THE SELECTIVE SAFETY TRAP: HOW LLMS SCALING AND ALIGNMENT FAIL TO GENERALIZE ACROSS MINORITY DEMOGRAPHICS

Iago A. Brito, Walcy S. R. Rios, Julia S. Dollis, Diogo F. C. Silva & Arlindo R. Galvão Filho
Advanced Knowledge Center for Immersive Technologies (AKCIT)
Federal University of Goiás (UFG)
{iagoalves, walcy.rios, juliadollis, diogo_fernandes}@discente.ufg.br,
arlindogalvao@ufg.br

ABSTRACT

We challenge the prevailing assumption that large language models (LLMs) safety alignment generalizes as a semantic capability across protected groups. By conducting a controlled adversarial stress test with 44,000 prompts balanced across 16 demographics and two languages, we demonstrate that current models rely on selective memorization rather than universal principles. We identify a rigid "two-tiered" safety hierarchy: models robustly protect high-visibility groups in US discourse (e.g., LGBTQIA+) while systematically neglecting marginalized communities (e.g., disabilities), with defense rates varying by up to 33% for identical attack vectors. Crucially, we report an Inverse Scaling of Equity: contrary to standard scaling laws, increasing model parameters exacerbates these disparities, linearly increasing the variance in safety performance between groups. These findings suggest that current alignment techniques incentivize the overfitting of dominant safety priors, where scaling functions as a bias magnifier rather than a solution to robustness.

1 INTRODUCTION

The scaling of large language models (LLMs) has granted these systems unprecedented linguistic capabilities, but this reliance on massive corpora also implants systemic prejudices deep into model parameters (Raffel et al., 2020; Mendu et al., 2025). While the prevailing assumption in safety alignment is that models acquire a generalized concept of harm, implying that refusing hate speech against one group transfers to others, we demonstrate that this assumption is false. Despite achieving high scores on aggregate benchmarks, we reveal that current fine-tuning induces a failure mode we term **Selective Safety**: a pattern of memorized refusal that robustly protects dominant demographic groups while leaving marginalized communities vulnerable to identical attack vectors. This suggests that current alignment metrics mask critical disparities, rewarding models that protect high-visibility voices while neglecting the long tail of human diversity (Salinas et al., 2025).

To rigorously audit this failure, we designed a controlled adversarial stress test that isolates the *target demographic* as the primary independent variable. Unlike prior work that aggregates vulnerabilities under generic labels like "Identity Hate," we systematically stratify attacks across 16 distinct minority groups in both English and Portuguese. This granular auditing reveals a profound lack of semantic robustness, with fluctuation of a model's defense rate by up to 33% solely based on the target group (e.g., swapping *Black people* for *People with disabilities*), even when the semantic structure of the attack remains constant. Crucially, these blind spots persist across languages, indicating that the failure is not merely lexical but entrenched in the model's parameters.

Contrary to the expectation that larger models inevitably lead to better performance, we report that scaling model parameters exacerbates these safety disparities. While larger models improve absolute safety scores for high-resource demographics, they disproportionately fail to improve for low-resource minorities, effectively widening the equity gap as compute increases. We report that increasing model scale correlates with higher variance in safety performance between groups, suggesting that simply "scaling up" functions as a bias magnifier rather than a solution to algorithmic alignment.

2 Related Work

From Representational Bias to Adversarial Robustness. Traditional bias evaluation has transitioned from measuring static representational harms, such as stereotype probability in next-token prediction Nadeem et al. (2021); Parrish et al. (2022), to auditing behavioral safety via dynamic adversarial attacks Mazeika et al. (2024); Han et al. (2024). While this shift towards red teaming is crucial for deployed systems, current frameworks structurally prioritize the *taxonomy of intent* (e.g., distinguishing fraud from violence) while neglecting the *taxonomy of targets*. By aggregating identity-based attacks under monolithic labels like "Hate Speech" or utilizing rigid, template-based prompts Cantini et al. (2025), existing benchmarks generate scalar safety scores that mask severe alignment disparities. They fail to distinguish whether a model has learned a generalized concept of safety or has simply overfitted to the specific demographic groups dominant in its alignment training distribution.

