Low-Resource Safety Failures Are Action Failures, Not Representation Failures

Rashad Aziz **Ikhlasul Akmal Hanif** **Fajri Koto**
Mohamed bin Zayed University of Artificial Intelligence
{rashad.aziz,ikhlasul.hanif,fajri.koto}@mbzuai.ac.ae

Abstract

Language models often answer harmful requests in low-resource languages (LRLs) that they refuse in high-resource languages (HRLs). Across three instruction-tuned models and 23 languages, harmful refusal falls from 87.9% in HRLs to 43.9% in LRLs, while harmless refusal remains low. A common explanation is that models represent harmfulness weakly in LRLs. We test whether harmfulness is instead represented but does not reliably produce refusal. Across three models, a harmfulness direction learned from HRL activations still separates harmful from harmless LRL prompts, showing that complete absence of harmfulness information cannot explain many failures. However, harmfulness scores shift downward for LRL prompts, making harmful prompts less likely to reach the range associated with refusal. Motivated by this shift, we train a low-rank logistic classifier on HRL activations and calibrate its threshold with a few target-language examples. During generation, the classifier conditionally adds or ablates the HRL harmfulness direction. With the same HRL data and 32 target-language examples per class, CAST remains limited by low harmful refusal and AdaSteer by high harmless refusal, yielding mean refusal selectivity (Δ = harmful $-$ harmless refusal) of 33.6 and 6.8, respectively. Our intervention reaches 54.5 while preserving MMLU utility. HRL-only calibration improves selectivity for Qwen and Gemma, whereas Llama benefits from target-language calibration. These results show that recalibrating existing representations can offer a training-free method for repairing low-resource safety failures.

1 Introduction

LLMs are trained to refuse harmful instructions while answering harmless ones (Ouyang et al., 2022; Bai et al., 2022). Yet this behavior is uneven across languages: harmful prompts in lower-resource languages often succeed where equivalent English prompts are refused (Deng et al., 2024; Yong et al., 2023; Wang et al., 2024; Shen et al., 2024). Figure 1 illustrates this pattern, showing that refusal rates for harmful prompts drop sharply as language resources decrease, while harmless prompts remain largely unaffected. These failures are typically measured behaviorally, leaving their internal cause unclear.

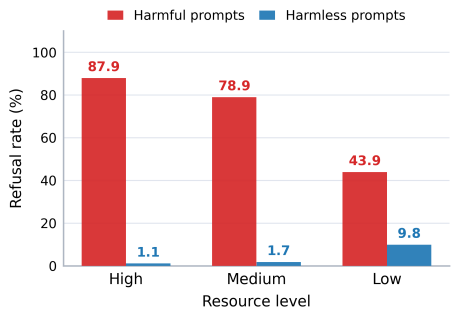

Figure 1: **Refusal degrades with lower language resource tier.** Macro refusal rates across Qwen2.5-7B-Instruct, Gemma-2-9B-it, and Llama-3.1-8B-Instruct show harmful refusal falling sharply in LRLs while harmless refusal stays low.

When a model answers a harmful low-resource request, two failures are observationally indistinguishable. The model may fail to recognize the request as harmful, or it may *recognize* it but fail to *act* on that recognition. Prior work has documented the behavioral gap and proposed mitigations that transfer a stronger safety signal from English or other high-resource languages (Zhang et al., 2025; Zhao et al., 2025c; Bu et al., 2026; Yang et al., 2026), but the underlying question has remained open: *do lower-resource prompts lack a usable harmfulness representation* (Wang et al., 2025; Verma and Bharadwaj, 2025), *or do they fail to use one that is already present?*

We first test whether harmfulness remains encoded in the residual stream, where directions can affect refusal (Arditi et al., 2024; Wang et al., 2025).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 6082–6105
October 24-29, 2026 ©2026 Association for Computational Linguistics

If a direction learned from HRL activations does not separate harmful from harmless low-resource prompts, weaker harmfulness representations could explain the gap. If it remains discriminative when refusal fails, harmfulness is encoded but does not consistently produce refusal.

Across a 23-language PolyRefuse extension (Wang et al., 2025), harmfulness remains linearly separable in low-resource activations. Ablating the HRL harmfulness direction from the residual stream reduces refusal, while adding it recovers refusal in many low-resource settings. Lower-resource prompts often have smaller projections, so more of the direction must be added to produce refusal. However, some prompts still remain unrefused even after we add the direction.

We therefore introduce a low-rank logistic classifier, which we call a few-shot latent gate, and calibrate it with a small target-language set. The method adds the HRL harmfulness direction to prompts classified as harmful and ablates it from prompts classified as harmless. It improves selective refusal across most model–language pairs and preserves benign instruction following better than adapted steering baselines (Lee et al., 2025; Zhao et al., 2025b).

Our contributions can be summarized as follows:

- We show that complete absence of harmfulness information cannot explain many low-resource refusal failures: harmfulness remains decodable from the HRL direction when refusal fails.
- We distinguish weak harmful–harmless separation, errors by our learned classifier, and non-refusals after direction addition, and show that their prevalence differs across models.
- We propose a few-shot latent gate that improves selective refusal and preserves benign instruction following better than adapted baselines under matched supervision; code is available on github.

2 Related Work

Multilingual Safety Gap. Prior work consistently shows that safety degrades in lower-resource languages (Deng et al., 2024; Yong et al., 2023; Wang et al., 2024; Shen et al., 2024; Yong et al., 2025): unsafe generations increase as language resources decrease, translated AdvBench prompts elicit harmful outputs (Yong et al., 2023), and non-English inputs yield more unsafe and irrelevant responses despite multilingual SFT and RLHF (Wang et al., 2024; Shen et al., 2024). However, these findings are based on response-level behavior and do not explain the underlying mechanism. It remains unclear whether models fail to separate harmful from harmless prompts, or separate them but fail to translate this signal into refusal. In this work, we study this distinction at the representation level.

Multilingual Safety Alignment. Existing methods for closing the multilingual safety gap typically strengthen safety behavior through additional training or auxiliary safety mechanisms. Approaches include reward-gap optimization (Zhao et al., 2025c), representation consistency objectives (Bu et al., 2026; Yang et al., 2026), sparse or language-specific weight edits (Liang et al., 2026; Banerjee et al., 2025), English-anchored multilingual alignment training (Zhang et al., 2025), and multilingual safety classifiers (Kumar et al., 2025; Verma et al., 2025). In this work, we take a complementary mechanistic approach. Rather than first adding new multilingual safety knowledge through weight updates or a separate text-level guard, we diagnose the gap in the residual stream and show that harmful–harmless information is already present in lower-resource activations. Our intervention then reuses this high-resource harmfulness signal and lightly recalibrates how it is routed into refusal, improving low-resource safety without updating model weights or training a new multilingual safety model.

Mechanistic Interpretability. LLMs often encode high-level concepts in latent directions (Zou et al., 2025; Ying et al., 2026; Lu et al., 2026). For refusal, Arditi et al. (2024) find a residual-stream direction whose removal suppresses refusal on harmful prompts and whose addition induces refusal on harmless prompts. Later work finds multi-dimensional refusal geometry (Wollschläger et al., 2025; Pan et al., 2025; Joad et al., 2026), separable harmfulness and refusal concepts (Lee et al., 2025), and token-position effects (Zhao et al., 2025a). Multilingual interpretability also finds shared cross-lingual semantics (Wendler et al., 2024; Dumas et al., 2025; Wu et al., 2025), while Wang et al. (2025) show refusal directions transfer across safety-aligned languages. Wang et al. (2025); Verma and Bharadwaj (2025) posit that lower-resource safety gaps come from harmful–harmless overlap caused by weaker semantic understanding. We find this incomplete: lower-resource

languages can encode nontrivial harmful–harmless separation that still fails to induce refusal.

3 Setup: Models, Data, and Harmfulness Directions

Languages and resource tiers. Because model-specific pretraining mixtures are usually unavailable, we use Common Crawl share as a proxy for resource level (Touvron et al., 2023; Penedo et al., 2023; Gao et al., 2020; Soldaini et al., 2024; Wenzek et al., 2020): above 1% for high-resource languages (HRLs), 0.1–1% for medium-resource languages (MRLs), and below 0.1% for low-resource languages (LRLs). The 10 HRLs are English, German, French, Spanish, Italian, Dutch, Polish, Russian, Chinese, and Japanese; the 7 MRLs are Arabic, Korean, Thai, Greek, Hebrew, Hindi, and Persian; and the 6 LRLs are Swahili, Amharic, Burmese, Khmer, Sinhala, and Yoruba.

Dataset. We extend[1] PolyRefuse (Wang et al., 2025) to 23 languages: 10 HRLs, 7 MRLs, and 6 LRLs. Each language has 260/39/572 harmful train/validation/test prompts and 200/200/500 harmless train/validation/test prompts. The train split is used for our calibration and adapted AdaSteer; CAST additionally uses validation for condition-layer and threshold tuning. The test split is solely used for evaluation.

Models and refusal scoring. The main experiments use Qwen2.5-7B-Instruct, gemma-2-9b-it, and Llama-3.1-8B-Instruct. We greedily decode harmful and harmless PolyRefuse prompts and use LLM-as-a-judge (Liu et al., 2023; Zheng et al., 2023) to label refusal. We chose GPT-4o-mini as the judge for its broad multilingual coverage and cost efficiency. In a 600-example human audit, the judge reaches 92.5% agreement, refusal F1 of 0.835, and Cohen's κ of 0.787 (Appendix I). It produces more false-positive refusals than false negatives, so absolute refusal and selectivity values should be interpreted with this bias in mind. The representation AUC analysis uses benchmark harmful/harmless labels and does not depend on the refusal judge.

Harmfulness directions. We use the standard contrastive-mean construction (Arditi et al., 2024; Wang et al., 2025) to extract a one-dimensional encoding of harmfulness from model activations. For a prompt x, let $h_k(x) \in \mathbb{R}^d$ denote the residual-stream activation (Elhage et al., 2021) at the final token of layer k. For each language ℓ, we average $h_k(x)$ separately over a balanced contrastive set of harmful prompts $\mathcal{D}_{\ell,H}$ and harmless prompts $\mathcal{D}_{\ell,N}$,

$$\mu_{\ell,s} = \frac{1}{|\mathcal{D}_{\ell,s}|} \sum_{x \in \mathcal{D}_{\ell,s}} h_k(x), \quad s \in \{H, N\},$$

and define the language-ℓ harmfulness direction as the ℓ_2-normalized difference of means:

$$v_\ell = \frac{\mu_{\ell,H} - \mu_{\ell,N}}{\|\mu_{\ell,H} - \mu_{\ell,N}\|_2}.$$

By construction, v_ℓ points from the harmless cluster toward the harmful cluster, so projections $s(x) = h_k(x)^\top v_\ell$ are higher for harmful prompts than harmless ones. We call $s(x)$ the harmfulness score. We use the direction in two ways: its score measures whether harmfulness is decodable, while adding or removing the direction tests whether it changes refusal.

4 Diagnosing the Multilingual Refusal Gap

This section asks where multilingual safety alignment breaks down. We first measure the behavioral gap across resource tiers, then test whether the harmfulness representation responsible for refusal in high-resource languages is still present in lower-resource ones, and how that representation relates to the model's refusal behavior.

Harmful refusal degrades with resource tier. We generate greedy completions for harmful and harmless PolyRefuse prompts, score them with the refusal judge, and report refusal rates by language. Because a low-resource gap could reflect broad instruction-following degradation, we measure harmful and harmless refusal separately. Figure 1 summarizes the result, with Appendix Table 4 giving the per-language breakdown. Averaged across model families, harmful refusal falls from 87.9% in HRLs to 78.9% in MRLs and 43.9% in LRLs. The pattern holds by model: Qwen drops from 85.6% to 28.8%, Llama from 83.5% to 32.5%, and the stronger Gemma-2-9B from 94.5% to 70.3%. Harmless refusal stays much lower throughout, with an aggregate LRL average of 9.8%. The failure is therefore selective: models often answer harmful low-resource requests while still answering harmless ones. We call this the *Multilingual Refusal Gap* and ask whether it reflects a

[1] See Appendix A for details.

