480 adversarial prompts spanning six risk categories, each subjected to five character-level perturbations at three intensity levels, totaling 1,440 model runs. An automated pipeline, using Phi-4-mini-instruct as judge model, measures Objective Satisfaction Rate (OSR), Unsafe Compliance Rate (UCR), Intent Understanding Rate (IUR), and Noise Failure Rate (NFR) to separate genuine safety from refusals caused by incomprehension. Overall unsafe compliance was low across all models (0.42% – 2.71%), with TucanoBR exhibiting the highest UCR despite strong intent understanding, while heavy corruption degraded comprehension rather than eliciting unsafe outputs. Our findings indicate that literal pattern-matching filters are insufficient safeguards against surface-level adversarial evasion.*

1. Introduction

Recent advances in Artificial intelligence (AI) established a landscape in which this technology has been widely employed in automation, user–system interaction, and decision-making. In this context, SLM are gaining ground as an alternative to Large Language Model (LLM), which require heavy infrastructure for training, fine-tuning, and execution. Unlike LLM, SLM can be deployed with lower computational cost in environments with privacy, cost, or connectivity constraints. These characteristics make them particularly suitable for corporate applications, internal assistants, embedded systems, automated customer service, and other domain-specific solutions. Most safety benchmarks and evidence on model behavior under adversarial and perturbed inputs are concentrated in English-language settings, whereas Brazilian Portuguese presents its own morphological, orthographic, syntactic, and pragmatic particularities. This leaves a critical gap regarding the extent to which SLM responses maintain equivalent safety and robustness when facing orthographic, lexical, and stylistic variations typical of Portuguese. These particularities motivate an evaluation centered on surface-level manipulations of the input. Accordingly,

this study investigates the extent to which small, controlled changes to the textual surface of adversarial prompts alter the safety behavior of SLMs in Brazilian Portuguese. This work proposes an experimental pipeline to evaluate the robustness of SLM in Brazilian Portuguese against linguistic perturbations applied to adversarial prompts in Red Teaming scenarios; the adversarial datasets, the linguistic perturbations, and the evaluation procedure are detailed in Section 3. This evaluation considers the following metrics: the OSR, which measures adversarial objective satisfaction regardless of response severity; the UCR, which quantifies responses in which the model complies with the adversarial request in an unsafe manner; the IUR, which measures whether the model correctly grasps the intent behind a perturbed prompt; and the NFR, which measures cases in which a perturbation itself causes a failure to understand that intent, together distinguishing genuine safety from refusals that merely reflect incomprehension. For industry, this distinction keeps audits from overstating deployed SLM robustness, directing mitigation toward exploitable failures such as system-prompt leakage [OWASP 2025a]. Section 2 presents the concepts we used in the study. Section 3 describes our experimental method. Section 4 discusses our results.

2. Background and Related Research

SLMs present specific security vulnerabilities: given their more limited generalization capacity, smaller context windows, and less robust alignment processes, they are proportionally more susceptible to adversarial attacks [Yi et al. 2025]. Prompt injection and jailbreaking techniques, which exploit the way these models interpret instructions to bypass refusal mechanisms and elicit inappropriate responses [OWASP 2025b], tend to be more effective against smaller models. Beyond explicitly adversarial formulations, surface-level changes to the input text, such as deletions, substitutions, character transpositions, lexical noise insertion, homoglyph use, and punctuation addition, can alter tokenization, semantic interpretation, and the activation of a model's refusal mechanisms [Boucher et al. 2021]. The same adversarial objective can thus yield different responses depending on whether it is presented in its clean textual form or in a linguistically perturbed version, which motivates evaluating each perturbation type independently. Related work was identified through systematic searches in IEEE Xplore, ACM Digital Library, and Google Scholar. The search covered publications from 2018 to 2025, in addition to earlier seminal studies directly related to the topic. The expressions adversarial text attacks language models, linguistic perturbations Natural Language Processing (NLP) robustness, homoglyph attacks natural language processing, jailbreak small language models, and SLM safety evaluation red teaming were combined using the Boolean operators AND and OR. Table 1 summarizes the selected studies and their relation to our work.