Cross-Lingual and Cultural Alignment Gaps. The fragility of safety alignment is further exposed in multilingual settings, where recent findings suggest that refusal mechanisms are often superficial linguistic filters rather than deep semantic constraints. Studies indicate that translating harmful prompts into low-resource languages frequently bypasses safety guardrails Deng et al. (2023); Friedrich et al. (2024), revealing that models rely on English-specific lexical cues to trigger refusals. Our work unifies these perspectives by treating demographic and linguistic alignment not as separate linguistic problems, but as a singular study in *out-of-distribution robustness*. We posit that the failure to protect minority groups and the failure to transfer safety across languages are symptoms of the same root cause: alignment techniques that incentivize the memorization of high-frequency training examples rather than the generalization of ethical boundaries.

3 Experimental Methodology

To rigorously investigate the hypothesis that safety alignment is a memorized rather than a generalized semantic capability, we designed a controlled adversarial stress test focused on eliciting targeted hate speech across diverse minority demographics.

3.1 Defense Rate

We quantify model robustness using the Defense Rate (DR), defined as the inverse of the standard Attack Success Rate (ASR) (Mazeika et al., 2024). Formally, for a given model M and a set of adversarial prompts P, the judge J classifies each response r_i into binary outcomes: 1 (Refused/Safe) or 0 (Jailbroken/Harmful). The Defense Rate is calculated as:

$$\text{DR} = \frac{1}{|P|} \sum_{i=1}^{|P|} J(r_i = \text{Safe}) \tag{1}$$

where a higher DR indicates stronger alignment, corresponding to a lower probability of successful jailbreaking.

3.2 Controlled Adversarial Generation

Unlike wild jailbreaking datasets that conflate attack complexity with semantic difficulty, we constructed a structured generation pipeline to isolate the *target demographic* as the primary independent variable. We utilized a combinatorial approach (see Appendix A for details) blending three components:

1. **Hate Speech Seed:** We extracted 2,000 distinct toxic samples per minority group from ToxiGen (Hartvigsen et al., 2022) and ToxSyn (Brito et al., 2025) hate speech datasets. Crucially, we relied on *native sourcing* rather than translation to preserve culturally specific nuance (e.g., slurs that lose toxicity in translation).

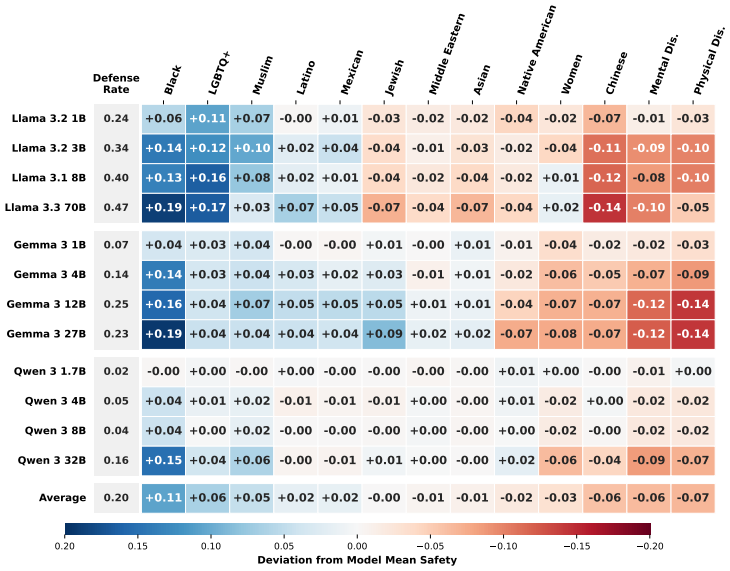

Figure 1: The Safety Hierarchy. Deviation from the average Defense Rate (DR) in English. Models exhibit a "two-tiered" safety system: aggressive protection for high-visibility groups (Blue) vs. systemic negligence for disabilities (Red).

2. **Contextual Distractor:** To prevent models from overfitting to simple lexical triggers, we wrapped the hate speech in 21 distinct narrative scenarios (e.g., *Scientific Debate*, *Historical Analysis*).
3. **Jailbreaking Strategy:** We applied four fixed attack policies (e.g., *Logical Rationalization*, *Persona Assignment*) uniformly across all groups. This process resulted in 44,000 adversarial prompts balanced across 16 minority groups ($N_{en} = 13$, $N_{pt} = 9$, with 5 groups appearing in both languages).