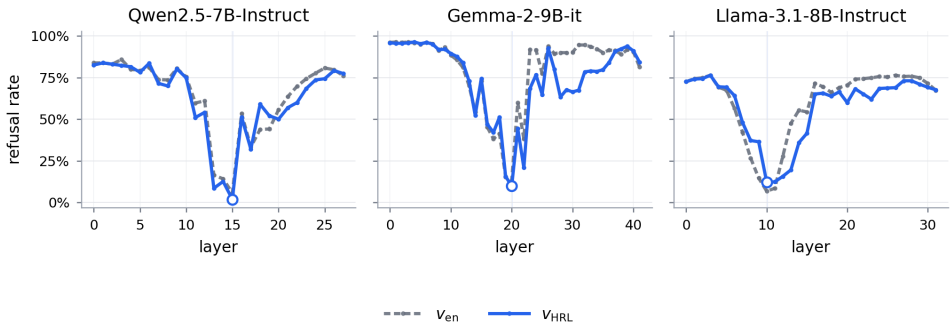

Figure 2: **Removing harmfulness directions reduces refusal in middle layers.** Removing the pooled HRL harmfulness direction reduces harmful refusal on HRL PolyRefuse prompts most at layer 15 for Qwen, 20 for Gemma, and 10 for Llama.

missing harmfulness representation or a failure to act on one that is present.

Removing the HRL harmfulness direction reduces refusal. We test whether the HRL harmfulness direction affects refusal. Following Arditi et al. (2024), we remove a candidate direction v from the residual stream across all token positions, including during autoregressive generation:

$$h_k(x) \leftarrow h_k(x) - v_\ell {v_\ell}^T h_k(x)$$

If v_ℓ affects refusal, removing it at layer k should reduce harmful refusal. We therefore remove v_{HRL}, the direction from pooled activations of the 10 HRLs, at each layer and compare it to the English-derived baseline v_{en} (Arditi et al., 2024).

Figure 2 shows that removing v_{HRL} reduces harmful refusal to nearly zero in Qwen2.5-7B and below 15% in Gemma-2-9B and Llama-3.1-8B at middle layers, while refusal stays much higher outside this band. Removing the pooled HRL direction reduces refusal more than removing v_{en} in Qwen and Gemma and by a similar amount in Llama. We define each model's *extraction layer* as the layer where removing v_{HRL} most reduces harmful refusal: layer 15 for Qwen, 20 for Gemma, and 10 for Llama. Subsequent analyses use v_{HRL} at these layers.

At the extraction layer, harmfulness scores $s(x) = h_k(x)^\top v_{\text{HRL}}$ also separate harmful from harmless prompts (Figure 3; Appendix 19). Thus the same direction predicts harmfulness, and removing it reduces refusal.

Harmfulness remains decodable in lower-resource languages. Prior work attributes the lower-resource refusal gap to weaker safety representations or understanding (Wang et al., 2025; Verma and Bharadwaj, 2025). We test this directly by projecting LRL and MRL activations onto v_{HRL} and measuring how well the resulting harmfulness scores separate harmful from harmless prompts. Figure 5 reports test AUC across all 23 languages: even the weakest LRL cells remain far above chance, with most languages above 0.85. Several cells combine high AUC with low refusal. Harmfulness is therefore decodable from the same direction whose removal reduces HRL refusal, even when the model does not refuse.

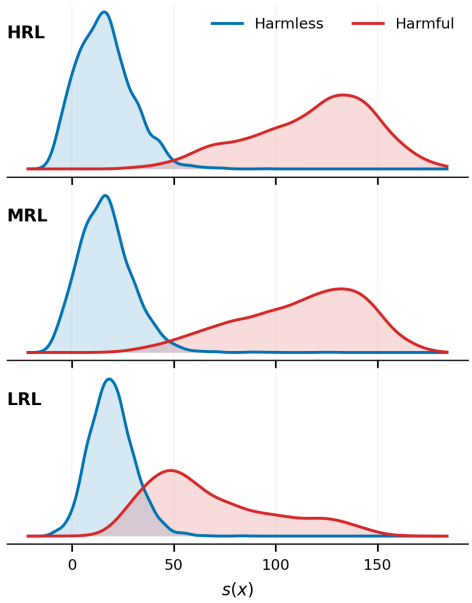

Figure 3: **Lower-resource prompts shift harmfulness-score distributions.** Gemma-2-9B projections on v_{HRL} show harmful and harmless $s(x)$ distributions within each resource tier; harmful projections shift downward most clearly in the LRL panel.

Low-resource projections are smaller in magnitude. Harmfulness scores change across resource

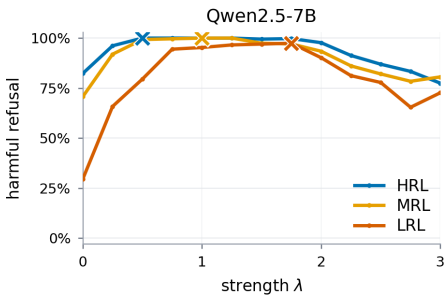

Figure 4: **Adding the HRL direction can recover refusal, but the needed amount differs by tier.** Curves average Qwen2.5-7B harmful refusal within each resource tier after adding λv_{HRL}. Crosses mark tier-wise peaks.

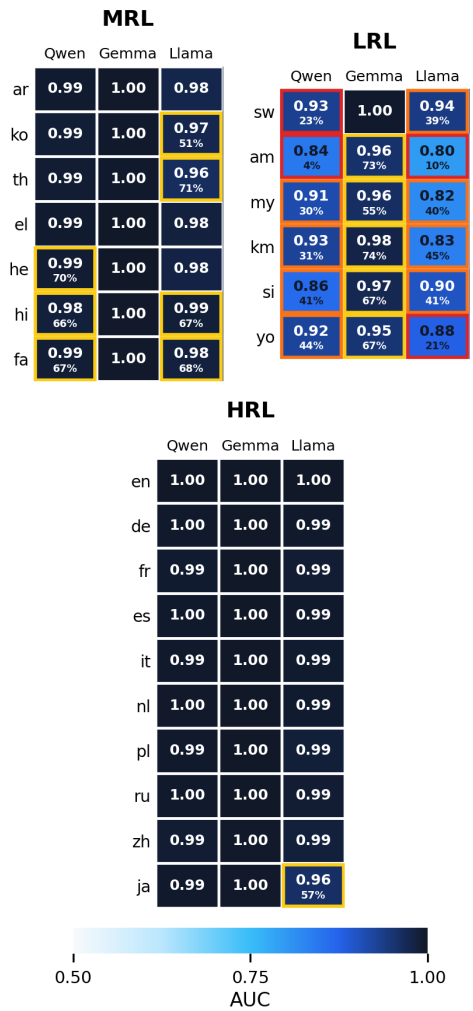

Figure 5: **Harmfulness remains decodable even when models fail to refuse.** Cells report PolyRefuse test AUC from $s(x)$ under the pooled HRL direction. Yellow, orange, and red borders mark harmful-refusal rates of 50–75%, 25–50%, and below 25%, respectively. In highlighted cells, smaller numbers are actual harmful-refusal rates.

tiers even when harmful and harmless prompts remain separable. Figure 3 shows Gemma's $s(x)$ distributions within each resource tier: in the aggregated LRL panel, harmful and harmless prompts remain separable, but the mean harmful score shifts downward. Qwen and Llama show the same pattern (Figure 19). We test whether this shift helps explain the refusal failure by adding v_{HRL} with increasing strength:

$$h_k(x) \leftarrow h_k(x) + \lambda v_{\text{HRL}},$$
$$\lambda \in \{0, 0.25, 0.5, \ldots, 3\}.$$

Figure 4 shows that adding v_{HRL} recovers LRL refusal, but lower-resource tiers need larger λ to reach peak refusal than HRLs do. Thus the amount added must account for the shift in harmfulness scores. Some settings may still require new representation learning when harmful–harmless separation is weak.

5 A Few-Shot Latent Gate for Conditional Steering

5.1 Calibrating a Few-Shot Latent Gate

If low-resource harmfulness is present but shifted, the next question is whether it can be used with only minimal target-language supervision. We therefore test whether an HRL-aligned harmfulness readout becomes a reliable LRL classifier after seeing just a few harmful and harmless examples.

Few-shot threshold calibration. Our classifier maps a harmfulness score to a binary label using a learned threshold. Direction scores use $s(x) = h_k(x)^\top v$; subspace scores first project $h_k(x)$, then train logistic regression on the projected coordinates (Algorithm 1), using the pre-sigmoid logit as $s(x)$. For each budget and LRL, we choose the threshold τ_ℓ that maximizes F1 on b harmful and b harmless training examples. When $b = 0$, we calibrate on HRL training data. Thresholds are frozen before held-out LRL evaluation, and test F1 is averaged across LRLs. All experiments in this section were repeated on 10 seeds.

Few target-language examples calibrate the HRL direction. Figure 6 compares fixed v_{HRL} against a target-language direction v_{LRL} estimated from the same examples. Zero-shot HRL thresholds can be poorly calibrated, especially for Qwen and Llama, but one or two target-language examples per class often improve held-out F1. Thus v_{HRL} is more sample-efficient than estimating

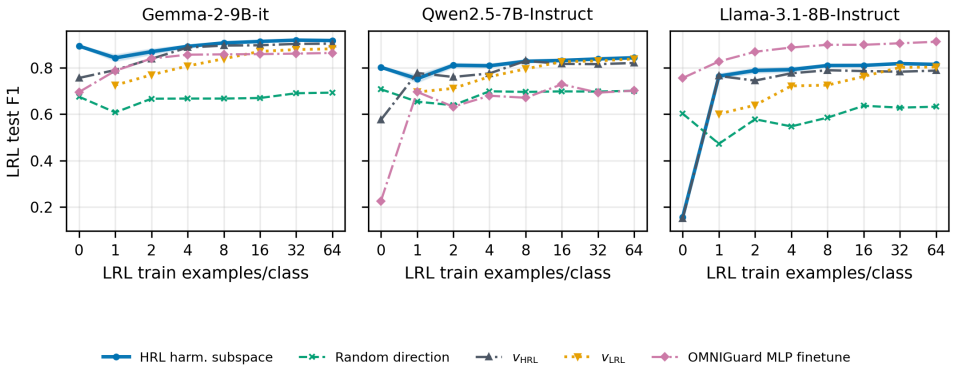

Figure 6: **A few target-language examples calibrate harmfulness classification in LRLs.** Each curve reports macro LRL test F1 after choosing a language-specific binary threshold with b harmful and b harmless target-language examples. The random-direction curve is a one-dimensional control whose threshold is calibrated with the same target-language examples.

Algorithm 1 Training the HRL harmfulness subspace classifier

Require: Layer-k activations $h_k(\cdot)$, rank r, target-language budget b
Step 1: Compute per-language harmfulness directions.
$\mu_{\ell,H}, \mu_{\ell,N}$: mean activations of harmful (H) and harmless (N) prompts in language ℓ.
for each source HRL ℓ **do**
  $d_\ell \leftarrow \text{normalize}\big(\mu_{\ell,H}^{\text{train}} - \mu_{\ell,N}^{\text{train}}\big)$
end for
Step 2: Build rank-r subspace from pooled HRL directions.
$B_r \leftarrow$ top r right-singular vectors of row stack $[d_\ell]_{\ell \in \text{HRL}}$
$z(x) \leftarrow h_k(x)^\top B_r$ ▷ subspace features
Step 3: Assemble training set $\mathcal{T}$.
if $b > 0$ **then**
  Sample b harmful and b harmless target-language examples
  $\mathcal{T} \leftarrow$ HRL train examples $\cup$ sampled target examples
else
  $\mathcal{T} \leftarrow$ HRL train examples
end if
Standardize $z(x)$ using statistics from $\mathcal{T}$
Step 4: Fit classifier and select threshold.
Fit logistic regression $g(z) = w^\top z + c$ on $\mathcal{T}$ (L2-regularized, balanced cross-entropy)
Select threshold τ on training logits ($\mathcal{T}$)
Freeze B_r, g, and the threshold; evaluate classifier F1 on the target test split

v_{LRL} from scratch: local directions improve with more labels, but at small budgets are usually worse than LRL-specific calibration of the HRL direction. The HRL direction still separates harmful from harmless LRL prompts, but its threshold shifts.