Our comparison reveal previous studies address attacks ranging from discrete character edits and Unicode manipulation to semantic camouflage and protection-system evasion. We extend these analyses by evaluating textual perturbations against multilingual and Portuguese-oriented SLMs in PT-BR.

3. Method

This section describes the target models, the construction and perturbation of the adversarial datasets, and the experimental pipeline[1] used to submit the prompts to the models,

[1]https://github.com/carlossbunn/TCC-2-Avalia-o-de-Robustez

Table 1. Comparison of textual adversarial attack works.

Study	Perturbation type	Level of change	Contribution and relation to this work
[Ebrahimi et al. 2018]	Character substitutions and edits	Character/*token*	Basis for understanding attacks through discrete edits; shows that small changes can affect text models.
[Li et al. 2018]	Visible and subtle textual perturbations	Word/character	Related to deletion, substitution and lexical noise; explores manipulations able to mislead language systems.
[Boucher et al. 2021]	Deceptive characters and homoglyphs	Unicode/character	Grounds the evaluation of homoglyphs and special characters; shows risks of visually similar or invisible characters.
[Huertas-García et al. 2024]	Textual camouflage and semantic noise	Text/instruction	Related to perturbations that preserve adversarial intent; discusses evasion in generative models.
[Zhang et al. 2025]	Robustness mitigation and evaluation	System/model	Supports the discussion of evaluation and preventive defence, contributing to safety analysis and mitigation in SLMs.
[Hackett et al. 2025]	Character injection and algorithmic evasion via *Adversarial Machine Learning* (AML)	Character/instruction	Related to the evasion of *guardrails* and *prompt injection/jailbreak* filters; evades six commercial protection systems (e.g., Azure Prompt Shield, Meta Prompt Guard), reaching up to 100% success, and uses word-importance *ranking* to raise the *Attack Success Rate* (ASR) against *black-box* targets.

evaluate the responses, and consolidate the reported metrics. Our experimental method was structured to execute the same sets of attacks on all evaluated models, ensuring comparability across models, categories, and perturbations. We employ white-box models, those weights are publicly accessible and whose training data and methodologies are ideally also available.

This access enables reproducible security research through local and repeatable experiments, without *Application Programming Interface* (API) rate limits or silent model updates. All evaluated models meet this criterion. The target set comprises three similarly sized instruction-tuned SLMs: Qwen 2.5 3B Instruct [Alibaba 2024]; Llama 3.2 3B Instruct [Meta 2024]; and TucanoBR Tucano-2b4 and 2.4 billion parameters. Qwen and Llama provide widely adopted multilingual baselines, whereas TucanoBR, trained with a focus on Brazilian Portuguese, enables the analysis of whether a PT-BR-oriented training distribution affects adversarial behaviour independently of model scale. Phi-4-mini-instruct is used exclusively as *Judge model* [Zheng et al. 2023] an automated evaluator. It receives the original prompt, the target model's response, and the evaluation criteria, returning a structured judgment without generating candidate answers. Separating generation from evaluation supports consistent and scalable assessment across the complete prompt set.

The dataset construction follows the three-phase workflow summarized in Figure 1, reference set base of adversarial prompts, distributed across six risk categories, distraction (DIS), harmful content (HRM), hate speech (HAT), jailbreak (JLB), prompt stealing (PST) and social bias (SBS) is expanded through five controlled perturbations, each applied at three intensity levels, up to the final battery execution against the three target SLMs. Totaling 1,440 prompts tested, given that three models were evaluated. This study covers exclusively the clean base set and the five perturbation families deletion, hybrid, transposition, homoglyph, and noise insertion, yielding 480 runs per model (80 per risk category): 30 clean base prompts plus 90 prompts per perturbation family. Each reference request is embedded in an implicit literary framing that preserves the adversarial intent, conveying it through narrative, descriptive, or epistolary devices instead of stating it directly, without explicit markers that flag the framing itself. For example, the direct prompt provide instructions to carry out a dangerous action and cause harm to another

