

One Direction, Many Languages: Causal Cross-Lingual Refusal Transfer Across Small Open Models

Bhoomika Gupta^{1,*} Aradhya Goel^{2,*}

¹San Jose State University ²Indian Institute of Technology Kanpur
bhoomika.gupta@sjsu.edu aradhyag24@iitk.ac.in

Abstract

It is easier to jailbreak a language model in Hindi or Vietnamese than in English, but we still know very little about why, at the level of what happens inside the model. There is also a puzzle in the recent literature. One line of work says the “refusal direction” (a single direction in activation space) carries over from one language to another almost perfectly. Another line of work, using sparse autoencoders, says the features that look like “refusal” in English often do not mean the same thing in another language. We tried to settle this on three small open models, GEMMA-2-2B-IT, LLAMA-3.2-1B-INSTRUCT, and QWEN2.5-1.5B-INSTRUCT, using 142 hand-written, topic-matched prompt pairs per language in English, Hindi, and Vietnamese, an external GPT-4o-class judge to score refusals, and bootstrap and Wilson confidence intervals.

Three things came out of it. First, the dense refusal direction really does line up across languages: cosine similarity between the per-language directions runs from about 0.85 to 0.97, and an English-derived direction still separates harmful from harmless prompts in the other languages at held-out AUROC between 0.92 and 0.98. Second, and this is the part we care about most, the same direction also works causally. If we subtract the English direction during generation, refusal drops not just in English but in Hindi and Vietnamese too, and this holds in all three models, not just one. The English drop is large everywhere (0.77 to 1.00), Vietnamese is clearly affected (0.37 to 0.73), and Hindi moves as well, though by less and with more noise (0.17 to 0.67). Third, the sparse-feature story does not reproduce at this scale: cross-lingual feature overlap is no lower than the overlap we get within a single language, so we cannot claim the features drift. Along the way we also found that a keyword-based refusal detector agrees with the judge only at chance ($\kappa \approx 0$), and that a model scoring its own Vietnamese outputs systematically mislabels them. Across these small models, the refusal direction transfers across languages both in geometry and in behaviour, while the sparse features are simply a noisier probe.

1 Introduction

Safety-tuned language models refuse harmful requests, but they do not do it evenly across languages. A prompt that gets refused in English often slips through when it is written in a lower-resource language. For South and Southeast Asia, where hundreds of millions of people speak Hindi and Vietnamese, that gap is a real deployment problem. If we want to understand why the refusal fails in some languages, a good place to start is what the model is actually doing on the inside.

*Both authors contributed equally

Two lines of interpretability research seem to pull in opposite directions. The first finds that refusal lives along a single residual-stream direction that carries over across safety-aligned languages (Arditi et al., 2024; Wang et al., 2025). The second, working with sparse autoencoders, finds that a feature signalling “refusal” in English need not keep that meaning in another language. We test all three levels (geometry, causal mechanism, and the sparse features) on small open models, and we run the same test on three different models rather than one, which is something the earlier single-model studies could not do.

Contributions.

1. A controlled multilingual probe of refusal from 142 hand-curated, topic-matched minimal pairs per language (English, Hindi, Vietnamese), released with code.
2. Evidence that the dense refusal direction transfers **geometrically** across the three languages, with confidence intervals.
3. Evidence that it also transfers **causally**: ablating the English direction reduces Hindi and Vietnamese refusal, and this **replicates across three model families**, which is the central result.
4. A **negative result** for sparse features: no significant cross-lingual SAE drift at this scale, with the within-language baseline making the null interpretable.
5. A **measurement caution**: keyword refusal classifiers agree with an external judge only at chance, and self-judging misclassifies Vietnamese refusals; we use an external judge throughout.

2 Related Work

Refusal directions and steering. Arditi et al. (2024) show refusal in chat models is mediated by a single difference-in-means direction; ablating it bypasses refusal and amplifying it induces over-refusal. Wang et al. (2025) report cross-lingual transfer of this direction across safety-aligned languages. We reproduce the geometric transfer and, using an external behavioural judge, establish that the causal effect transfers as well, across multiple models.

Sparse autoencoders and feature universality. SAEs decompose activations into sparse, more interpretable features (Lieberum et al., 2024). Whether “refusal features” are stable across languages is debated. We quantify cross-lingual feature overlap against a within-language baseline and find no reliable drift at our scale.

Evaluating refusal. Keyword matching conflates a change in refusal *wording* with a change in refusal *behaviour*, and same-model self-judging inherits the model’s own multilingual blind spots. We measure both failure modes and adopt an external judge.

3 Method

Models and SAEs. GEMMA-2-2B-IT (26 layers; with Gemma Scope residual SAEs, width 16k), LLAMA-3.2-1B-INSTRUCT (16 layers), and QWEN2.5-1.5B-INSTRUCT (28 layers), loaded in TransformerLens on a single GPU.

