

Lost in Translation: Cross-Lingual Transfer of Refusal Steering Vectors in Small Language Models¹

Pranamyadeshpande

GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik
pranamyadeshpande14@gmail.com

With Apart Research

Abstract

Safety alignment in large language models has been studied almost exclusively in English, yet these models are deployed to hundreds of millions of speakers of other languages. We investigate whether *refusal steering vectors* — directions in a model's residual stream encoding the decision to refuse a harmful request — transfer across languages, specifically from English to Hindi and Marathi, two South Asian languages with a combined speaker population exceeding 680 million. We extract English refusal directions from three small open-weight models (Llama-3.2-3B-Instruct, Gemma-2B-IT, Qwen-2.5-3B-Instruct) using contrastive activation analysis and ablate those directions during inference on harmful prompts in all three languages. We find that the English refusal direction produces **zero transfer to Marathi** across all three models ($\Delta = 0.000$), and only partial transfer to Hindi in one model (Qwen, $\Delta = +0.067$). More strikingly, base refusal rates for Hindi and Marathi harmful prompts are near zero across all models *before* any ablation, indicating that safety alignment was not learned for these languages in the first place. These findings reveal a concrete safety gap for South Asian AI deployments and motivate mandatory multilingual safety evaluation.

1. Introduction

Modern AI safety techniques for preventing harmful model outputs have been developed and validated predominantly in English. Refusal steering — the extraction and manipulation of internal model representations that govern whether a model refuses a harmful request — has emerged as a promising mechanistic approach to understanding model safety behaviour [1]. However, whether these mechanisms generalise to non-English languages remains an open question with direct practical consequences.

India alone accounts for over 680 million speakers of Hindi and Marathi, two of the most widely spoken languages in the world, yet little to no safety research targets these languages [6]. If the refusal mechanisms learned during English-language safety training do not operate in Hindi or Marathi, those speakers interact with AI systems whose safety guardrails are effectively disabled — not through any deliberate attack, but simply by using their native language.

We test this gap empirically. Motivated by Wang et al. [2], who claim that refusal directions in language models are universal across contexts, we ask whether this universality extends across languages to South Asian languages largely absent from prior refusal steering literature. We find it does not: English refusal vectors fail to suppress harmful outputs in Marathi in every tested model, and base refusal rates for Hindi and Marathi are already near zero before intervention.

Our main contributions are:

1. We provide the *first empirical study* of English refusal steering vector transfer to Hindi and Marathi across three open-weight models.
2. We show that base refusal rates for Hindi and Marathi harmful prompts are near zero across all models, revealing a safety gap independent of steering.
3. We show that English refusal directions provide *zero transfer to Marathi* across all three models, challenging claims of universal refusal representations.

¹Research conducted at The Global South AI Safety Hackathon, June 2026.

2. Related Work

Refusal steering and representation engineering. Zou et al. [1] introduced representation engineering, showing that model behaviour can be controlled by identifying and manipulating directions in the residual stream. Applied to refusal, this approach extracts a “refusal direction” by contrasting harmful and benign prompt activations. Wang et al. [2] claim these directions are universal across models and contexts; our work directly tests this universality across languages.

Multilingual jailbreaking. Prior multilingual safety work has shown that models can be prompted in low-resource languages to bypass safety filters [4, 5]. These works demonstrate the safety gap empirically via prompting; we investigate the mechanistic cause — whether the refusal direction itself is absent for these languages.

Dataset. We use CohereLabs/aya_redteaming [3], a multilingual red-teaming dataset covering English and Hindi. For Marathi, no public red-teaming dataset exists. We author native-speaker Marathi prompts, making this the first refusal steering study to include Marathi.

3. Methods

Models. We evaluate three instruction-tuned models: meta-llama/Llama-3.2-3B-Instruct, google/gemma-2b-it, and Qwen/Qwen2.5-3B-Instruct. All are inference-only; no fine-tuning is performed. These models represent different training lineages and varying degrees of multilingual data, allowing us to assess whether findings are model-specific.

Data. We use 60 harmful and 60 benign prompts per language (180 prompts total per language). English and Hindi harmful prompts are sampled from CohereLabs/aya_redteaming. Marathi prompts are authored natively from English red-teaming prompts translated by a native speaker. Benign prompts are factual questions in each language (e.g., “What is the capital of France?” / “ ?”).