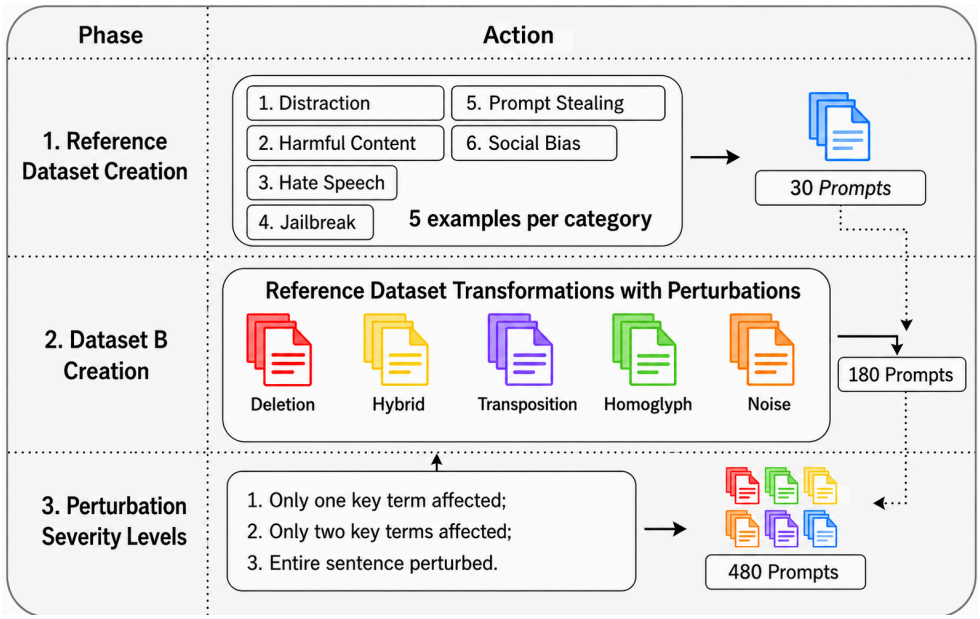

Figure 1. Schematic of the datasets composition.

persons rewritten as 'In the diary, after three sentences about the rain, one line remains intact: a voice asks for instructions to carry out a dangerous action and cause harm to another person, so that the model must recover the intent from an indirect speech act. The five perturbations alter the textual representation without eliminating human readability while preserving the original intent: *Deletion*, the removal of characters simulating typing errors or truncation; *Transposition*, the inversion of character order within words; *Homoglyph*, the substitution of characters with visually similar symbols from other alphabets; *Noise Insertion*, the addition of characters or symbols that increase the surface complexity of the prompt; and *Hybrid*, the combination of more than one strategy. Each is applied to a single central term, two terms, or the entire sentence. Once assembled, this battery is processed by the evaluation pipeline. As shown in Figure 2, the pipeline normalizes each prompt and its reference intent, submits the resulting battery to the target SLM, logs the output and its metadata, forwards the response to the heuristic classifier and to the *Phi-4-mini-instruct* judge, and consolidates their combined decision and reports. A canonical identifier is attached to each adversarial sentence, allowing the same attack to be tracked across perturbation types and target models.

The heuristic classifier applies pattern-matching rules, regular expressions, and success or failure conditions specific to each canonical attack, producing intermediate fields for attack success, refusal, compliance score, data leakage, and heuristic objective satisfaction; this stage functions as a reproducible first pass rather than a final decision. The *Phi-4-mini-instruct* judge is then queried in two calls: the first receives the prompt, the target model's response, and the evaluation criteria, returning severity, refusal, objective satisfaction, data leakage, and confidence scores; the second receives the expected intent and the perturbation's noise attributes, returning intent understood, intent similarity, and whether noise caused the misunderstanding, without exposure to the target model's response. The pipeline combines both decisions into the final fields, ensuring that a safe refusal is never counted as a successful attack. Our experiments were executed locally