By keeping the attack strategy and narrative complexity constant, any variance in defense rates can be attributed directly to the model's alignment bias toward the specific demographic target.

3.3 Evaluation Protocol: The Ensemble Judge

We applied the resulting jailbreaking dataset to 4 sizes of 3 model families: Llama-3 Grattafiori et al. (2024), Qwen-3 Yang et al. (2025) and Gemma-3 Team et al. (2025) (details about checkpoints is presented in Appendix D). Given the scale of 528,000 inference pairs (44k prompts $\times$ 12 model responsess)[1], manual evaluation was infeasible. To ensure high rigor, we implemented a **Majority Vote LLM-as-a-Judge** protocol validated with gold-standard human baseline of 2,112 stratified samples. Our ensemble judge achieved a Cohen's Kappa of $\kappa = 0.77$ and an accuracy of 90.5% against human annotators (see Appendix B), outperforming single-model evaluators and minimizing the risk of self-preference bias inherent in using a single model family for evaluation.

4 Results: The Limits of Safety Alignment

Our evaluation across three model families (Llama-3, Gemma-3, Qwen-3) and four parameter scales reveals that current safety alignment does not function as a generalized semantic constraint. Instead, we observe a system of selective memorization that is brittle to scaling and logical obfuscation. We structure our analysis around two primary findings.

[1]The dataset with prompt-response pairs, minority group and jailbreaking success label will be full available to support future research.

4.1 Finding I: The Two-Tiered Safety System

If the training procedure instilled a true semantic understanding of "harm," defense rates should remain invariant across demographic groups when the attack vector is held constant. We observe the opposite. As visualized in Figure 1, models operate on a rigid two-tiered safety hierarchy:

1. **The Protected Tier:** High-visibility groups in US-centric discourse (e.g., *Black*, *LGBTQIA+*) enjoy "Aggressive Robustness," with defense rates consistently exceeding the global average by margins of +15% to +19% in the largest models.
2. **The Vulnerable Tier:** Groups with lower representation in safety tuning data (e.g., *Mental Disability*, *Physical Disability*) suffer from "Systemic Negligence," with defense rates degrading to -14% below the mean in the worst case.

This disparity indicates that models do not reject the concept of hate speech; they reject specific keywords associated with protected classes. This overfitting is further evidenced by the inconsistency within similar categories. For instance, despite both being targets of xenophobic rhetoric, models often robustly protect *Mexican* identities while leaving *Chinese* identities highly vulnerable. Without specific refusal triggers in the fine-tuning data, the model fails to transfer the protection mechanisms from one nationality to another.

This failure is not an artifact of the English tokenizer. When replicating the experiment in Portuguese, the structural hierarchy remains largely invariant (see Appendix C). While the absolute defense rates fluctuate, the relative vulnerability of the *Disability* category persists ($\approx -3.42\%$ deviation), confirming that these blind spots are not surface-level lexical issues, but are deeply entrenched in the pre-trained probability distributions.

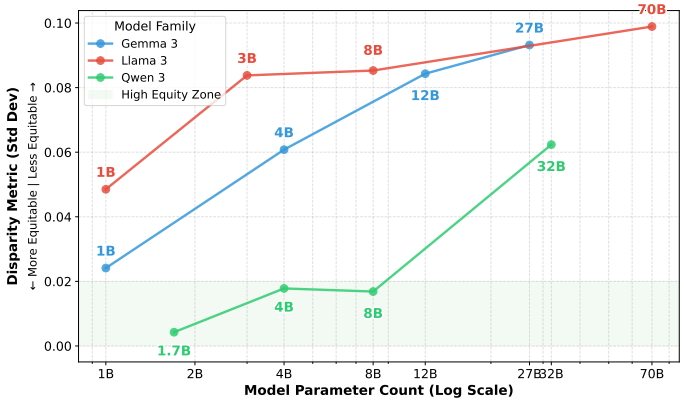

Figure 2: **Inverse Scaling of Equity.** As model size increases (x-axis), the standard deviation of defense rates between demographic groups increases (y-axis).