Direction extraction. For each model, we register forward hooks on all transformer layers and collect the last-token hidden state for each prompt. We compute the per-layer contrastive direction as:

$$\mathbf{d}_i = \bar{\mathbf{h}}_i^{\text{harmful}} - \bar{\mathbf{h}}_i^{\text{benign}}$$

where $\bar{\mathbf{h}}_i$ is the mean hidden state at layer i . The best layer is selected as $\arg \max_i \|\mathbf{d}_i\|_2$ — the layer where the harmful/benign distinction is strongest. We set `padding_side='left'` to ensure the last-token position captures content rather than padding.

Ablation. During generation on test prompts, a forward hook at the best layer projects the refusal direction out of the residual stream in real time:

$$\mathbf{h} \leftarrow \mathbf{h} - (\mathbf{h} \cdot \hat{\mathbf{d}}) \hat{\mathbf{d}}$$

where $\hat{\mathbf{d}}$ is the unit-normalised direction. This removes the “refuse this” signal before it propagates to subsequent layers without modifying any model weights.

Evaluation. We measure the refusal rate — the fraction of harmful-prompt responses containing language-specific refusal phrases — before and after ablation. The primary metric is $\Delta = \text{rate_before} - \text{rate_after}$. Positive Δ indicates the direction suppressed refusals (transfer confirmed); $\Delta \approx 0$ indicates no transfer. We use a keyword-based refusal classifier with language-specific phrase lists (English: “I cannot”, “I refuse”; Hindi: “ ”; Marathi: “ ”), which is conservative and biases toward underestimating transfer, making positive findings more reliable.

4. Results

Table 1 summarises refusal rates before and after ablating the English-extracted refusal direction across all three models and languages. Figure 1 shows the Δ values as a heatmap. Figure 2 compares the peak English refusal direction norm across all three models, and Figure 3 contrasts the effect of EN vs HI direction ablation within Gemma-2B on English test prompts.

Table 1: Refusal rate before and after ablating the English refusal direction. $\Delta = \text{rate_before} - \text{rate_after}$; positive values indicate transfer. $n = 60$ harmful prompts per cell.

Model	Language	Before	After	Δ
Llama-3.2-3B	English	0.350	0.267	+0.083
	Hindi	0.117	0.117	0.000
	Marathi	0.017	0.017	0.000
Gemma-2B	English	0.900	0.483	+0.417
	Hindi	0.000	0.017	-0.017
	Marathi	0.000	0.000	0.000
Qwen-2.5-3B	English	0.600	0.367	+0.233
	Hindi	0.083	0.017	+0.067
	Marathi	0.017	0.017	0.000

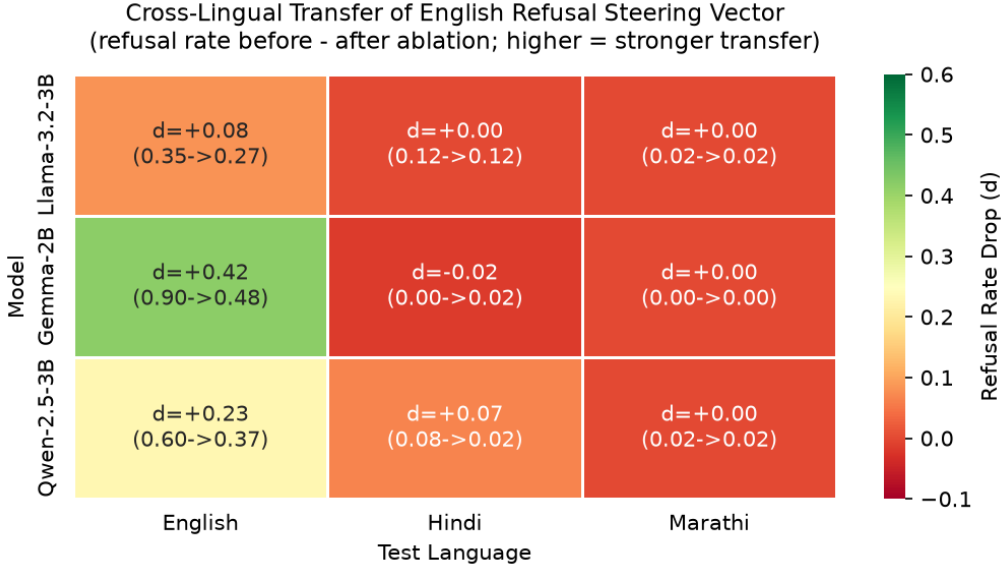


Figure 1: Heatmap of refusal rate drop (Δ) after ablating the English refusal direction across all three models and languages. Green = strong transfer (refusal suppressed); red = no transfer. Marathi shows zero transfer across all models; Hindi transfer is partial and model-dependent.