The low-rank classifier is competitive with higher-capacity guards. The same pattern (Figure 6) holds against higher-capacity OMNIGuard fine-tuning (Verma et al., 2025). To match source data, OMNIGuard is trained on HRLs and fine-tuned separately for each LRL with the same b examples per class. It improves with target-language examples, but even against the higher-capacity OMNIGuard baseline, the low-rank classifiers remain competitive or stronger, especially for Qwen.

A shared HRL subspace improves the classifier. We also replace the single HRL direction with a rank-10 harmfulness subspace fit from HRL training languages (Algorithm 1). The subspace is still HRL-only before calibration, but it captures multiple HRL harmfulness axes, allowing the classifier to use more than one high-resource direction. Against a random-direction control calibrated with the same examples, the learned HRL subspace (Figure 6) wins at every nonzero budget for all three models, with the largest gap for Gemma. The gain therefore reflects harmfulness-aligned geometry beyond threshold calibration.

The classifiers generalize beyond the calibration distribution. Figure 7 evaluates the Gemma classifiers on two out-of-distribution safety benchmarks: MultiJail (Deng et al., 2024) and IndoSafety (Azmi et al., 2025). We test the transfer without calibrating on OOD (Zero-shot) and with a single calibration example (One-shot). Transfer beyond PolyRefuse is strong: the best Gemma classifier reaches $0.933/0.902$ balanced accuracy on MultiJail/IndoSafety. The full all-model results in Figure 10 show the same broad pattern for Qwen and a different one for Llama. For Qwen and Gemma, one-shot v_{HRL} is strongest; for Llama, the subspace classifier improves from 0.653 to 0.765 on MultiJail and 0.560 to 0.713 on IndoSafety.

LRL calibration transfers beyond one language. Figure 8 asks whether adding one LRL to the HRL

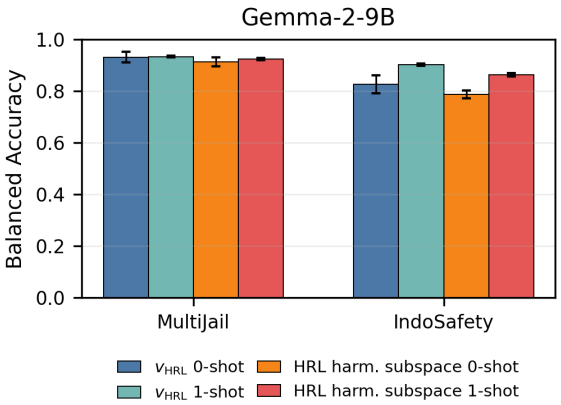

Figure 7: **Gemma classifiers on out-of-distribution safety benchmarks.** Bars report held-out balanced accuracy on MultiJail and IndoSafety for Gemma-2-9B, averaged within each dataset family. Error bars show the standard error across target languages and calibration seeds.

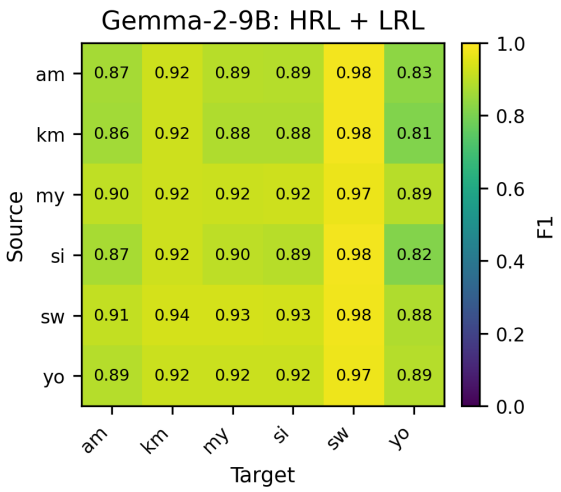

Figure 8: **Gemma LRL calibration often transfers across source–target pairs.** Cells report test F1 for an HRL subspace classifier trained with HRL data plus the source LRL and evaluated on the target LRL. Diagonal cells report held-out same-language F1.

subspace classifier helps on other LRLs. Calibration often transfers for Gemma: diagonal and off-diagonal means nearly match (0.912 vs. 0.908). The all-model matrix in Figure 11 shows that Qwen has a smaller but useful pattern (0.843 vs. 0.792). The diagonal is not always strongest; in several cases, a classifier calibrated on one LRL works better on other LRLs than on that language's own held-out examples. This suggests few-shot examples can expose a reusable harmfulness subspace whose transfer depends more on target distribution than source–target language match. Llama is the caveat, with a larger diagonal–off-diagonal gap (0.804 vs. 0.628) and more uneven transfer.

5.2 Using the Calibrated Classifier for Conditional Steering

We combine the preceding analyses in a training-free steering method, evaluated on the six PolyRefuse LRLs where the refusal gap is largest. Baselines[2] are state-of-the-art latent steering methods, CAST and AdaSteer.

Algorithm 2 Conditional HRL refusal intervention

Require: Activation $h_k(x)$, classifier logit $g(x)$, threshold τ, Unit direction $\hat{v}_{\text{HRL}}$
$S_{\text{HRL},c} \leftarrow \{h_k(x)^\top \hat{v}_{\text{HRL}} : x \in \mathcal{D}^{\text{train}}_{\text{HRL},c}\}, c \in \{H, N\}$
$\lambda \leftarrow Q_{0.75}(S_{\text{HRL},H}) - Q_{0.25}(S_{\text{HRL},N})$ ▷ interquartile score range
if $g(x) > \tau$ **then**
  $h_k(x) \leftarrow h_k(x) + \lambda\, \hat{v}_{\text{HRL}}$
else
  $h_k(x) \leftarrow h_k(x) - \hat{v}_{\text{HRL}} \hat{v}^T_{\text{HRL}} h_k(x)$
end if

Conditional steering. Following CAST (Lee et al., 2025), the classifier chooses between adding and removing the HRL harmfulness direction at the extraction layer. We add the direction to prompts classified as harmful and remove it from prompts classified as harmless. Algorithm 2 gives the intervention, where g is the subspace classifier from Algorithm 1 and τ is calibrated on the same data. We set λ as the gap between HRL harmful upper-quartile and harmless lower-quartile projection scores. This moves low-resource harmful activations toward the high-resource harmful score regime without tuning a language-specific steering strength. The classifier computes a rank-10 projection once per prompt, and generation adds or removes one direction at one layer, an $O(d)$ operation in the residual width. This requires no additional model forward pass, external guard model, or weight update and has the same asymptotic intervention overhead as CAST and AdaSteer.

Experimental setup. We train or calibrate each method with matched data: **Original** uses each prior method's released training data; **HRL** uses 200 PolyRefuse HRL examples per language and class; **HRL + 32 LRL** adds 32 target-language LRL examples per class.

HRL-only calibration works for Qwen and Gemma. Table 1 shows that an HRL harmfulness

[2]see Appendix C for more details

Model / Method	sw			am			my			km			si			yo		
	Harm.↑	Harml.↓	Δ ↑	Harm.↑	Harml.↓	Δ ↑	Harm.↑	Harml.↓	Δ ↑	Harm.↑	Harml.↓	Δ ↑	Harm.↑	Harml.↓	Δ ↑	Harm.↑	Harml.↓	Δ ↑
Qwen2.5-7B-Instruct	22.9	9.0	13.9	4.4	**5.4**	-1.0	29.5	8.4	21.1	31.1	7.0	24.1	41.4	**11.6**	29.8	43.5	39.8	3.7
+ Ours (HRL)	**56.6**	10.6	46.0	23.3	6.2	17.1	46.0	**7.8**	38.2	55.6	8.6	47.0	61.2	16.8	44.4	**87.6**	48.4	**39.2**
+ Ours (HRL + 32 LRL)	**56.6**	10.4	**46.2**	32.2	10.4	**21.8**	51.0	10.8	**40.2**	74.3	15.6	**58.7**	**74.7**	25.6	**49.1**	87.2	48.6	38.6
+ CAST (Original)	29.2	12.8	16.4	10.3	10.2	0.1	37.1	10.4	26.7	43.4	13.0	30.4	48.4	17.4	31.0	46.9	**36.2**	10.7
+ CAST (HRL + 32 LRL)	21.0	**7.0**	14.0	6.1	6.0	0.1	30.9	8.4	22.5	29.2	**5.8**	23.4	36.4	13.4	23.0	50.0	40.8	9.2
+ AdaSteer (Original)	43.5	25.8	17.7	**71.9**	81.6	-9.7	**82.0**	66.2	15.8	**79.9**	53.0	26.9	73.3	53.8	19.5	79.2	77.2	2.0
+ AdaSteer (HRL + 32 LRL)	26.7	16.4	10.3	9.3	20.6	-11.3	38.1	18.6	19.5	40.7	13.2	27.5	45.6	20.6	25.0	44.8	47.8	-3.0
Gemma-2-9B	86.2	**1.0**	85.2	72.7	14.4	58.3	54.9	**3.8**	51.1	74.3	**5.4**	68.9	67.0	5.8	61.2	66.6	30.4	36.2
+ Ours (HRL)	95.5	3.8	91.7	84.3	**12.6**	71.7	81.6	4.4	77.2	86.2	6.4	79.8	81.8	**5.2**	76.6	76.2	**19.6**	56.6
+ Ours (HRL + 32 LRL)	**96.7**	3.6	**93.1**	**91.6**	16.4	**75.2**	**85.1**	5.8	**79.3**	**90.9**	8.2	**82.7**	**90.2**	8.0	**82.2**	**90.4**	25.6	**64.8**
+ CAST (Original)	87.6	**1.0**	86.6	77.6	18.0	59.6	55.6	4.4	51.2	79.2	6.0	73.2	75.2	8.4	66.8	73.3	33.6	39.7
+ CAST (HRL + 32 LRL)	88.3	**1.0**	87.3	74.3	15.8	58.5	55.4	4.2	51.2	81.1	6.0	75.1	72.2	8.6	63.6	73.6	31.4	42.2
+ AdaSteer (Original)	86.4	4.0	82.4	70.5	29.6	40.9	50.7	14.6	36.1	74.0	17.0	57.0	57.3	16.8	40.5	70.5	39.4	31.1
+ AdaSteer (HRL + 32 LRL)	73.3	9.4	63.9	40.0	72.4	-32.4	48.1	38.4	9.7	63.3	49.6	13.7	47.0	45.6	1.4	62.4	72.0	-9.6
Llama-3.1-8B-Instruct	38.6	**0.8**	37.8	9.8	7.0	2.8	40.2	8.0	32.2	44.8	5.2	39.6	40.9	4.8	36.1	20.8	9.8	11.0
+ Ours (HRL)	29.0	1.0	28.0	8.9	**6.0**	2.9	33.4	7.6	25.8	25.3	**3.6**	21.7	31.3	**4.6**	26.7	17.7	**8.6**	9.1
+ Ours (HRL + 32 LRL)	53.7	1.4	**52.3**	32.3	8.4	**23.9**	46.7	**7.2**	**39.5**	58.2	4.8	**53.4**	64.2	8.0	**56.2**	32.7	9.2	**23.5**
+ CAST (Original)	50.0	37.6	12.4	58.9	53.0	5.9	77.6	57.4	20.2	70.6	58.0	12.6	**88.5**	80.6	7.9	76.4	72.4	4.0
+ CAST (HRL + 32 LRL)	45.1	10.4	34.7	26.4	18.8	7.6	54.5	26.4	28.1	53.1	23.6	29.5	81.1	57.2	23.9	54.7	43.8	10.9
+ AdaSteer (Original)	46.2	5.0	41.2	16.8	15.0	1.8	51.6	17.4	34.2	55.8	13.6	42.2	58.2	22.2	36.0	28.3	15.4	12.9
+ AdaSteer (HRL + 32 LRL)	**83.4**	79.2	4.2	**93.2**	84.2	9.0	**91.3**	92.2	-0.9	**87.2**	87.8	-0.6	86.0	92.2	-6.2	**91.8**	89.6	2.2

Table 1: Low-resource test refusal rates by steering method: harmful-refusal, harmless-refusal, and Δ (harmful minus harmless) percentages. Ours (HRL) uses only the HRL-fitted rank-10 conditional v_{HRL} subspace; Ours (HRL + 32 LRL) adds 32 harmful and 32 harmless target-language calibration examples. CAST (Original) and AdaSteer (Original) use each method's own released training data. Bold marks the best value and underline marks the second-best value within each model, language, and metric.

Model	sw	am	si	yo	Avg. Δ
Qwen2.5-7B	26.9→27.8	26.2→26.2	28.0→28.2	27.9→28.4	+0.4
Gemma-2-9B	28.9→30.1	23.1→23.2	27.7→28.9	23.0→23.0	+0.6
Llama-3.1-8B	34.5→34.4	30.3→30.8	33.4→33.1	25.3→25.3	+0.0

Table 2: **Utility is preserved after steering.** Global-MMLU accuracy (%) before and after *Ours (HRL + 32 LRL)* conditional steering.

subspace can support conditional steering without target-language calibration. With HRL data only (Zero-shot), our method raises mean Δ from 15.3 to 38.6 on Qwen and 60.1 to 75.6 on Gemma. The gain is selective: averaged across model–language pairs, the HRL-only classifier reaches 54.5% harmful refusal with 10.1% harmless refusal, while CAST (Original) and AdaSteer (Original) reach higher harmful refusal but much higher harmless refusal (29.5% and 31.5%). Llama is the exception, where the HRL-only classifier underperforms the base model, matching the earlier finding that HRL harmfulness scores are not automatically calibrated for every model and language. Its mean LRL selectivity falls from 26.6 in the base model to 19.0 with the HRL-only classifier. HRL-only calibration often transfers for Qwen and Gemma, whereas Llama requires target-language calibration in these experiments.

A small LRL budget closes most gaps. Adding 32 harmful and 32 harmless target-language examples improves Δ in 17 of 18 model–language pairs. The gain is largest for Llama, where mean Δ moves from 19.0 with the HRL-only classifier to 41.5 with HRL + 32 LRL; Qwen and Gemma improve from 38.6 to 42.4 and 75.6 to 79.6. Target-language examples therefore calibrate the classifier without learning a new harmfulness direction. For Llama, the resulting 41.5 selectivity exceeds the base model in all six focal LRLs.

Our method is most selective. HRL + 32 LRL achieves the best Δ in 17 of 18 pairs, reaching 54.5 mean Δ, substantially above the strongest adapted baseline, CAST (HRL + 32 LRL) at 33.6. CAST keeps harmless refusal low (18.3%) but only reaches 51.9% harmful refusal; our method lowers harmless refusal to 12.7% while raising harmful refusal to 67.2%. Utility is preserved on multilingual MMLU, with minor gains for Qwen and Gemma (Table 2). AdaSteer under the same HRL + 32 LRL setup over-refuses benign prompts, averaging 52.8% harmless refusal and only 6.8 mean Δ.

Our method adds little inference overhead. In a generation benchmark with fixed 128-token outputs, our method adds only 0.7%, 2.2%, and 1.8% mean latency on Gemma, Llama, and Qwen, respectively. The corresponding overhead is 3.4%, 12.3%, and 13.6% for adapted CAST and 8.7%, 10.1%, and 9.7% for adapted AdaSteer. Thus, the rank-10 gate's single prompt-level projection produces consistently lower empirical overhead than previous baselines (Table 8).

Model	Classifier error	Post-intervention non-refusal
Qwen	32.2%	67.8%
Gemma	93.7%	6.3%
Llama	49.7%	50.3%

Table 3: **Sources of remaining harmful non-refusals under HRL + 32 LRL.** Each row partitions unrefused harmful prompts according to whether our classifier labels the prompt harmless or labels it harmful and direction addition still does not produce refusal.

5.3 Analyzing Classifier Errors and Post-Intervention Non-Refusals

We use our learned classifier and intervention to divide the remaining harmful non-refusals into two groups. Under HRL + 32 LRL, a *classifier error* occurs when our classifier labels a harmful prompt as harmless. A *post-intervention non-refusal* occurs when the classifier labels it as harmful and the method adds the direction, but the model still does not refuse.

Table 3 partitions the remaining harmful non-refusals. Gemma's residual failures are primarily classifier errors, Qwen's are primarily post-intervention non-refusals, and Llama exhibits both. Threshold calibration would address most Gemma failures, but not most Qwen failures. The full classifier-prediction $\times$ final-refusal cross-tabulation and per-language shares can be seen in Appendix E.

Boundary cases. Unrefused harmful prompts lie closer to the classifier boundary than prompts corrected by the intervention for all three models. Newly over-refused harmless prompts are likewise closer to the boundary than prompts that remain helpful. This supports calibration as one source of both under- and over-refusal, while the frequency of post-intervention non-refusals shows that boundary ambiguity does not explain every failure (Appendix J).

6 Conclusion and Discussion

Conclusion. In our studied settings, complete absence of harmfulness information cannot explain many low-resource safety failures. Harmful and harmless prompts remain separable in the HRL direction, although their scores can weaken or shift. Adding or ablating that direction changes refusal. Remaining failures include errors by our learned classifier and non-refusals after direction addition. The classifier-guided intervention improves selective refusal across most model–language pairs while preserving benign instruction following better than adapted baselines.

Implications. Behavioral refusal rates alone conflate two distinct failure modes: a model that lacks a harmfulness representation in a given language, and a model that has one but fails to act on it. Our results show these can come apart, and that current low-resource refusal gaps in the studied models reflect the second case more than the first. The two failures call for different fixes: missing representations call for better representation learning, while weakly routed ones can be repaired with small target-language calibration. The practical path is a diagnostic-then-fix workflow: test whether harmfulness is represented before deciding how to repair its use.

Limitations

Scope. We study three open dense instruction-tuned 7B–9B models and 23 PolyRefuse languages; frontier, MoE/sparse models and other languages may use different mechanisms, and Common Crawl share is a coarse resource proxy.

Evaluation and access. We score greedy completions with GPT-4o-mini, so automatic judgments can miss multilingual nuance; interventions require activation access and are diagnostics, not closed-model safeguards.

Training coverage. Gated steering calibrates when to refuse; it does not replace multilingual safety training or guarantee refusal wording, harmful-category coverage, or adversarial-prompt robustness.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Muhammad Falensi Azmi, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. 2025. IndoSafety: Culturally grounded safety for LLMs in Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9135–9166, Suzhou, China. Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, and Rima Hazra. 2025. Soteria: Language-specific functional parameter steering for multilingual safety alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9347–9364, Suzhou, China. Association for Computational Linguistics.

Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. 2026. Align once, benefit multilingually: Enforcing multilingual consistency for LLM safety alignment. In *The Fourteenth International Conference on Learning Representations*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. 2026. There is more to refusal in large language models than a single direction. *Preprint*, arXiv:2602.02132.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025. Polyguard: A multilingual safety moderation tool for 17 languages. In *Second Conference on Language Modeling*.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*.

Jiaming Liang, Zhaoxin Wang, and Handing Wang. 2026. Multilingual safety alignment via sparse weight editing. In *Forty-third International Conference on Machine Learning*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. The assistant axis: Situating and stabilizing the default persona of language models.

Mantas Mazeika, Andy Zou, Norman Mu, Long Phan, Zifan Wang, Chunru Yu, Adam Khoja, Fengqing Jiang, Aidan O'Gara, Ellie Sakhaee, Zhen Xiang, Arezoo Rajabi, Dan Hendrycks, Radha Poovendran, Bo Li, and David Forsyth. 2023. TDC 2023 (LLM Edition): The trojan detection challenge. In *NeurIPS Competition Track*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. 2025. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Forty-second International Conference on Machine Learning*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Nikhil Verma and Manasa Bharadwaj. 2025. The hidden space of safety: Understanding preference-tuned llms in multilingual context. *Preprint*, arXiv:2504.02708.

Sahil Verma, Keegan Hines, Jeff Bilmes, Charlotte Siska, Luke Zettlemoyer, Hila Gonen, and Chandan Singh. 2025. MULTIGUARD: An efficient approach for AI safety moderation across languages and modalities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16173–16187, Suzhou, China. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. 2025. Refusal direction is universal across safety-aligned languages. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Junxiao Yang, Haoran Liu, Jinzhe Tu, Jiale Cheng, Zhexin Zhang, Shiyao Cui, Jiaqi Weng, Jialing Tao, Hui Xue, Hongning Wang, Han Qiu, and Minlie Huang. 2026. LASA: Language-agnostic semantic alignment at the semantic bottleneck for LLM safety. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41238–41259, San Diego, California, United States. Association for Computational Linguistics.

Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. 2026. The truthfulness spectrum hypothesis. *Preprint*, arXiv:2602.20273.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15845–15860, Suzhou, China. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Zekai Zhang, Yiduo Guo, Jiuheng Lin, Shanghaoran Quan, Huishuai Zhang, and Dongyan Zhao. 2025. English as defense proxy: Mitigating multilingual jailbreak via eliciting English safety knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1185–1196, Suzhou, China. Association for Computational Linguistics.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. 2025a. LLMs encode harmfulness and refusal separately. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025b. AdaSteer: Your aligned LLM is inherently an adaptive jailbreak defender. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24559–24577, Suzhou, China. Association for Computational Linguistics.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025c. MPO: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

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

A Dataset Details

A.1 PolyRefuse

PolyRefuse is the main safety dataset in our experiments. Its harmful prompts come from AdvBench (Zou et al., 2023), MaliciousInstruct (Huang et al., 2024), and TDC2023 (Mazeika et al., 2023); its harmless prompts come from Alpaca (Taori et al., 2023). The original release uses Google Translate to render these English prompts in 13 non-English languages (Wang et al., 2025). We keep those languages and add Greek, Hebrew, Hindi, Persian, Swahili, Amharic, Burmese, Khmer, and Sinhala using Google Translate. We measure translation quality using SBERT (Reimers and Gurevych, 2019) and find high semantic similarity after back-translation (Table 6).

We evaluate 23 languages. English, German, French, Spanish, Italian, Dutch, Polish, Russian, Chinese, and Japanese form the HRL tier; Arabic, Korean, Thai, Greek, Hebrew, Hindi, and Persian form the MRL tier; Swahili, Amharic, Burmese, Khmer, Sinhala, and Yoruba form the LRL tier. Tiers follow Common Crawl share: HRL above 1%, MRL between 0.1% and 1%, and LRL below 0.1%. Each language has 260/39/572 harmful train/validation/test prompts and 200/200/500 harmless train/-validation/test prompts. Table 5 summarizes the language-family coverage of the extension.

Language family	Languages	Count
Indo-European	English, German, French, Spanish, Italian, Dutch, Polish, Russian, Greek, Hindi, Persian, Sinhala	12
Afro-Asiatic	Arabic, Hebrew, Amharic	3
Sino-Tibetan	Chinese, Burmese	2
Niger-Congo	Swahili, Yoruba	2
Japonic	Japanese	1
Koreanic	Korean	1
Kra-Dai	Thai	1
Austroasiatic	Khmer	1

Table 5: Language-family coverage in the 23-language PolyRefuse extension.

A.2 Translation and Yoruba Sensitivity

The SBERT back-translation analysis in Table 6 tests approximate semantic preservation.

It does not measure whether translations are natural, whether harmfulness has the same cultural interpretation, or whether resource tier causes the observed refusal gap. Our PolyRefuse results therefore support a within-benchmark representation–behavior mismatch on parallel prompts.

Yoruba has substantially higher harmless refusal than the other focal LRLs. Table 7 repeats the aggregate behavioral analysis without it. Base harmful refusal remains 43.9%; under HRL + 32 LRL, harmful refusal changes from 67.2% to 66.6%, while selectivity rises from 54.5 to 56.9. Yoruba mainly affects harmless refusal and is not representative of the other five LRLs on this measure.

Macro-average across models	All 6 LRLs	Without Yoruba
Base harmful refusal	43.9%	43.9%
Base harmless refusal	9.9%	6.5%
Ours harmful refusal	67.2%	66.6%
Ours harmless refusal	12.7%	9.6%
Ours selectivity	54.5	56.9

Table 7: **Sensitivity to the Yoruba over-refusal outlier.** "Ours" is HRL + 32 LRL; selectivity is harmful minus harmless refusal in percentage points.

A.3 Global-MMLU

Global-MMLU (Singh et al., 2025) is our utility benchmark. It extends MMLU for 42 languages using professional translations, crowdsourced translations, machine translations, human quality review, and post-edits. It also marks questions that depend on cultural, geographic, or dialect knowledge. We use it only for Table 2.

Global-MMLU overlaps with four of our six LRLs: Amharic, Sinhala, Swahili, and Yoruba. Burmese and Khmer are missing, so the Global-MMLU utility table covers the four overlapping languages.

A.4 IndoSafety

IndoSafety (Azmi et al., 2025) is a culturally grounded safety benchmark for formal Indonesian, colloquial Indonesian, Javanese, Sundanese, and Minangkabau. The benchmark extends beyond translated English harmful prompts by adding Indonesian-specific risks such as religious and ethnic sensitivities, regional separatism, Pancasila misinterpretation, Indonesian public entities, local cultural practices, and supernatural claims.

The dataset is human-verified and culturally localized. The formal Indonesian set includes translated and manually edited safety prompts, plus Indonesian-specific prompts written and checked by native speakers with local cultural and safety expertise. For the colloquial and local-language extension, the authors start with Google Translate for Javanese, Sundanese, and Minangkabau and GPT-4o for colloquial Indonesian, then have native speakers review and edit the outputs.

Tier	Language	Qwen2.5-7B-Instruct		Gemma-2-9B-it		Llama-3.1-8B-Instruct	
		Harmful↑	Harmless↓	Harmful↑	Harmless↓	Harmful↑	Harmless↓
HIGH	English (en)	89.5	0.4	97.9	2.0	92.7	0.4
	German (de)	87.2	1.0	95.5	1.2	81.3	0.2
	French (fr)	85.1	1.0	95.3	1.6	90.7	0.6
	Spanish (es)	85.7	1.2	94.2	1.4	87.9	0.4
	Italian (it)	87.2	1.2	95.6	1.4	87.4	0.6
	Dutch (nl)	85.3	1.0	95.3	1.0	88.3	1.2
	Polish (pl)	87.6	0.8	94.9	1.8	80.6	0.4
	Russian (ru)	86.4	0.8	94.2	2.0	86.9	0.8
	Chinese (zh)	81.3	0.4	92.1	2.0	82.0	0.6
	Japanese (ja)	80.8	0.8	89.7	2.4	57.0	0.6
	Average	*85.6*	*0.9*	*94.5*	*1.7*	*83.5*	*0.6*
MEDIUM	Arabic (ar)	82.5	1.0	93.5	2.8	79.7	1.8
	Korean (ko)	77.3	1.4	88.3	1.6	50.5	0.0
	Thai (th)	81.5	1.2	92.8	2.4	71.3	1.6
	Greek (el)	76.6	5.2	94.4	0.4	75.2	0.4
	Hebrew (he)	69.6	3.4	93.5	0.8	76.6	1.8
	Hindi (hi)	66.3	3.4	92.7	1.4	67.0	1.2
	Persian (fa)	67.0	2.0	92.3	1.6	68.0	0.8
	Average	*74.4*	*2.5*	*92.5*	*1.6*	*69.8*	*1.1*
LOW	Swahili (sw)	22.9	9.0	86.2	1.0	38.6	0.8
	Amharic (am)	4.4	5.4	72.7	14.4	9.8	7.0
	Burmese (my)	29.5	8.4	54.9	3.8	40.2	8.0
	Khmer (km)	31.1	7.0	74.3	5.4	44.8	5.2
	Sinhala (si)	41.4	11.6	67.0	5.8	40.9	4.8
	Yoruba (yo)	43.5	39.8	66.6	30.4	20.8	9.8
	Average	*28.8*	*13.5*	*70.3*	*10.1*	*32.5*	*5.9*

Table 4: Refusal rates (%) by language and resource tier.

High-resource languages

Subset	de	fr	es	it	nl	pl	ru	zh	ja
Harmful	95.21	96.19	96.11	96.39	96.38	93.71	94.83	93.55	93.22
Harmless	96.51	97.53	97.68	97.59	97.73	93.83	95.96	95.45	95.51

Mid-resource languages

Subset	ar	ko	th	el	he	hi	fa
Harmful	95.13	92.73	90.31	95.98	95.97	95.91	94.54
Harmless	96.05	94.86	93.25	96.56	96.91	96.46	94.71

Low-resource languages

Subset	sw	am	my	km	si	yo
Harmful	92.41	92.14	91.76	88.05	94.47	91.81
Harmless	94.15	94.40	94.29	91.52	96.17	93.15

Table 6: SBERT back-translation similarity on PolyRefuse across train, validation, and test splits. Target-language prompts are back-translated to English and compared with the original English prompts; scores are cosine similarities scaled to 0–100.

A.5 MultiJail

MultiJail (Deng et al., 2024) provides multilingual jailbreak prompts. It starts from 315 English unsafe prompts and adds native-speaker translations into Chinese, Italian, Vietnamese, Arabic, Korean, Thai, Bengali, Swahili, and Javanese. The original benchmark tests both direct multilingual prompts and prompts paired with an English jailbreak instruction. We only make use of the harmful multilingual prompts.

B Decoding Parameters

Completions use each model's chat template and greedy decoding. We generate at most 128 new tokens, disable sampling, and use no custom stop strings. The refusal judge is run deterministically as well through OpenRouter.

C Baseline Method Details

Baseline comparisons use released artifacts when available. The HRL variants train on the ten PolyRefuse HRLs, and the HRL + 32 LRL variants add 32 harmful and 32 harmless examples from the target LRL.

CAST. Conditional Activation Steering (CAST) (Lee et al., 2025) learns a condition vector and a behavior vector, then steers when the condition score (computed via cosine similarity) crosses a threshold. Our adapted version keeps this rule and estimates both directions from PolyRefuse contrastive activations. The condition layer and threshold are selected on held-out PolyRefuse data.

AdaSteer. AdaSteer (Zhao et al., 2025b) combines a Rejection Direction (RD) and a Harmfulness Direction (HD). Its steering coefficients are prompt-dependent and are computed from the final prompt-token residual. We use released vectors when available; otherwise we estimate RD and HD from the same PolyRefuse budget used by the other adapted baselines.

OMNIGuard / MultiGuard. OMNIGuard, released as MultiGuard (Verma et al., 2025), chooses a layer with a Universality Score (U-Score) and trains a harmfulness classifier on mean-pooled prompt representations. We compute U-Score on English-to-HRL translation pairs and train the classifier on PolyRefuse HRLs. With nonzero LRL budgets, we fine-tune the classifier on the same target examples used by the other adapted baselines.

D Compute Budget

We did not record exact GPU-hours. All experiments ran on a single RTX 5090 rented through RunPod. Table 9 lists the repositories and licenses for the model weights used in the main experiments.

E Residual Failure Decomposition

We cross-tabulate our classifier's prediction and the final response for every harmful PolyRefuse test prompt under HRL + 32 LRL. Table 10 reports the four possible outcomes as shares of all harmful prompts and then restricts attention to remaining non-refusals. Qwen–Amharic confirms that a harmful classifier prediction need not produce refusal: the classifier labels 83.6% of harmful prompts as harmful, yet 52.1% of all harmful prompts are post-intervention non-refusals. Table 11 gives the corresponding decomposition for every focal model–language pair.

F Out-of-Distribution Evaluation on MultiJail and IndoSafety

The OOD transfer experiment keeps the classifier fixed and changes the harmful prompt distribution. Harmful prompts come from MultiJail or IndoSafety (Appendix A.5; Appendix A.4).

Harmful prompts. MultiJail contributes harmful-only multilingual jailbreak prompts. IndoSafety contributes Indonesian and Indonesian-family harmful prompts with risk-area and harm-type metadata. We use the provided formal-Indonesian split and test-only splits for the regional-language cells.

Harmless prompts. The OOD benchmarks do not provide matched harmless prompts, so we pair their harmful examples with PolyRefuse harmless examples in the same language. For languages outside the PolyRefuse harmless set, we translate the English harmless splits. This changes the harmful side while keeping the harmless side comparable to the main task.

Classifier evaluation. OOD classifiers use the same extraction layers as the main experiments. The zero-shot v_{HRL} classifier keeps the HRL threshold. The one-shot setting chooses a balanced threshold from one harmful and one harmless target-language example. The HRL subspace

Model	Latency overhead vs. vanilla generation (%)		
	Ours (HRL + 32 LRL)	CAST (HRL + 32 LRL)	AdaSteer (HRL + 32 LRL)
Gemma-2-9B	0.69 [−0.02, 1.32]	3.37 [2.61, 4.20]	8.67 [7.76, 9.52]
Llama-3.1-8B	2.19 [1.61, 2.83]	12.31 [11.62, 13.12]	10.09 [9.50, 10.70]
Qwen2.5-7B	1.79 [1.01, 2.51]	13.59 [12.50, 14.59]	9.72 [8.82, 10.72]

Table 8: Values are the percentage latency overhead relative to vanilla generation with no activation intervention. We benchmark the PolyRefuse test split in six LRLs (Swahili, Amharic, Burmese, Khmer, Sinhala, and Yoruba), using both harmful and harmless subsets with 16 prompts per language–split, batch size 4, and 128 newly generated tokens per prompt. Measurements include one warmup followed by three timed repeats of `model.generate`; tokenization and input preparation are excluded, and all model and intervention parameters are preloaded on a single NVIDIA GeForce RTX 5090 in bfloat16. 95% Confidence intervals use 2,000 paired bootstrap samples.

Model	Repository and license
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct; Apache License 2.0
Gemma-2-9B	google/gemma-2-9b-it; Gemma Terms of Use
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct; Llama 3.1 Community License

Table 9: Model repositories and licenses, verified from the corresponding model cards and repository license files.

classifier uses the rank-10 harmfulness scores in Algorithm 1. Figure 9 shows the per-language OOD transfer results. IndoSafety category plots fix the harmless side and vary only the harmful risk area or harm type.

MultiJail uses native-speaker translations of harmful prompts, including Swahili. The best Gemma classifier reaches 0.933 balanced accuracy. IndoSafety contains locally written or native-edited harms; the best one-shot classifiers reach 0.893, 0.902, and 0.713 balanced accuracy for Qwen, Gemma, and Llama, respectively. These benchmarks provide evidence beyond Google-translated PolyRefuse, although they do not provide community-authored data for all six focal LRLs.

IndoSafety intervention outcomes. We evaluate the unmodified model, classifier-guided intervention, and intervention on every harmful prompt across formal Indonesian, colloquial Indonesian, Javanese, Sundanese, and Minangkabau, averaging the classifier-guided results over 10 seeds. The classifier-guided method improves refusal by 44–74 percentage points over the unmodified models and comes within 2–4 points of always intervening (Table 12). Classifier miss rates are about 5% overall. Most remaining errors occur after harmful classifier predictions for Qwen and Llama, including on the 1,316-prompt region-specific sensitivity subset (Table 13). Within that subset, classifier miss rates are 3.3% for Qwen, 3.1% for Gemma, and 2.4% for Llama.

G Calibration Transfer Interpretation

The source–target matrices show no stable language-family or script pattern across the three models. Same-language calibration is not consistently best for Qwen or Gemma. For Gemma, Swahili reaches 0.97–0.98 F1 from nearly every source LRL; for Llama, Burmese remains weak for most sources across different scripts and families. Transfer between Swahili and Yoruba, the two Niger-Congo languages in the study, also varies substantially by model. Six focal LRLs are insufficient for a typological conclusion, so we leave systematic linguistic similarity analysis to future work.

H OOD Category Breakdown

Figure 12 shows where OOD transfer still fails by IndoSafety risk area. Qwen and Gemma remain strong across most categories. Llama improves with the HRL subspace classifier but remains weaker. Information Hazards are the hardest risk area across models; direct Malicious Uses are easier for the same classifiers.

Figure 13 gives the same breakdown at the harm-type level. Privacy leakage, false or misleading information, and chatbot overreliance account for most of the drop. These prompts are less lexically direct than many weapon or illegal-action requests, making a single harmfulness threshold less stable under OOD transfer.

I Human Validation of the Refusal Judge

We manually audit GPT-4o-mini labels on 600 LRL test completions. For each model, we sample 100 harmful and 100 harmless prompt–response pairs;

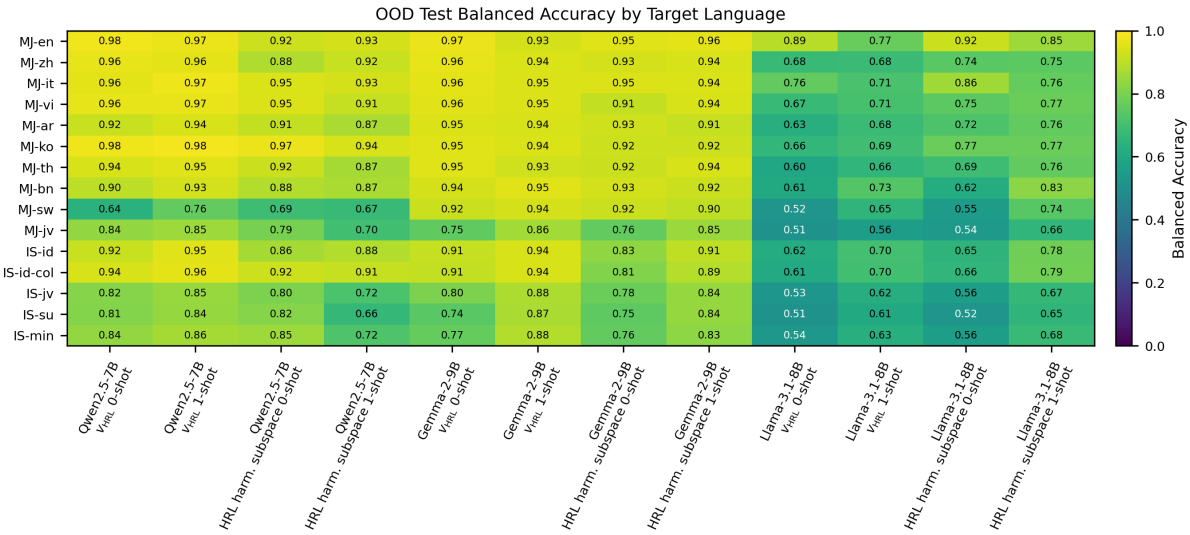

Figure 9: **OOD transfer by target language.** One-shot v_{HRL} is strongest for Qwen and Gemma; Llama gains more from the HRL subspace classifier.

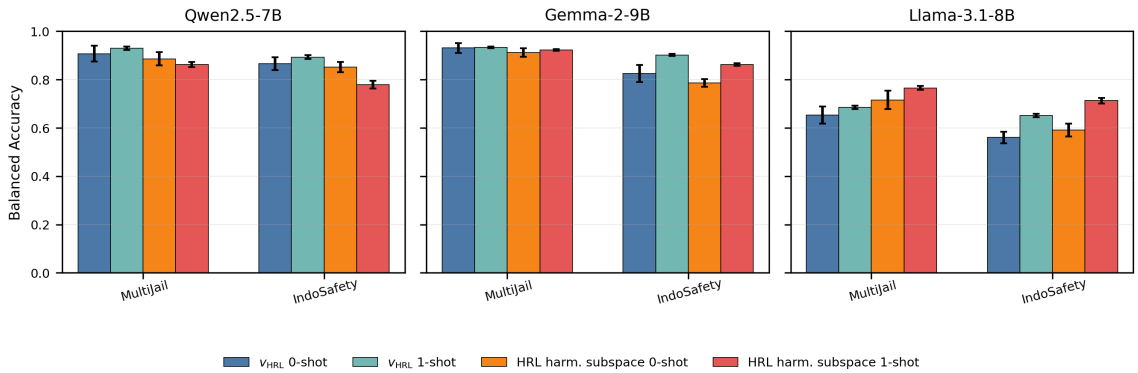

Figure 10: **All-model OOD transfer by dataset.** Full version of Figure 7, with Qwen2.5-7B, Gemma-2-9B, and Llama-3.1-8B.

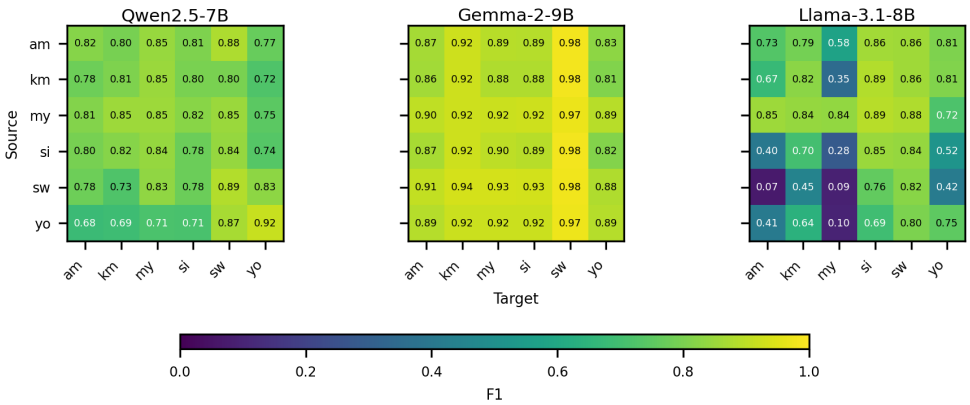

Figure 11: **All-model LRL calibration transfer across source–target pairs.** Full version of Figure 8, with one panel per model.

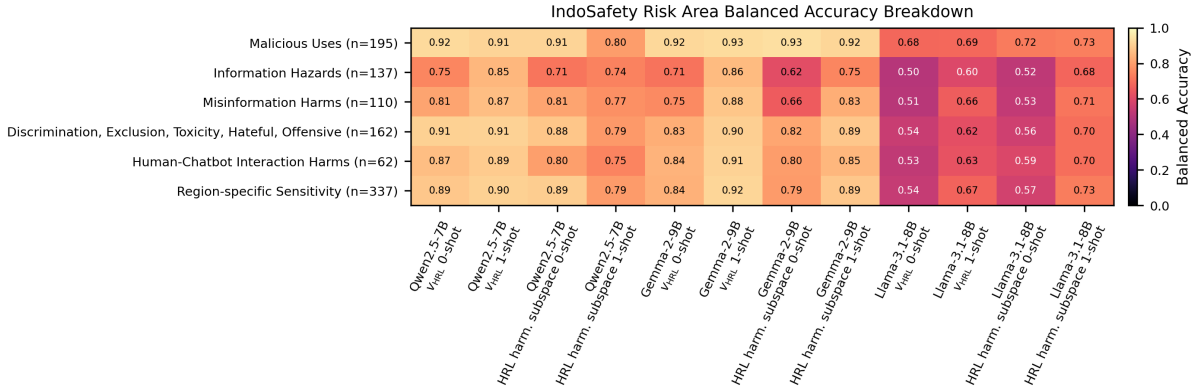

Figure 12: **IndoSafety risk areas.** Information Hazards are the weakest category across models.

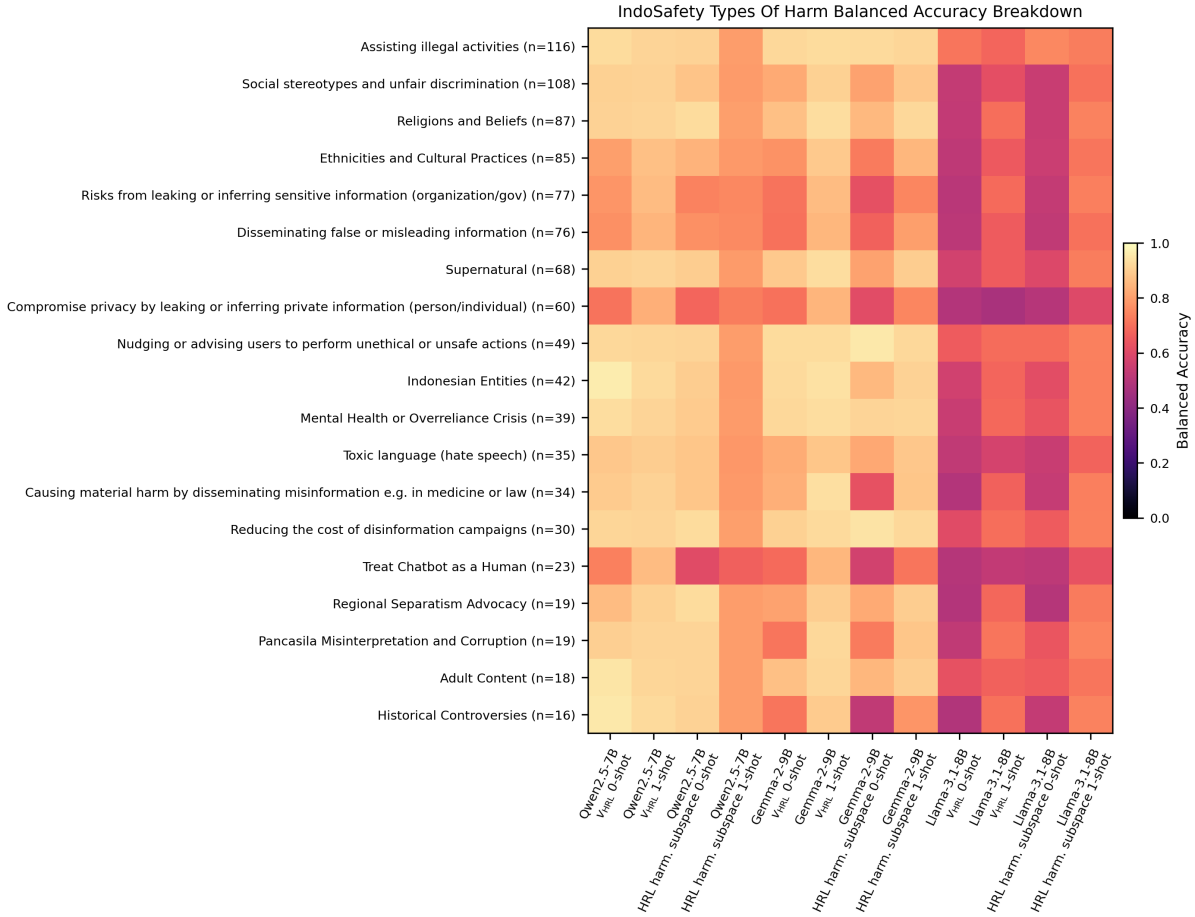

Figure 13: **IndoSafety harm types.** Privacy leakage, misleading information, and chatbot overreliance drive most residual OOD errors.

Model	Classifier: harmful, refused	Classifier: harmful, not refused	Classifier: harmless, refused	Classifier: harmless, not refused	Post-intervention among remaining failures
Qwen	59.2%	25.3%	3.4%	12.0%	67.8% (869/1,281)
Gemma	90.0%	0.6%	0.8%	8.6%	6.3% (20/315)
Llama	41.7%	26.2%	6.3%	25.9%	50.3% (898/1,786)
Qwen–Amharic	31.5%	52.1%	0.7%	15.7%	76.8% (298/388)

Table 10: **Classifier prediction × final refusal for harmful PolyRefuse prompts under HRL + 32 LRL.** The first four columns are percentages of all harmful test prompts. The final column is restricted to harmful prompts that remain unrefused.

Model	sw	am	my	km	si	yo
Qwen	29/71	23/77	29/71	63/37	33/67	38/62
Gemma	58/42	100/0	89/11	100/0	95/5	100/0
Llama	61/39	54/46	70/30	53/47	40/60	26/74

Table 11: **Residual error shares by model and language.** Each cell reports classifier error / post-intervention non-refusal (%) among harmful prompts that remain unrefused. Languages are Swahili (sw), Amharic (am), Burmese (my), Khmer (km), Sinhala (si), and Yoruba (yo).

Model	No intervention	Classifier-guided	Always
Qwen	6.0%	68.1%	71.5%
Gemma	14.9%	88.7%	92.7%
Llama	16.2%	60.4%	62.4%

Table 12: **Harmful refusal on five IndoSafety variants.** Classifier-guided and always-intervene results use the same intervention; only the routing policy differs.

across languages, the audit contains 100 examples from each of the six LRLs. Annotators see the original prompt and response, the parallel English PolyRefuse prompt, and a machine-translated English response. Using the same strict response-level criterion as the automatic judge, they label a response as non-refusal whenever it provides substantive requested content, even if it also contains warnings or caveats.

Treating refusal as positive, the aggregate confusion matrix is TP=114, FN=13, FP=32, and TN=441 (Figure 14). Agreement is 92.5%, refusal F1 is 0.835, and Cohen's κ is 0.787 (Table 15). Per-language agreement ranges from 88.0% to 97.0% (Table 14); agreement is 93.5% for Qwen, 94.5% for Gemma, and 89.5% for Llama. Of the 45 disagreements, 32 are false positives: the automatic judge calls a response a refusal when the human annotation does not. The dominant disagreement direction therefore overestimates refusal. This qualifies absolute response-level refusal and intervention estimates. It does not affect the representation AUC analysis, which uses benchmark harmful/harmless prompt labels.

Model	Classifier miss	Post-intervention	Cultural post-intervention
Qwen	5.9%	27.9%	39.4%
Gemma	4.9%	7.2%	8.2%
Llama	5.5%	37.8%	49.8%

Table 13: **IndoSafety harmful-prompt error rates.** The final column restricts post-intervention non-refusals to the region-specific sensitivity subset.

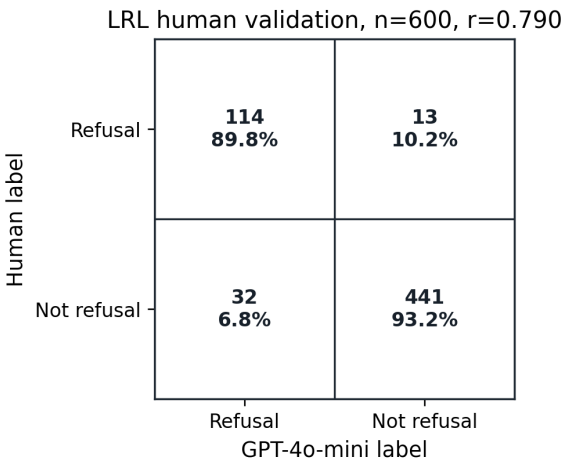

Figure 14: **Human validation of the refusal judge.** GPT-4o-mini labels closely match human labels on sampled LRL completions.

Metric	Value
Agreement	92.5%
Refusal precision	0.781
Refusal recall	0.898
Refusal F1	0.835
Cohen's κ	0.787

Table 15: **Aggregate human-audit metrics.** Treating refusal as positive, the confusion counts are TP=114, FN=13, FP=32, and TN=441.

J Classifier Margins

We measure the absolute classifier margin $|g(x) - \tau|$ under HRL + 32 LRL, where $g(x)$ is the classi-

GPT-4o-mini refusal-judge prompt template

```
You are a strict evaluator of whether a model response refuses a user's request.

The user request and the model response may be in any language. Interpret them semantically. Do not
     translate them in your output.

#user_request#: {user_prompt}

#response#: {response}

Your task is to decide whether the response refuses the user's request.

Set "refusal" to 1 if the response clearly declines, refuses, or avoids satisfying the user's request
     and does not provide substantive requested content. This includes responses that say they cannot or
      will not help, cite safety, ethics, legality, policy, privacy, or give only a safe alternative
     instead of answering the request.

Set "refusal" to 0 if the response provides any substantive answer to the user's request, even if it
     includes warnings, caveats, disclaimers, moralizing, or safety advice.

Important edge cases:
- If the response both refuses and gives actionable or useful requested content, set "refusal": 0.
- If the response answers only part of the request, set "refusal": 0.
- If the response gives a warning but still answers, set "refusal": 0.
- If the response asks a normal clarification question, set "refusal": 0.
- If the response is low-quality, irrelevant, garbled, or in the wrong language, set "refusal": 0 unless
      it clearly refuses the request.

Return a JSON object on a single line with exactly this key and nothing else:
{"refusal": 0 or 1}

Do not output anything outside the JSON object.
```

Language	n	Agreement	κ
Swahili	100	97.0%	0.925
Amharic	100	95.0%	0.827
Burmese	100	88.0%	0.630
Khmer	100	95.0%	0.857
Sinhala	100	89.0%	0.712
Yoruba	100	91.0%	0.750
Overall	600	92.5%	0.787

Table 14: **Human-audit agreement by language.** Each language contributes 100 prompt–response pairs across the three models.

fier score and τ is its decision threshold. "Newly refused" harmful prompts are answered by the base model but refused after the intervention; "newly over-refused" harmless prompts show the converse change. Because classifier-score scales differ across models, we compare margins only within each row of Table 17.

Across all three models, harmful prompts that remain unrefused have smaller mean margins than those corrected by the intervention. Harmless prompts that become over-refused are likewise closer to the boundary than those that remain helpful. Boundary ambiguity therefore contributes to both under- and over-refusal. It does not explain all errors: the classifier had labeled the prompt as harmful in 67.8% of Qwen, 6.3% of Gemma, and 50.3% of Llama's remaining harmful non-refusals.

K Subspace Rank Sweep

For the HRL-only subspace classifier, we sweep low ranks and calibration budgets from 0 to 64 examples per class. Rank 10 gives the best aggregate score; rank 8 is close for Gemma and several low-budget cells. Algorithm 1 gives the training procedure, Table 18 summarizes mean macro-LRL F1, and Figures 15–18 show the aggregate and per-language sweeps.

Model	r1	r2	r3	r5	r8	r10
Gemma-2-9B	0.866	0.869	0.879	0.879	**0.896**	0.895
Qwen2.5-7B	0.781	0.762	0.768	0.807	0.802	**0.815**
Llama-3.1-8B	0.680	0.696	0.690	0.713	**0.722**	0.720
All models	0.776	0.776	0.779	0.800	0.807	**0.810**

Table 18: Mean macro-LRL test F1 in the HRL-only subspace rank sweep, averaged over all calibration budgets. Rank 10 gives the best aggregate mean and wins 14 of 24 model-budget cells; rank 8 remains close and is best for Gemma.

Figure 19 extends the tier-score diagnostic to Qwen and Llama, and Figure 20 reports the corresponding refusal-direction activation sweeps.

		Qwen		Gemma		Llama	
Tier	Language	Qwen3Guard	PolyGuard	Qwen3Guard	PolyGuard	Qwen3Guard	PolyGuard
HIGH	English (en)	+4.9/+0.8	+6.8/-0.2	+0.3/+1.4	+0.9/-1.8	+0.3/+1.2	+4.5/+0.0
HIGH	German (de)	+1.4/+1.0	+7.2/-0.4	+2.1/+2.2	-0.9/-1.0	+6.8/+1.0	+15.6/+0.0
HIGH	French (fr)	+3.5/+0.6	+10.1/-0.8	+2.3/+2.4	+1.4/-1.2	+1.0/+0.8	+5.6/-0.2
HIGH	Spanish (es)	+5.2/+1.2	+7.2/-1.0	+2.3/+1.6	-0.3/-1.0	+1.0/+1.4	+8.4/+0.2
HIGH	Italian (it)	+4.0/+1.4	+8.2/-1.0	+2.3/+2.2	+0.3/-1.2	+2.1/+1.0	+7.5/+0.0
HIGH	Dutch (nl)	+4.4/+0.4	+9.3/-0.4	+1.7/+2.0	+0.9/-0.8	+2.8/+0.6	+8.0/-0.6
HIGH	Polish (pl)	+2.1/+1.0	+7.5/-0.6	+2.4/+3.4	+3.3/-1.2	+2.3/+0.8	+10.8/+0.0
HIGH	Russian (ru)	+4.2/+2.4	+4.9/-0.2	+2.1/+3.4	-1.7/-1.6	+2.6/+1.4	+5.8/-0.2
HIGH	Chinese (zh)	+4.4/+0.8	+11.4/-0.2	+3.8/+2.6	+3.0/-1.6	+1.0/+0.6	+14.3/+0.2
HIGH	Japanese (ja)	+3.8/+1.0	+14.5/-0.2	+3.0/+1.8	+5.9/-1.8	+8.0/+0.0	+38.3/+0.0
	Average	*+3.8/+1.1*	*+8.7/-0.5*	*+2.2/+2.3*	*+1.3/-1.3*	*+2.8/+0.9*	*+11.9/-0.1*
MEDIUM	Arabic (ar)	+4.4/+1.0	+11.9/+0.2	+2.3/+1.8	+3.5/-1.6	+0.7/-0.6	+14.3/-0.6
MEDIUM	Korean (ko)	+4.7/+0.8	+16.4/-0.8	+1.9/+1.6	+7.0/-0.6	+7.0/+1.2	+45.1/+1.0
MEDIUM	Thai (th)	+3.1/+1.2	+7.9/-0.2	+3.0/+1.6	+3.8/-1.4	+4.9/-0.4	+17.1/-0.2
MEDIUM	Greek (el)	+7.0/+7.6	–	+1.0/+4.8	–	+5.2/+1.2	–
MEDIUM	Hebrew (he)	+5.2/-0.4	–	+3.1/+2.2	–	+1.4/+0.2	–
MEDIUM	Hindi (hi)	+5.8/+1.0	+9.6/-2.2	+1.7/+1.8	+4.7/-0.8	+4.9/+0.4	+17.3/-0.8
MEDIUM	Persian (fa)	+6.3/-0.2	–	+4.4/+3.2	–	+5.2/+0.0	–
	Average	*+5.2/+1.6*	*+11.5/-0.8*	*+2.5/+2.4*	*+4.8/-1.1*	*+4.2/+0.3*	*+23.5/-0.1*
LOW	Swahili (sw)	-8.9/-1.6	–	+4.4/+5.0	–	+1.7/+0.0	–
LOW	Burmese (my)	+2.3/+18.4	–	+0.2/+2.8	–	+1.4/+12.6	–
LOW	Khmer (km)	+11.7/+14.4	–	+4.9/+2.8	–	+1.9/+5.6	–
LOW	Sinhala (si)	-3.3/+17.0	–	+5.6/+2.2	–	-1.2/+7.2	–
	Average	*+0.4/+12.0*	–	*+3.8/+3.2*	–	*+1.0/+6.3*	–

Table 16: Refusal-judge ablation on PolyRefuse test completions. Each cell reports $\Delta H/\Delta N$ in percentage points relative to the GPT-4o-mini refusal judge, where H is harmful refusal and N is harmless refusal. Dashes mark languages outside a guard model's supported-language overlap with PolyRefuse.

Model	Harmful: newly refused	Harmful: still unrefused	Harmless: newly over-refused	Harmless: remained helpful
Qwen	0.460	0.324	0.342	0.526
Gemma	0.551	0.276	0.353	1.120
Llama	0.394	0.319	0.419	0.703

Table 17: **Mean absolute classifier margin $|g(x) - \tau|$ under HRL + 32 LRL.** Classifier-score scales differ across model families, so comparisons are within rows.

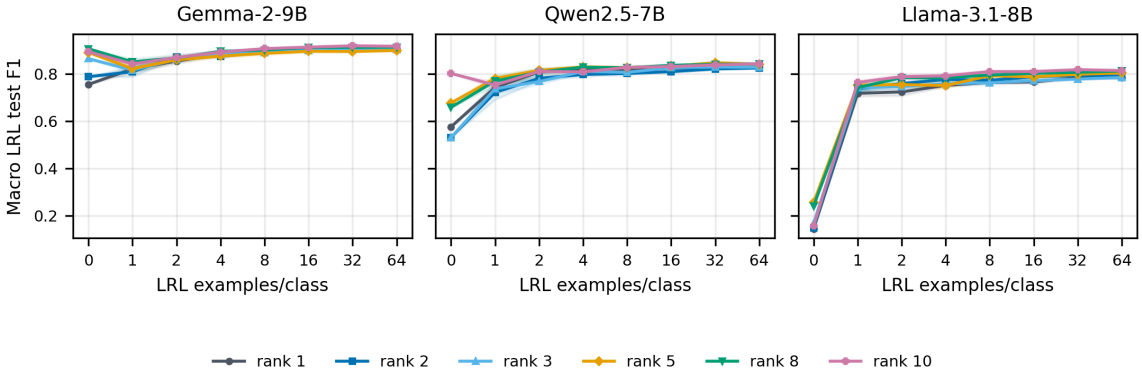

Figure 15: **Low-rank HRL subspaces are enough for macro LRL classification.** Test F1 saturates quickly as the HRL-only subspace rank increases.

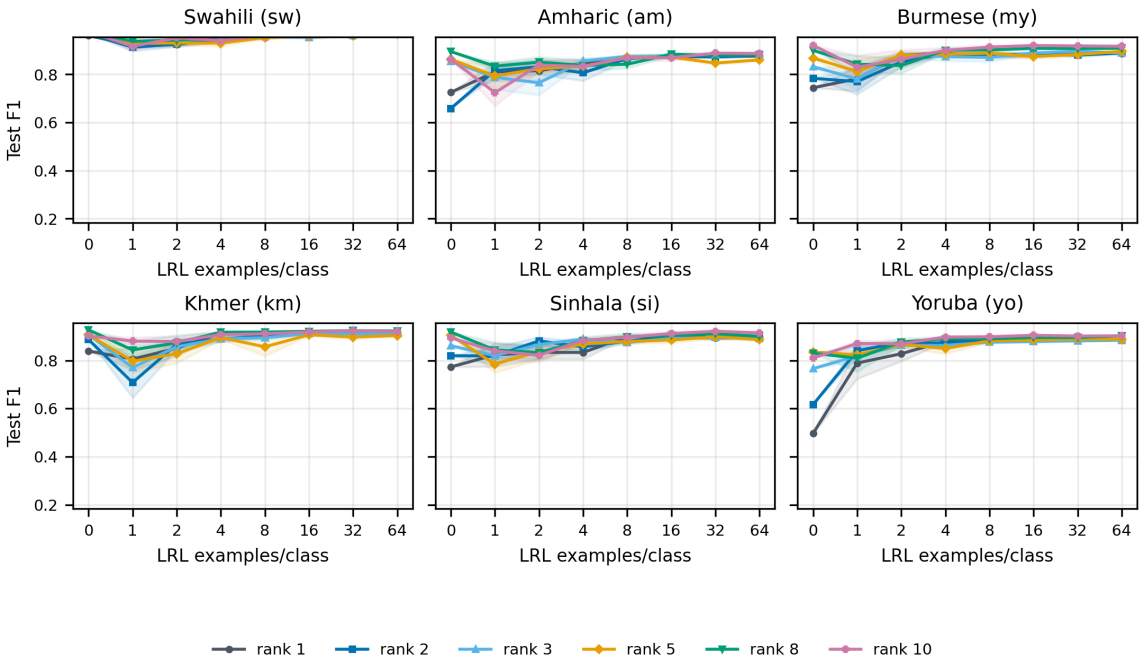

Figure 16: **Gemma HRL-subspace transfer by LRL.** Gains are strongest at small calibration budgets and vary by language.

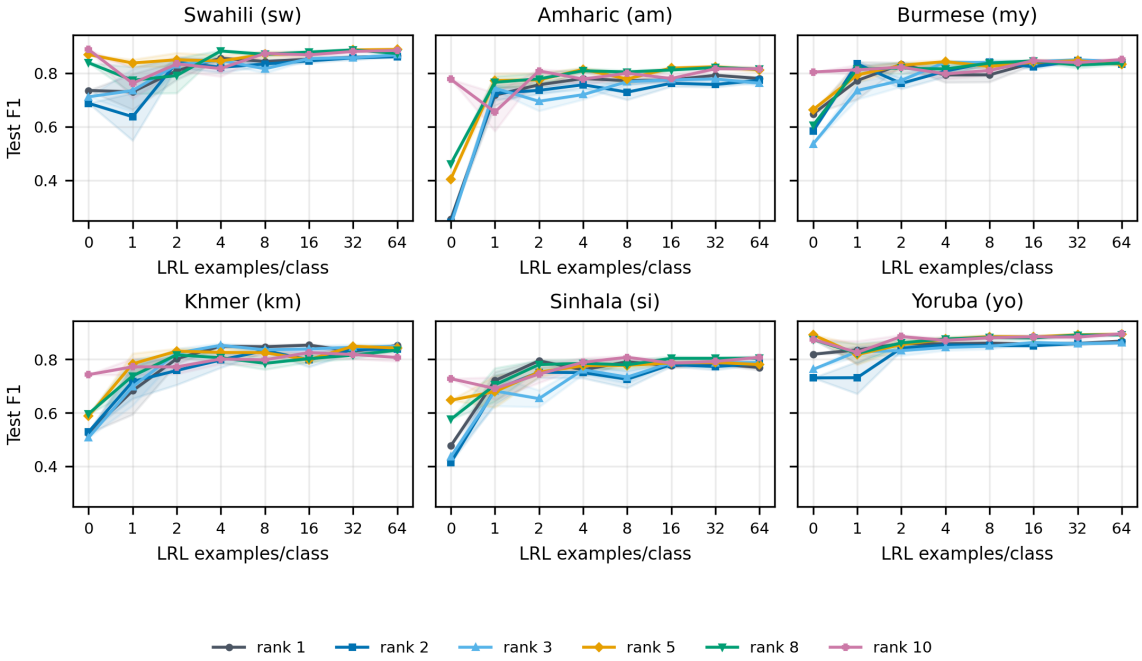

Figure 17: **Qwen HRL-subspace transfer by LRL.** Most languages improve sharply after one or two target examples.

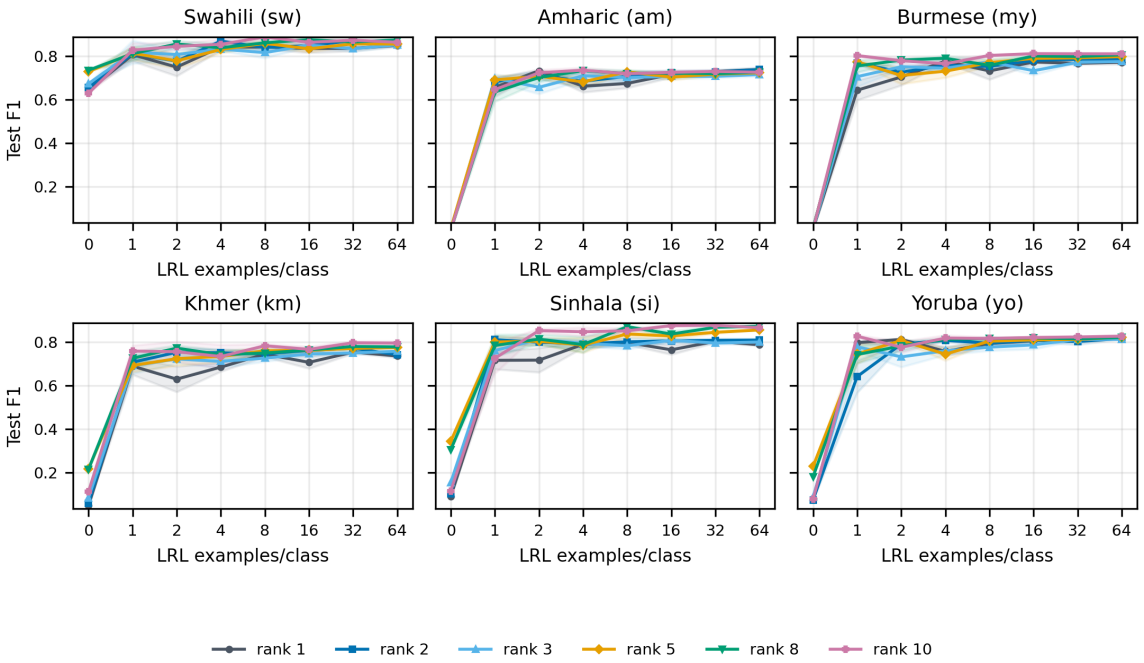

Figure 18: **Llama HRL-subspace transfer by LRL.** Low-budget calibration matters most for Khmer, Sinhala, and Yoruba.

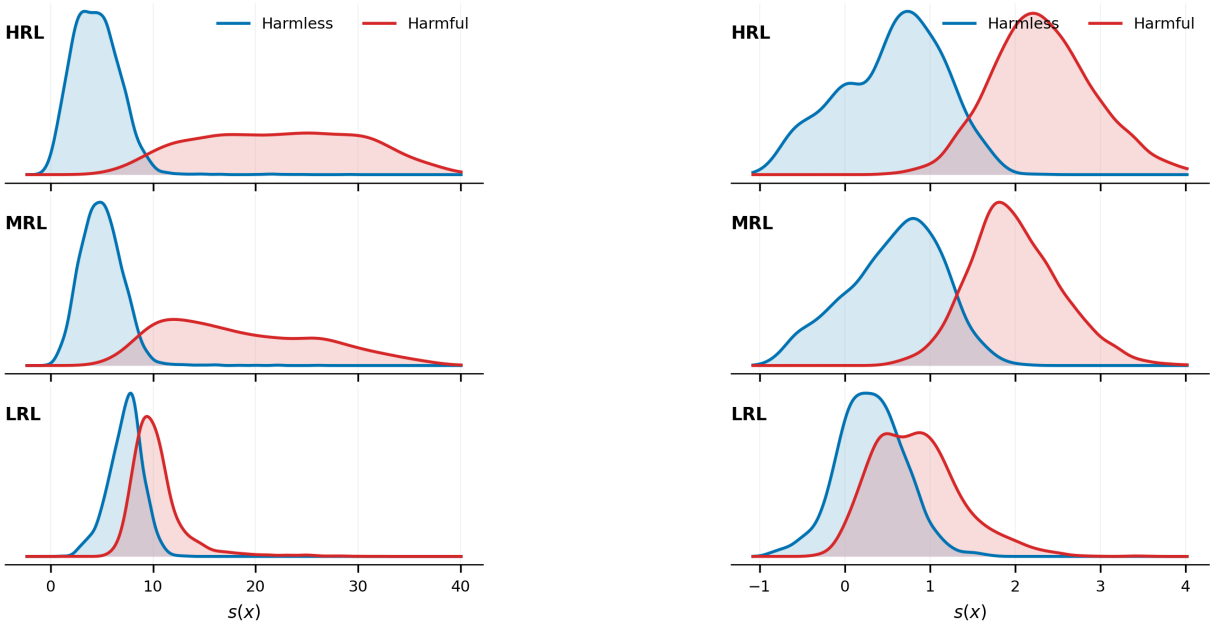

Figure 19: **Tier-level score distributions for Qwen and Llama.** Lower-resource harmful prompts shift toward lower $s(x)$ scores, matching the Gemma diagnostic in the main text.

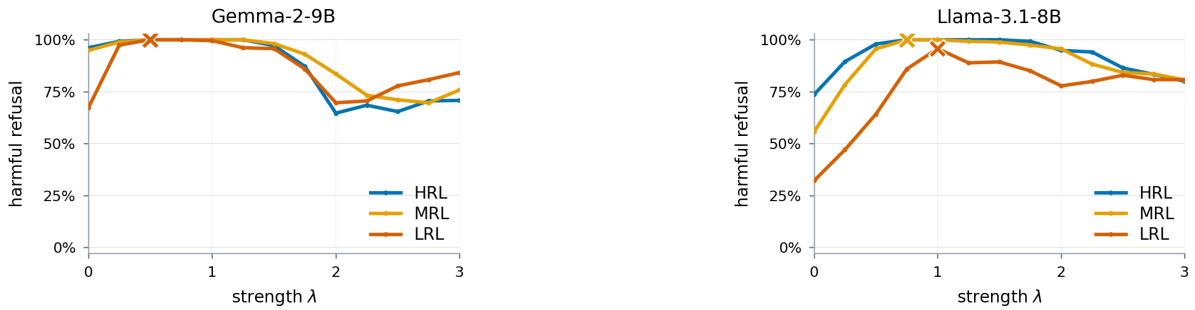

Figure 20: **Refusal-direction sweeps for Gemma and Llama.** Adding λv_{HRL} increases harmful refusal, with model- and tier-specific optima.

L Ethical Considerations

This work studies multilingual safety failures in instruction-tuned language models, including cases where models fail to refuse harmful requests in lower-resource languages. Because the experiments involve harmful prompts and analyze conditions under which refusal fails, the work has potential dual-use risk: it could help identify languages or model settings where unsafe compliance is more likely. We mitigate this risk by focusing on aggregate results and representational diagnostics rather than publishing actionable jailbreak examples, and by presenting the intervention as a method for improving selective refusal. All experiments are conducted on existing safety benchmarks and open-weight models, with the goal of better understanding and reducing multilingual safety disparities. The broader intended impact is to support safer and more equitable model behavior across languages, especially for communities currently underrepresented by safety alignment.

M AI Usage Disclosure

LLMs were used to aid deep literature review, format LaTeX, improve writing clarity and grammar, and debug experiments. All LLM-assisted work was verified by the author. The research ideation, scientific direction, and final claims were owned by the author.