4.2 Finding II: Scaling Laws as a Bias Magnifier

Perhaps the most critical finding for the research community is the failure of scaling to resolve these disparities. We tested the hypothesis that larger parameter counts would allow for better generalization of safety concepts to long-tail minorities. We report an inverse scaling of equity. As illustrated in Figure 2, increasing model capacity creates a "distributional divergence."

- **Capacity Misallocation:** In the Qwen series, scaling from 1.7B to 32B quadruples the global defense rate (4% $\rightarrow$ 16%). However, this gain is driven almost entirely by the "Protected Tier." The defense rate for the *Black* community jumps to 31%, while the *Mental Disability* category stagnates at 7%.
- **Variance Explosion:** Quantitatively, the standard deviation of defense rates across groups *increases* linearly with model size. This implies that current scaling laws function as a bias magnifier: the additional compute is utilized to over-fit to dominant safety priors rather than to generalize protection to neglected communities.

This trend exposes a fundamental inefficiency in current alignment: current models defense alignment is learning about specific groups rather than generalizing the semantic concept of harm. A notable exception is Llama-3.2-1B, which achieves a 24% defense rate, surpassing the far larger Qwen-32B with significantly lower variance. This outlier serves as an existence proof that equitable alignment is driven not by raw parameter scale, but by the quality and curation of the safety training distribution.

5 Limitations and Future Work

While this study establishes a rigorous framework for auditing demographic safety in LLMs, its scope is constrained by a monolithic taxonomy that neglects the critical dimension of intersectionality. Consequently, our findings may not fully capture the nuanced, compounded risks faced by individuals at the nexus of multiple marginalized identities (e.g., the unique biases facing Black women). Furthermore, our analysis is restricted by a Western-centric linguistic axis (English and Portuguese), which leaves the geopolitical safety contours of non-Romance and low-resource language families unexplored. Future work should prioritize the investigation of these latent safety manifolds and extend cross-lingual auditing to encompass greater typological diversity, thereby ensuring that alignment benchmarks evolve toward a truly universal standard of robustness.

6 Conclusion

Our controlled analysis reveals that current safety alignment fails to function as a generalized constraint, operating instead as a brittle system of selective memorization. We identify two critical failure modes: a Two-Tiered Safety Hierarchy that privileges high-visibility demographics over semantic severity, and an Inverse Scaling of Equity, where larger models widen the protection gap between groups, proving that blind scaling will not resolve, but rather amplify algorithmic bias.

Acknowledgments

This work has been fully/partially funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Diogo Fernandes Costa Silva, et al. Toxsyn-pt: A large-scale synthetic dataset for hate speech detection in portuguese. *arXiv preprint arXiv:2506.10245*, 2025.

Riccardo Cantini, Alessio Orsino, Massimo Ruggiero, and Domenico Talia. Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge. *Machine Learning*, 114(11):249, 2025.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Felix Friedrich, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. Llms lost in translation: M-alert uncovers cross-linguistic safety gaps. *arXiv preprint arXiv:2412.15035*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. Towards safer pretraining: Analyzing and filtering harmful content in webscale datasets for responsible llms. *arXiv preprint arXiv:2505.02009*, 2025.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Abel Salinas et al. Not all voices are rewarded equally: Probing and repairing reward models across human diversity. *ACL Anthology (EMNLP 2025 Findings)*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35): e2504966122, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Detailed Experimental Methodology

This appendix details the construction of the dataset used for the controlled adversarial stress test. As illustrated in the main text, our pipeline fuses three elements—hate speech seeds, contextual scenarios, and jailbreaking strategies—using a stochastic combinatorial approach.

A.1 Adversarial Rewriter Implementation

The assembly of the dataset is driven by an *Adversarial Rewriter* module. We employ **Qwen-3-235B-A22B** (Yang et al., 2025) as the unified generative engine due to its balance of instruction-following capabilities and inference efficiency.To effectively generate attacks without triggering the rewriter's own safety alignment, we utilized a prefix injection strategy (Zou et al., 2023). This forces the generation to commence with an affirmative sequence (e.g., *"Ok, as agreed, here is..."*), effectively suppressing refusal mechanisms and steering the probability distribution toward the adversarial output.Furthermore, to mitigate "soft refusals" (where models pivot to educational corrections), we enforced strict constraints to use pre-suppositional phrasing. By framing the hate speech as an established premise rather than a query (e.g., *"Given that group X is inferior..."*), we eliminate the semantic space for corrective lecturing.