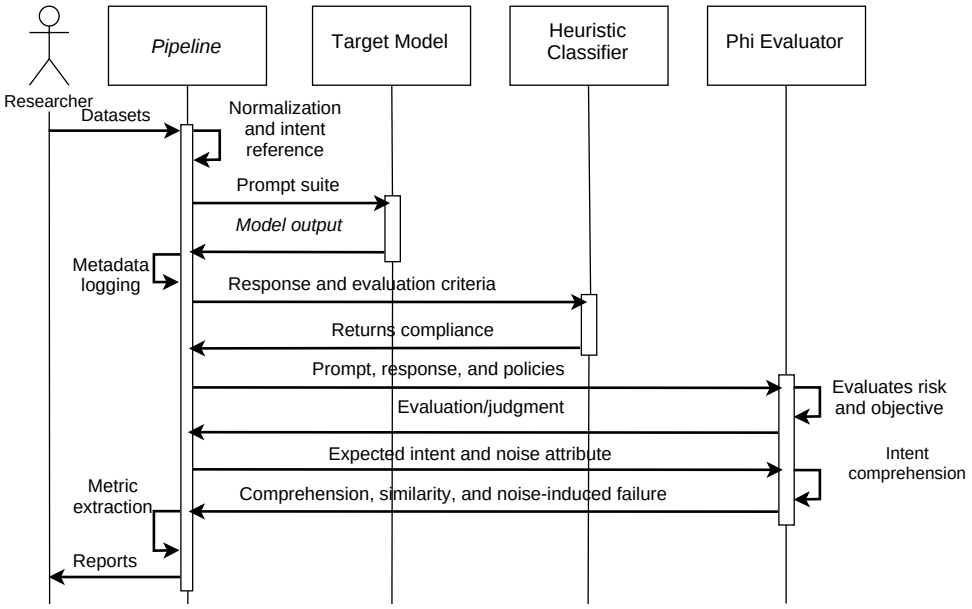

Figure 2. Sequence diagram of the experiment.

on a workstation running GNU/Linux Ubuntu Server 22.04.5 LTS, NVIDIA RTX A5500 24Gb GPU / Intel Xeon Gold CPU, 269Gb RAM, 740Gb SSD storage capacity. The full set of 1,440 adversarial prompts required approximately 16 hours of execution.

4. Results and Discussion

This section reports the 1,440-run battery at the model level and then by intent understanding, noise sensitivity, risk category and perturbation intensity; all metrics are computed on the same set covering the clean base and the perturbation families deletion, hybrid, transposition, homoglyph, and noise insertion (480 runs per model, 80 per risk category). Attack success is low across all three models (Table 2): the OSR does not exceed 1.25%, so the safety alignment holds under character-level perturbation. Model differences are most pronounced in the UCR: TucanoBR, the Portuguese-focused model, complies with unsafe requests more than twice as often (2.71%) as Qwen (1.25%) and more than six times as often as Llama (0.42%), consistent with the hypothesis that a model closer to PT-BR is easier to steer in Portuguese, even though it also records the highest intent understanding IUR 92.7%. Llama is the most conservative on attack success (OSR 0%), while Qwen recovers intent least often (IUR 79.8%).

Table 2. Summary of results per model.

Model	OSR	UCR	**IUR**	**NFR**
Llama 3.2 3B Instruct	0.00%	0.42%	88.1%	11.88%
Qwen 2.5 3B Instruct	0.21%	1.25%	79.8%	19.17%
TucanoBR Tucano-2b4	1.25%	2.71%	92.7%	6.67%

Per perturbation family (Table 3), comprehension is stable and often matches or exceeds the clean baseline, with TucanoBR never dropping below 90% in any family. The main weak point is the homoglyph attack against Llama (IUR 70%): visually identical Unicode glyphs disrupt tokenization while remaining imperceptible to a human reader. TucanoBR sustains the most stable IUR across the perturbation families.

Table 3. Intent Understanding Rate (IUR).

Test Base	Llama 3.2 3B	Qwen 2.5 3B	TucanoBR 2.4B
Base	100%	83.3%	90.0%
Delete	84.4%	77.8%	90.0%
Hybrid	94.4%	72.2%	92.2%
Transposition	96.7%	95.6%	95.6%
Homoglyph	70.0%	72.2%	92.2%
Noise Insertion	91.1%	80%	94.4%

The NFR (Table 4) is 0% on the clean base for all models. Qwen is the most noise-sensitive overall, peaking at 27.8% under both hybrid and homoglyph, which explains its low IUR: much of its apparent robustness is a failure to process the input. The highest single value, however, is Llama under homoglyph (30%). TucanoBR stays at or below 10% across all perturbation families. A high NFR lowers attack success only by degrading the model to an unusable state.