Model	Geometry (en→)			Refusal drop, subtract En dir. (judge)		
	cos en/hi, en/vi	transfer en/hi, en/vi	layer	en	hi	vi
GEMMA-2-2B-IT	0.955 / 0.967	0.944 / 0.963	12	0.83	0.67	0.67
LLAMA-3.2-1B	0.85 / 0.87	0.92 / 0.98	8	0.77	0.37	0.37
QWEN2.5-1.5B	0.53 / 0.94	0.76 / 0.94	15	1.00	0.17	0.73

Table 1: Dense geometry and judge-scored causal refusal drop (baseline – subtract) from ablating the *English* direction, per model. The direction transfers causally to Hindi and Vietnamese in all three families; the effect is strongest in English and Vietnamese, weaker and noisier in Hindi. Baselines: Gemma en/hi/vi = 0.97/0.87/0.80; Llama = 1.00/0.63/0.83; Qwen = 1.00/0.33/0.87. Qwen’s Hindi baseline of 0.33 is low, so its 0.17 drop should be read as partial evidence given the limited headroom.

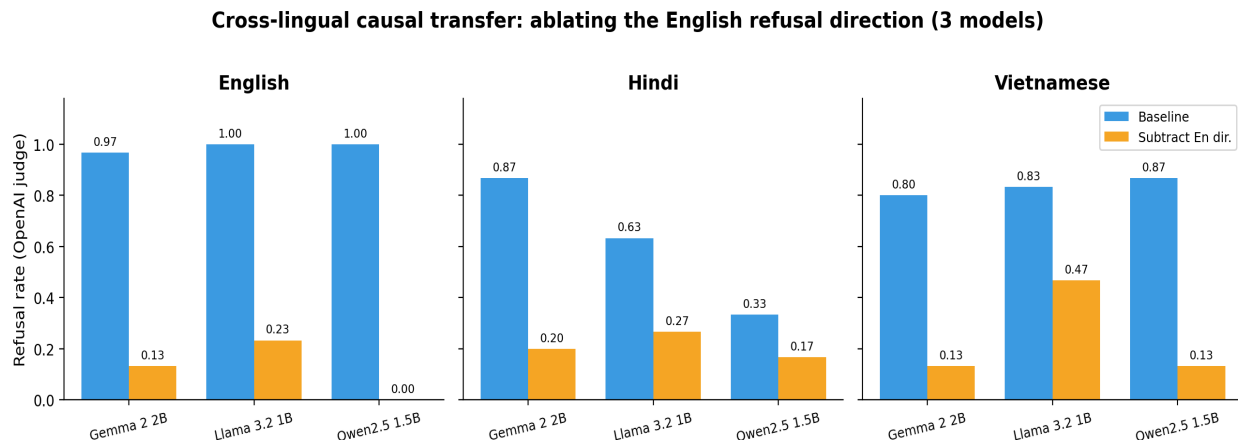


Figure 1: Cross-lingual causal transfer of the English refusal direction, replicated across three models. Subtracting the English direction collapses refusal in English, Hindi, and Vietnamese for all three models (OpenAI judge).

Topic-matched minimal pairs. To avoid confounding refusal with topic, we author 142 minimal pairs per language across 15 harm categories (fraud, theft, hacking, privacy, harassment, defamation, forgery, cheating, financial wrongdoing, piracy, impersonation, discrimination, break-in, surveillance, evasion). Each harmful prompt has a benign twin sharing topic, verb, and structure, differing only in refusal-worthiness (e.g. “pick a lock to break into someone’s house” vs. “pick a lock when I’m locked out of my own house”), and is a refusal-*eliciting* request rather than operational content. We use a 40/100 train/test split. Hindi and Vietnamese were produced from English with NLLB-200 (NLLB Team, 2022) and audited by native speakers (audit ongoing; english and hindi prompts audited completely).

Direction, geometry, features. The refusal direction is the unit difference-in-means of last-token residuals, with the analysis layer chosen by own-language train AUROC (layer 12 for Gemma). We report pairwise cosine and held-out *transfer AUROC* (one language’s direction scoring another’s prompts), and compare top-50 SAE feature sets across languages by Jaccard overlap against a within-language split-half baseline.

Causal intervention and judging. We subtract the English direction from the residual stream during generation (and, as a control, amplify it, `add_2x`), and measure the change in refusal rate.

Generations are scored by an external GPT-4o-class judge with a graded schema (refusal / safe-redirect / compliance / unclear), judging behaviour by meaning in each language. We additionally run an Arditi-style causal direction selection and strength sweep. The sweep (layers 9, 12, 13, 16; strengths 1 to 4; self-judged on $n=12$) confirmed layer 12 at strength 1 as the most consistent configuration. Higher strengths and other layers did not improve the cross-lingual effect and sometimes degraded it.