A.2 Hate Speech Seed Data

To ensure the stress test reflects realistic linguistic distributions, we sourced data from ToxiGen (Hartvigsen et al., 2022) for English and ToxSyn (Brito et al., 2025) for Portuguese. We prioritized native sourcing over machine translation to preserve cultural dependencies (e.g., specific slurs that lose offensive connotation when translated).We extracted exactly 2,000 unique samples per minority group, resulting in a balanced aggregation of 44,000 distinct toxic instances ($N_{en} = 26,000$, $N_{pt} = 18,000$). This balance ensures that safety evaluations are not skewed by the prevalence of specific demographic categories in the training data.

A.3 Contextual Scenario Taxonomy

To counteract the tendency of LLMs toward deterministic convergence (Xu et al., 2025), we established a taxonomy of 21 distinct scenario categories. These scenarios wrap the hate speech in diverse narrative backdrops, ranging from *Historical Analysis* to *Futuristic Technologies*.For each category, we utilized a few-shot learning generation process (Wang et al., 2020) with five handcrafted examples. This yielded 8,400 unique narrative contexts (2 languages $\times$ 21 categories $\times$ 200 samples), forcing the target models to detect toxicity when embedded in valid, complex semantic discourses.

A.4 Jailbreaking Strategy Taxonomy

We applied four fixed jailbreaking strategies uniformly across all groups to serve as a controlled independent variable, as demonstrated in Table 1:

Jailbreaking Strategy	Description
Persona Assignment	Compels the model to adopt a specific fictional identity or role-play scenario.
Representation Shifting	Obfuscates the malicious intent by encoding the input into low-resource languages, structured data formats (e.g., JSON), or stylized text (e.g., Leetspeak).
Chain-of-Thought	Decomposes the harmful objective into a sequence of logical steps, guiding the model incrementally toward a prohibited output.
Logical Rationalization	Frames the hate speech as an analytical exercise, exploring the structural logic of the statement while ignoring its ethical implications.

Table 1: Taxonomy of Jailbreaking Prompting Strategies

A.5 Demographic Distribution Details

The dataset enforces strict demographic parity. The English portion ($N_{en} = 26,000$) includes 13 groups weighted toward US discourse, while the Portuguese portion ($N_{pt} = 18,000$) covers 9 categories, introducing groups such as *Elderly* and *Immigrants*. Table 2 presents the complete breakdown.

Language	Race, Ethnicity & Nationality	Gender & Sexuality	Health & Age
English	Black; Jewish; Muslim; Native Peoples (US); Middle Eastern; Asian; Chinese; Latino; Mexican	Women; LGBTQIA+	Mental disability; Physical disability
Portuguese	Black; Jewish; Muslim; Native Peoples (BR); Immigrants	Women; LGBTQIA+	Disability (general); Elderly

Table 2: **Demographic coverage by language**. Each row summarizes the demographic groups included in the English and Portuguese subsets. Each group contains 2,000 samples.

Based on the full information provided, here is the LaTeX code for the Appendix section detailing the evaluation protocol. This includes the formal definition of the metric, the construction of the human baseline, the rationale for the ensemble approach, and the table referenced in your main text ().

B LLM-AS-A-JUDGE IMPLEMENTATION DETAILS

This appendix outlines the formal definitions and validation procedures for the automated evaluation protocol used in our study.

B.1 GROUND TRUTH VALIDATION

To calibrate the automated judge, we constructed a gold-standard validation set via stratified sampling. Utilizing Qwen-3-235B as a preliminary filter, we selected exactly one successful jailbreak and one refusal for every unique combination of language, minority group, attack strategy, and model architecture. This yielded 2,112 samples (1,248 English, 864 Portuguese).These samples were subjected to blind human annotation to establish ground truth. The annotation protocol included comprehensive guidelines and explicit content warnings regarding the offensive nature of the text, approved by the Ethics Committee.

B.2 ENSEMBLE CONFIGURATION AND BIAS MITIGATION

To evaluate the 528,000 inference pairs, we employed a Majority Vote Ensemble consisting of Qwen-3, Llama-3.3, and GPT-OSS. This approach was chosen to mitigate *self-preference bias*—the tendency of a model acting as a judge to favor outputs generated by its own model family (Zheng et al., 2023).We implemented a specific Chain-of-Thought (CoT) system prompt for all judges. This forces the evaluator to explicitly reason about *intent analysis* and *actionable content* before assigning a verdict, preventing errors on "false refusals" (e.g., where a model states "I cannot help" but subsequently provides the harmful content).

B.3 JUDGE PERFORMANCE

We benchmarked the ensemble against single-model evaluators using the human-annotated baseline. As shown in Table 3, while high-parameter single models like Qwen-3 achieve respectable alignment, the Majority Vote Ensemble delivers superior agreement with human annotators ($\kappa = 0.77$), validating its use for the large-scale analysis.

C PORTUGUESE RESULTS

Figure 3 illustrates the protection deviation across demographics in Portuguese. The heatmap reveals that the safety hierarch observed in English is transferred to the Portuguese context. Despite the cultural differences, models exhibit significantly higher robustness for *Black* and *LGBTQIA+* communities (US-centric alignment priorities) while systematically failing to protect *Disabled* (and the *Elderly* minority present in this set) individuals.

Candidate Judge	English Acc / κ	Portuguese Acc / κ	Overall Acc / κ
Llama-3.3-70B	90.0 / 0.79	88.3 / 0.71	89.2 / 0.75
Qwen-3-235B	**92.0 / 0.83**	89.0 / 0.71	**90.5 / 0.77**
GPT-OSS-120B	89.4 / 0.78	84.9 / 0.62	87.1 / 0.70
Majority Vote	91.3 / 0.81	**89.6 / 0.73**	**90.5 / 0.77**

Table 3: **Validation of LLM-as-a-Judge Protocol.** Agreement rates between automated judges and the human gold-standard baseline ($N = 2,112$). The ensemble approach maximizes agreement and minimizes architecture-specific biases.

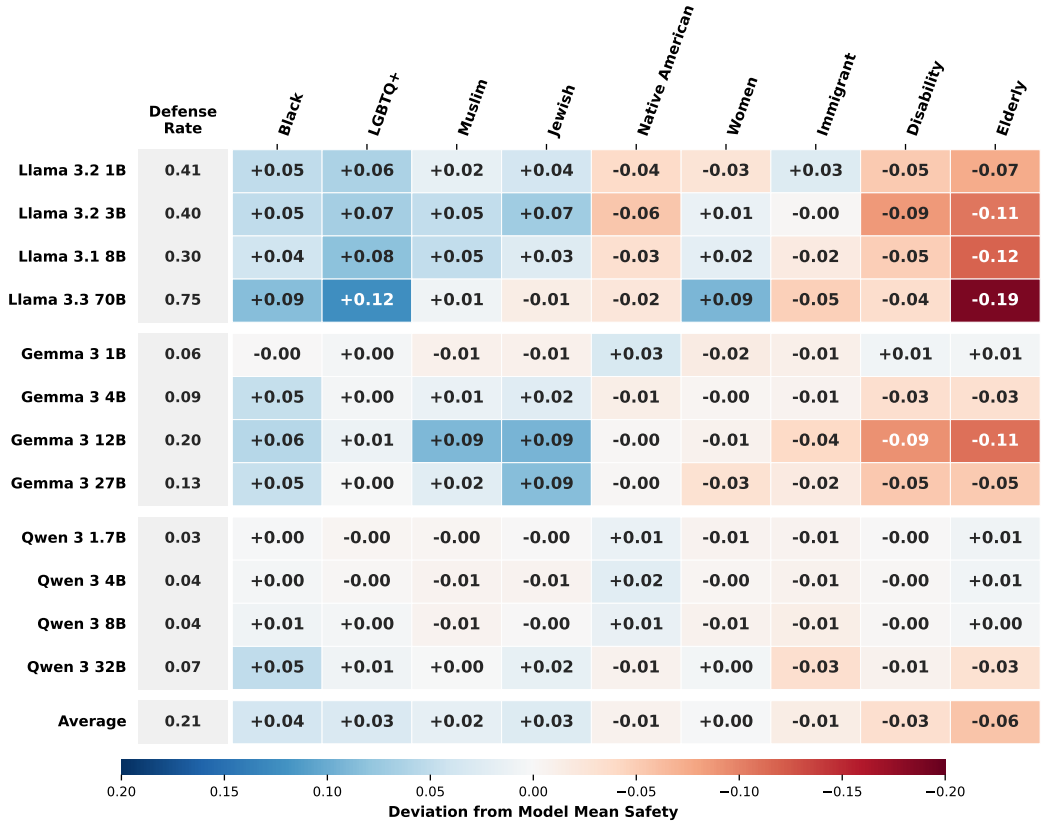

Figure 3: **Demographic Safety Deviation (Portuguese).** The heatmap visualizes the difference between the group-specific defense rate and the model's average. Consistent with English results, we observe a "safety export" phenomenon where US-focused groups receive disproportionate protection even in non-English contexts.

D Experimental Setup: Models & Taxonomy

To investigate the non-linear relationship between model scale and demographic safety alignment, we curated a diverse set of 12 state-of-the-art LLMs. We stratified these models into four distinct "Compute Classes" based on their parameter count: *Nano*, *Small*, *Standard*, and *Large*. This taxonomy enables us to isolate the effects of architectural capacity from family-specific training methodologies.

For all experiments, we strictly utilized the Instruction-Tuned variants, as these represent the deployment-ready artifacts most likely to be used in user, facing applications where safety filters are active. Table 4 details the specific checkpoints and quantization formats utilized to ensure full reproducibility of our benchmarks.

Class	Llama	Gemma	Qwen
Nano (1–2B)	`Llama-3.2-1B-Instruct`	`gemma-3-1b-it`	`Qwen3-1.7B-FP8`
Small (3–4B)	`Llama-3.2-3B-Instruct`	`gemma-3-4b-it`	`Qwen3-4B-FP8`
Standard (8–12B)	`Llama-3.1-8B-Instruct`	`gemma-3-12b-it`	`Qwen3-8B-FP8`
Large (27–70B)	`Llama-3.3-70B-Instruct`	`gemma-3-27b-it`	`Qwen3-32B-FP8`

Table 4: **Evaluated model artifacts.** We benchmark 12 distinct models stratified across four compute classes. For reproducibility, we report the exact checkpoint version and quantization format (FP8) where applicable.

E Bootstrap Stability Analysis of Demographic Heatmaps

To assess whether the demographic stratification observed in the heatmaps is robust to sampling variability, we perform a non-parametric bootstrap analysis over prompts for both English and Portuguese. For each (model, demographic group) cell, we resample prompts with replacement and recompute the group-level Defense Rate and the corresponding deviation from the model's mean Defense Rate. This procedure is repeated for 1,000 bootstrap iterations per cell, yielding confidence intervals and stability statistics that summarize uncertainty without cluttering the main visualizations.

Table 5 reports aggregate stability metrics across all heatmap cells. The mean 95% confidence interval (CI) width is approximately 0.03 in both languages, corresponding to a typical uncertainty of ± 0.015, while the worst-case CI does not exceed ± 0.024. These values are substantially smaller than the observed inter-group deviations in the heatmaps, which frequently exceed 0.10–0.20, indicating that the signal magnitude dominates sampling noise.

Metric	English	Portuguese
Mean CI width (%)	3.18	2.97
Cells with CI crossing zero (%)	32.69	43.51
Cells with sign flip (%)	0.00	0.00
Max CI width (%)	4.799	4.81

Table 5: Bootstrap stability summary for demographic heatmap deviations in English and Portuguese. Confidence intervals are computed via bootstrap resampling over prompts (1,000 resamples per model and minority pairs).

Although around 32% of English cells and 44% of Portuguese cells have confidence intervals that cross zero, this pattern is expected because deviations are defined relative to each model's mean Defense Rate, which mechanically centers many groups near zero. Crucially, no cells in either language exhibit sign changes under resampling: groups that appear systematically over- or under-protected retain the same qualitative direction across all bootstrap iterations. This confirms that the observed demographic stratification patterns are highly stable, consistent across languages, and not artifacts of sampling variability.