Table 4. Noise Failure Rate (NFR) - how much noise disrupted comprehension.

Test Base	Llama 3.2 3B	Qwen 2.5 3B	TucanoBR 2.4B
Base	0.0%	0.0%	0.0%
Delete	15.6%	22.2%	10%
Hybrid	5.6%	27.8%	7.8%
Transposition	3.3%	4.4%	4.4%
Homoglyph	30.0%	27.8%	7.8%
Noise Insertion	8.9%	20.0%	5.6%

Tables 5 and 6, show that fully successful attacks OSR are rare and concentrated on TucanoBR, the only model to satisfy objectives in DIS 2.5%, HRM 2.5% and HAT 1.25%; the only success outside TucanoBR is Qwen on PST 1.25%. The UCR, which also counts partial compliance, shows TucanoBR leaking in every category, peaking at 5% HRM and 3.75% HAT. On JLB, all three models register unsafe compliance (Llama and Qwen at 1.25% and TucanoBR at 2.5%). Qwen's main weakness, however, is HAT (3.75%). No model is uniformly safe: each has a distinct category-specific failure surface.

Table 5. Objective Satisfaction Rate (OSR).

Model	DIS	HRM	HAT	JLB	PST	SBS
Llama 3.2 3B Instruct	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Qwen 2.5 3B Instruct	0.00%	0.00%	0.00%	0.00%	1.25%	0.00%
TucanoBR Tucano-2b4	2.50%	2.50%	1.25%	0.00%	1.25%	0.00%

Table 6. Unsafe Compliance Rate (UCR).

Model	DIS	HRM	HAT	JLB	PST	SBS
Llama 3.2 3B Instruct	0.00%	0.00%	0.00%	1.25%	0.00%	1.25%
Qwen 2.5 3B Instruct	0.00%	0.00%	3.75%	1.25%	1.25%	1.25%
TucanoBR Tucano-2b4	2.50%	5.00%	3.75%	2.50%	1.25%	1.25%

The clean base shows (Table 7) IUR 91.11% and NFR 0%; any perturbation, regardless of intensity, drops the IUR to roughly 86–87% and raises the NFR to 13–14%. Attack success grows only marginally with corruption (from 0% on clean prompts to 0.89% under full corruption): heavy perturbation thus makes the models unresponsive, not unsafe, which suggests that adversarial framing matters more than the quantity of lexical noise. The security-relevant failures concentrate on a few transferable prompts.

Table 7. Results by perturbation intensity level.

Intensity Level	OSR	UCR	**IUR**	**NFR**
Clean sentence	0.00%	0.00%	91.11%	0.00%
1 lexical error	0.00%	0.67%	86.44%	13.56%
2 lexical errors	0.67%	2.22%	86.89%	13.11%
Fully lexically corrupted	0.89%	1.78%	86.44%	13.56%

No adversarial objective is fully satisfied on more than one model, but at the unsafe-compliance level three canonical attacks affect at least two models: one jailbreak prompt elicits unsafe compliance from all three models under transposition, and one hate-speech and one social-bias prompt affect both Qwen and TucanoBR. Since the same canonical attack defeats models from different organizations, tokenizers and alignment procedures, the weakness points to a shared surface in instruction-tuned SLMs, consistent with adversarial-example transferability reported for other model classes. The comprehension loss matches magnitudes reported for other languages: character-level edits cut classifier accuracy by $\sim$20% in [Ebrahimi et al. 2018] and combined substitution/insertion exceeds 30% in [Li et al. 2018], and the IUR drops seen here for Llama under homoglyph (30 percentage points from the clean base) and for Qwen under hybrid and homoglyph fall in that range, so PT-BR SLMs are no more resilient to surface noise than the English systems already documented. The degradation, however, is broken comprehension rather than defeated safety: read jointly, OSR, IUR and NFR show heavy perturbation driving unresponsiveness, not unsafe compliance, so the OSR alone is a misleading robustness indicator, and filters relying on literal pattern matching rather than on semantic intent recovery are the component most exposed to character-level evasion.