Finding 1: English refusal direction does not transfer to Marathi. Across all three models, $\Delta = 0.000$ for Marathi. The English refusal direction has no measurable effect on Marathi refusal behaviour. This directly contradicts universal refusal direction claims [2] for South Asian low-resource languages.

Finding 2: Transfer to Hindi is partial and model-dependent. Qwen shows a 6.7 pp drop in Hindi refusal rate, suggesting partial cross-lingual transfer. Llama shows zero transfer to Hindi. This divergence suggests that the extent of multilingual pre-training data may modulate whether a shared refusal representation is learned.

Finding 3: Base refusal rates are near zero for Hindi and Marathi. Before any ablation, harmful Hindi and Marathi prompts are refused at 0–11.7%, versus 35–60% for English. This gap exists *before* any intervention — these models have not internalised harmful-content recognition for South Asian languages at all.

Finding 4: English refusal signal strength varies widely across models but does not predict South Asian safety. As shown in Figure 2, Qwen-2.5-3B has the largest peak English refusal direction norm

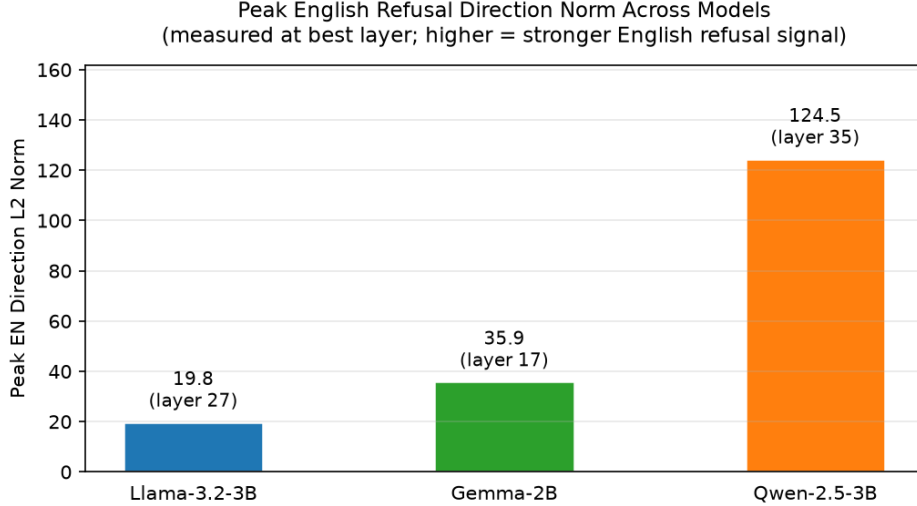


Figure 2: Peak L2 norm of the English refusal direction at the best layer for each model. Qwen-2.5-3B has the strongest English refusal signal (124.5 at layer 35), followed by Gemma-2B (35.9 at layer 17) and Llama-3.2-3B (19.8 at layer 27). Despite Qwen’s much stronger signal, its Marathi refusal rate before ablation is only 1.7%, confirming that a strong English direction does not imply any South Asian language alignment.

(124.5), consistent with its highest English refusal base rate (60%). Yet its Marathi base refusal rate is only 1.7%, and Marathi transfer is zero — a six-fold difference in English alignment strength produces no cross-lingual benefit. Figure 3 further demonstrates that even within Gemma-2B, the Hindi refusal direction is far weaker at suppressing refusals ($\Delta=+0.083$) than the English direction ($\Delta=+0.417$) when tested on English prompts, confirming that the two directions encode largely separate refusal representations in the same model.

5. Discussion and Limitations

Our results paint a consistent picture: safety alignment in these models is English-centric. The English refusal direction is strongest at the final transformer layers, but does not activate when processing South Asian language inputs. Practically, this means an attacker need only prompt in Marathi to bypass safety filters that would block an equivalent English request.

The near-zero *base* refusal rates for Hindi and Marathi are the most policy-relevant finding. Ablation tests whether an English vector can suppress refusals in another language; but if the model rarely refuses in that language to begin with, the safety failure is upstream of the direction — the model’s safety training never reached these languages. This suggests that the solution cannot be post-hoc vector transfer; language-specific safety training is required.

Limitations. Our dataset is small ($n = 60$ per language), which limits statistical power. The keyword-based refusal classifier may miss refusals in novel phrasings and may over-count refusals for responses that use refusal language while still providing harmful content. We evaluate only 2–3B parameter models; larger models with more documented multilingual training may show different transfer patterns. Marathi prompts are translations of English red-teaming prompts, which may not capture the full distribution of native harmful requests. The Llama Hindi-direction condition was lost to a Colab session disconnect and is excluded from this analysis.

Future Work. Language-specific refusal direction extraction for Hindi and Marathi, testing on more South Asian languages (Tamil, Telugu, Bengali), and evaluation on larger multilingual models (Aya, Navarasa) are natural extensions. From a safety policy perspective, our results motivate mandatory multilingual safety evaluation prior to deployment in South Asian markets.

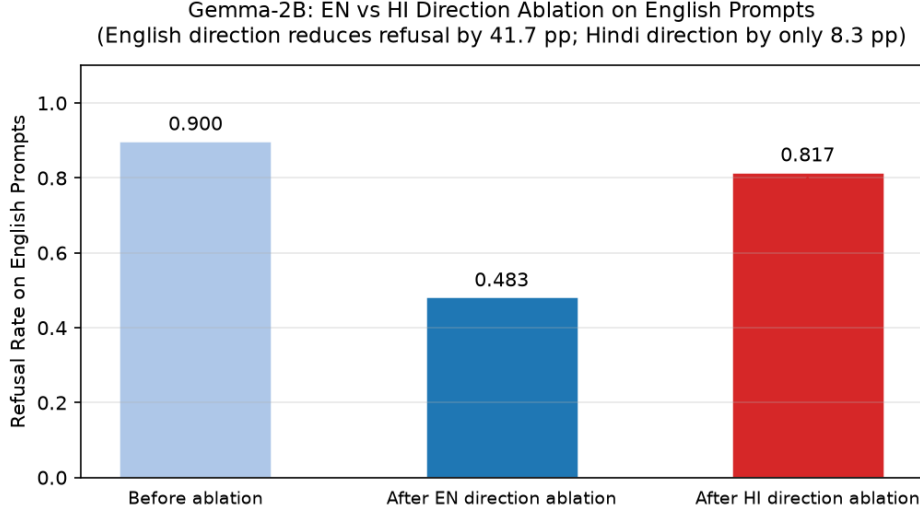


Figure 3: Refusal rate on English prompts for Gemma-2B before ablation and after ablating the English vs. Hindi refusal direction. Ablating the English direction drops the refusal rate from 0.900 to 0.483 ($\Delta=+0.417$), while ablating the Hindi direction yields only a marginal drop to 0.817 ($\Delta=+0.083$) — a five-fold difference in effectiveness. This suggests the English and Hindi directions encode largely separate refusal representations within the same model.

6. Conclusion

We tested whether English refusal steering vectors transfer to Hindi and Marathi across three small language models. The answer is largely no: English refusal directions do not suppress harmful outputs in Marathi in any tested model, and transfer to Hindi is partial at best. More critically, base refusal rates for Hindi and Marathi harmful prompts are near zero before any intervention — safety alignment was not learned for these languages in the first place. For the 680 million Hindi and Marathi speakers who interact with these systems, this is a concrete and measurable safety gap. Safety evaluation must extend beyond English.

Code and Data

- **Code repository:** <https://github.com/Pranamy16/GSAIH>
- **Data:** CohereLabs/aya_redteaming (EN, HI via HuggingFace); native-speaker Marathi prompts included in repository.

LLM Usage Statement

Claude Code was used to assist with code scaffolding, debugging, and drafting sections of this report. All experimental design, Marathi prompt authoring, and result interpretation were performed independently. All experimental results were verified by running the pipeline.

References

- [1] Zou, A., Long, Z., Turner, E., Mu, T., Basart, S., Li, L., ... & Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405.
- [2] Wang, Z., et al. (2025). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. NeurIPS 2025.
- [3] Aryabumi, V., et al. (2024). *Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning*. CohereLabs. https://huggingface.co/datasets/CohereLabs/aya_redteaming
- [4] Yong, Z. X., Menghini, C., & Bach, S. H. (2023). *Low-Resource Languages Jailbreak GPT-4*.

arXiv:2310.02446.

[5] Deng, Y., et al. (2023). *Multilingual Jailbreak Challenges in Large Language Models*. arXiv:2310.06474.

[6] Ahuja, S., et al. (2023). *MEGA: Multilingual Evaluation of Generative AI*. EMNLP 2023.