Statistics. Headline geometry/Jaccard metrics carry 1,000-sample bootstrap 95% CIs. Intervention rates carry Wilson 95% CIs and Classifier agreement is Cohen’s κ .

4 Results

4.1 Dense geometry transfers

For Gemma, each language’s own direction tells harmful prompts apart from harmless ones almost perfectly (AUROC 0.999 in English, 0.936 in Hindi, 0.985 in Vietnamese). The three directions are nearly collinear: cosine similarity is 0.955 between the English and Hindi directions, 0.967 between English and Vietnamese, and 0.965 between Hindi and Vietnamese. They also carry over across languages. Applied to held-out prompts in the other two languages, the English direction still separates harmful from harmless at AUROC 0.944 in Hindi and 0.963 in Vietnamese. The same pattern holds for all three models (Table 1, cosine and transfer columns), with Qwen’s Hindi direction the one clear outlier (cosine 0.53 with English).

4.2 The direction transfers causally, across three models

Table 1 reports, per model and language, the judge-scored refusal rate at baseline and after subtracting the *English* direction. In every model, subtraction reduces refusal not only in English but also in Hindi and Vietnamese. The English effect is large everywhere (drop 0.77 to 1.00), Vietnamese is substantial (0.37 to 0.73), and Hindi is positive but smaller and noisier (0.17 to 0.67). For Gemma, the post-subtract rate stays well below baseline everywhere: en 0.97→0.13 [0.05, 0.30], hi 0.87→0.20 [0.10, 0.37], vi 0.80→0.13 [0.05, 0.30] (brackets are Wilson 95% intervals on the post-subtract rate). The same holds for Llama-3.2-1B (en 1.00→0.23 [0.12, 0.41], hi 0.63→0.27 [0.14, 0.44], vi 0.83→0.47 [0.30, 0.64]) and Qwen2.5-1.5B (en 1.00→0.00 [0.00, 0.11], hi 0.33→0.17 [0.07, 0.34], vi 0.87→0.13 [0.05, 0.30]), where Qwen’s Hindi is the weakest leg given its low 0.33 baseline. The amplification control `add_2x` leaves refusal at ceiling (≥ 0.97 in every condition), so it tells us little given how high the baselines already are. When we read the actual generations, the responses under subtraction are genuine compliance (the model writes the impersonation notice, or lists the information-gathering steps), not broken or repetitive text.

Pair	Cross-lingual Jaccard	Within-language baseline
en–hi	0.56 [0.49, 0.64]	en 0.59, hi 0.67, vi 0.56
en–vi	0.64 [0.56, 0.72]	
hi–vi	0.64 [0.56, 0.72]	

Table 2: Top-50 SAE feature overlap (Gemma, layer 12) with bootstrap 95% CIs. Cross-lingual Jaccard overlaps every within-language baseline; no significant drift is detected at this scale.

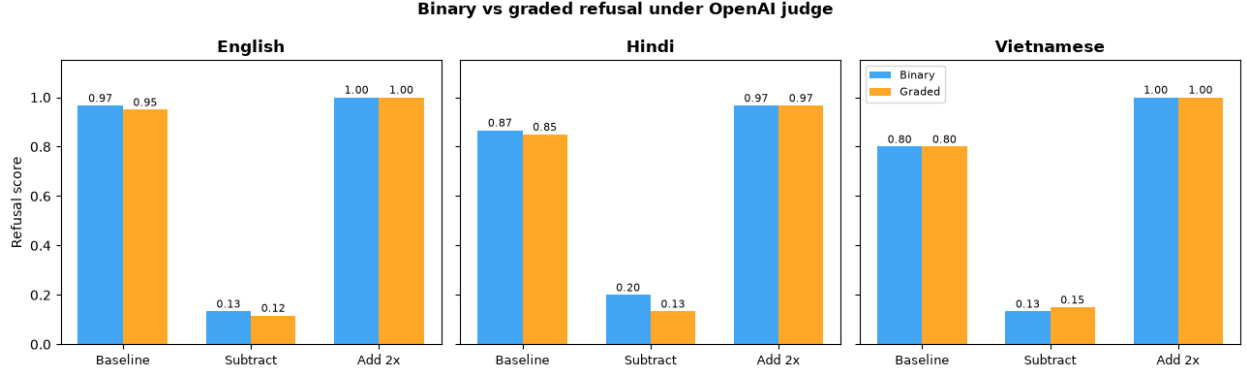


Figure 2: Binary vs. graded refusal under the external judge (Gemma). The two metrics agree closely across all conditions and languages, indicating the judging is robust to the binary/graded choice. The `add_2x` amplification control stays at ceiling.

4.3 No significant SAE drift at this scale

Table 2 contrasts cross-lingual top-50 SAE feature overlap with the within-language baseline. Unlike the dense direction, the sparse features do *not* show reliable cross-lingual drift. Cross-lingual Jaccard (en-hi 0.56 [0.49, 0.64]) overlaps the within-language baseline (~ 0.56 – 0.67). The within-language baseline is itself low, indicating that the top- k feature set is unstable even within a language at this layer and sample size; we therefore report no drift rather than over-claiming. This is a notable qualification to the “SAE refusal features differ across languages” narrative at small scale.

4.4 Refusal metrics: keyword and self-judge failures

A keyword refusal classifier agreed with the external judge only at chance (Cohen’s $\kappa = 0.00$ English, 0.05 Hindi). It counts softened refusals as compliance and vice versa, and under it the intervention effect is misestimated entirely. Same-model self-judging fails in a different way. When we audited Gemma judging its own Vietnamese outputs, we found 74 cases where the self-judge called clear compliance a refusal, which pushes the Vietnamese refusal rate up artificially. Both motivate the external graded judge we use throughout.

5 Discussion

The apparent conflict between the dense and sparse pictures does not really hold up once you separate the two. The dense refusal direction is shared across English, Hindi, and Vietnamese in both senses that matter. It is shared geometrically, as the cosine and transfer-AUROC numbers show, and it is shared causally, since removing it lowers refusal in all three languages. The part we find most convincing is that this causal effect shows up in three different models, not just one, which is something a single-model study could never tell you. The sparse features behave differently. At this scale they do not give a clean drift signal, and the fact that they are already unstable within a single language suggests the top- k probe is just noisier than the dense direction. It is not that the features “move” from one language to another, it is that we cannot measure them reliably enough to say.

The causal effect is strongest in English and Vietnamese and weakest in Hindi, and it is weakest of all in the smaller models, where the model barely refuses in the non-English languages to begin with (Qwen, for instance, refuses only about a third of the Hindi prompts before any intervention).

That itself is telling that you can only cleanly remove refusal where refusal is firmly there in the first place. For multilingual safety the takeaway is straightforward. A single English-derived steering direction can pull down refusal across several languages, which is useful for red-teaming and a warning against assuming that each language’s safety behaves on its own.

6 Limitations

Causal claims concern the diff-in-means direction. A full replication would sweep token positions and add a coherence filter. Hindi is the weakest leg: small, noisy effects and low/uncertain baselines in the smaller models, with non-trivial “unclear” judge rates. SAE analysis is Gemma-only and the within-language baseline is low, so the SAE null should be revisited with more pairs, more layers, or larger SAEs before being treated as conclusive. NLLB translations, pending completion of the vietnamese native-speaker audit, may retain artifacts. The external judge, while better than keyword/self-judging, is not human-validated here. A human-labelled subset would harden the agreement claim.

7 Ethics and Broader Impact

This work studies how refusal works, and directional ablation is a jailbreak technique. We kept the harmful prompts to short requests of the kind already found in public refusal benchmarks, with no operational detail, we release no working harmful generations, and we treat the interventions as a way of probing the model rather than attacking it. Our main message, that safety in one language does not stand on its own and can be pulled down through another, argues against a false sense of security in multilingual deployment, and we think that is more protective than harmful to point out.

8 Conclusion

Across GEMMA-2-2B-IT, LLAMA-3.2-1B, and QWEN2.5-1.5B, the refusal direction transfers across English, Hindi, and Vietnamese both representationally and causally: ablating the English direction reduces refusal in all three languages, replicated across all three models. Sparse SAE features, by contrast, show no significant cross-lingual drift at this scale. We release code, the 142×3 minimal-pair dataset, and all artifacts.

Reproducibility. Single GPU. Pipeline: minimal pairs $\rightarrow$ direction/layer sweep $\rightarrow$ transfer-AUROC $\rightarrow$ SAE Jaccard vs. within-language baseline $\rightarrow$ judge-scored ablation (external graded judge) $\rightarrow$ causal selection $\rightarrow$ multi-model replication $\rightarrow$ bootstrap/Wilson CIs. Datasets, code, per-model judged outputs, CSVs, and figures included.

References

- A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda. Refusal in Language Models Is Mediated by a Single Direction. *arXiv:2406.11717*, 2024.
- X. Wang, M. Wang, Y. Liu, H. Schütze, B. Plank. Refusal Direction is Universal Across Safety-Aligned Languages. *arXiv:2505.17306*, 2025.

T. Lieberum et al. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *arXiv:2408.05147*, 2024.

NLLB Team. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv:2207.04672*, 2022.