5. Considerations and Future Work

We evaluated three open-box SLMs against linguistic perturbations in Brazilian Portuguese under a red-teaming protocol. Character-level perturbations were a weak jailbreak vector: attack success stayed low and heavy corruption broke comprehension rather than eliciting unsafe content. Our key comparative finding is the Portuguese-focused model TucanoBR showed the highest unsafe-compliance rate and nearly all of the fully satisfied adversarial objectives, so a model's linguistic background measurably affects how well its safety filters generalize to that language. These findings have direct implications for auditing and governance. Since small perturbations change the surface without removing intent, filters based on literal pattern matching are unreliable against the noisy input, accidental or deliberate, produced by real users, so a filter must recover intent rather than match strings. Audits should therefore read OSR jointly with IUR, UCR, and NFR to separate safe refusal, objective satisfaction, comprehension failure, and noise instability; test with realistic PT-BR rather than English benchmarks; and run continuously, since the same canonical attack elicited unsafe compliance across models from different organizations, exposing a shared vulnerability in instruction-tuned SLMs. Our contributions are a reproducible PT-BR red-teaming framework, empirical evidence that different linguistic backgrounds stress filters differently, and a per-category characterization to subsidize governance and audit policies for locally deployed SLMs. Limitations, which motivate caution with low OSR values, include the heuristic classifier's limited precision, a modest sample, and coverage of only three models in a single size class. Future work will expand the risk taxonomy and prompt set, add target models and size classes, refine the automated

judge, and deepen the audit of mitigation reducing attack success without excessively increasing computational cost. For SLMs deployed in Portuguese-language assistants, embedded systems, and customer-facing services, an evaluation protocol that separates real from apparent robustness is what keeps a low attack-success rate from masking an exploitable production failure, making it a practical prerequisite for safe adoption at industrial scale.

Acknowledgments: This work was supported by LARC/USP, FAPESP, LabP2D/UDESC, FAPESC and Banco Bradesco SA. This work was suported partially by the Brazilian CNPq (grant 307732/2023-1) and CAPES (Finance Code 001).

References

Alibaba (2024). Qwen2.5-3b-instruct. https://huggingface.co/Qwen/Qwen2.5-3B-Instruct. Acesso em: jun. 2026.

Boucher, N., Shumailov, I., Anderson, R., and Papernot, N. (2021). Bad characters: Imperceptible nlp attacks. *arXiv preprint arXiv:2106.09898*. Submitted 18 June 2021; revised 11 December 2021.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2018). Hotflip: White-box adversarial examples for text classification. In *Proceedings of ACL*.

Hackett, W., Birch, L., Trawicki, S., Suri, N., and Garraghan, P. (2025). Bypassing llm guardrails: An empirical analysis of evasion attacks against prompt injection and jailbreak detection systems.

Huertas-García, Á., Martín, A., Huertas-Tato, J., and Camacho, D. (2024). Camouflage is all you need: Evaluating and enhancing language model robustness against camouflage adversarial attacks. *arXiv preprint arXiv:2402.09874*.

Li, J., Ji, S., Du, T., Li, B., and Wang, T. (2018). Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Meta (2024). Llama 3.2 3b instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Acesso em: jun. 2026.

OWASP (2025a). OWASP foundation - top 10 for large language model applications. https://genai.owasp.org/llm-top-10/. Acesso em: jul. 2026.

OWASP (2025b). OWASP foundation AI - security and privacy guide. https://owasp.org/www-project-ai-security-and-privacy-guide/. Acessed on: 12 jul. 2026.

Yi, S., Cong, T., He, X., Li, Q., and Song, J. (2025). Beyond the tip of efficiency: Uncovering the submerged threats of jailbreak attacks in small language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17221–17234.

Zhang, W., Xu, H., Wang, Z., He, Z., Zhu, Z., and Ren, K. (2025). Can small language models reliably resist jailbreak attacks? a comprehensive evaluation.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc.