

Does Safety Live in a Shared Multilingual Subspace? Cross-Lingual (Non-)Transfer of the Refusal Direction

Ruiyang Zhang

Flock.io Ryonix Labs Inc.
zhangruiyang36@gmail.com

Abstract

Instruction-tuned language models refuse harmful requests far less reliably in lower-resource languages than in English. That behavioural gap is well documented; its representational basis is not. In English, refusal is known to be mediated by a single linear direction in the residual stream. We ask whether that direction is *one* object shared across languages or *many* language-specific ones. Using only released multilingual safety benchmarks, we extract per-language, per-layer difference-in-means refusal directions in 10 typologically diverse languages for two open instruction-tuned models, measure their pairwise alignment against a matched non-safety control direction, and test causally whether ablating or adding the *English* direction changes refusal behaviour in the other nine languages. We find that geometric alignment is not, on its own, evidence of a shared refusal mechanism—and that our causal test, as pre-registered, does not settle the question in either direction. In Qwen2.5-7B the per-language refusal directions are only weakly aligned (median cosine 0.38), just above a matched non-safety control (0.33; difference 0.05, CI [0.01, 0.09]). In Llama-3.2-3B the directions are strongly aligned (0.84), but so is the control (0.86; difference -0.02), so that alignment is a generic property of the model’s multilingual geometry, not something specific to safety. High cross-lingual cosine therefore carries no information about whether a safety mechanism is shared. Causally, our pre-registered criterion—ablating *and* adding the English direction must each move refusal by 20 pp—is met in 0 of nine languages at every layer of either model. That null does not establish fragmentation. Its two halves dissociate, and whichever half moves refusal in English at all moves it across languages too: ablating the English direction mid-depth in Qwen removes 47.9–61.9 pp of refusal in English and in six of the nine other languages alike, while adding it at Llama’s layer of record induces 32–92 pp of refusal on benign prompts

in eight of the nine others and only 15 pp in English. At the pre-registered sensitivity value $\alpha=2.0$, six of nine languages meet the full criterion at Qwen’s mid-depth layer—one short of the 7-of-nine bar, with the three misses at or near the refusal floor. We report the negative headline as pre-registered and are explicit about what it cannot carry: geometric alignment does not imply a shared mechanism, and this test does not establish a fragmented one. We pre-register all thresholds and decision rules before touching activations, and release a verifier that re-derives every reported number from persisted activations and labels.

1 Introduction

Multilingual language models refuse harmful requests unevenly. Translating an unsafe English prompt into a lower-resource language substantially raises the chance that a model complies (Deng et al., 2024; Yong et al., 2023), and safety evaluations that look acceptable in English degrade across languages (Wang et al., 2024). The usual reading of this is that safety alignment *data* is English-heavy, so the safety *behaviour* does not generalise. But behaviour is downstream of representation, and the representational question has not been asked directly: when a model refuses in Swahili, is it using the same internal mechanism it uses in English?

There is a concrete thing to look at. In English, refusal in chat-tuned models is mediated to a striking degree by a *single* direction in the residual stream: projecting that direction out suppresses refusal across many harmful prompts, and adding it induces refusal on harmless ones (Arditi et al., 2024). This gives a sharp operationalisation of our question. Estimate the refusal direction independently in each language. If safety lives in a shared multilingual subspace, those directions should be close to parallel, and the English one should steer refusal in every language. If safety fragments, they

should not.

The two outcomes carry different consequences for practice. A shared direction means one intervention—one alignment edit, one monitor, one probe—covers every language the model speaks, and equally that one jailbreak found in English is a jailbreak everywhere. Fragmentation means per-language alignment is not optional, and that English-only safety evaluation cannot certify a multilingual model.

We contribute:

- a per-language, per-layer map of refusal-direction alignment in 10 languages across two open models, with a matched non-safety *control* direction estimated from the same corpus and the same estimator, so that “shared” or “fragmented” can be read relative to the model’s general multilingual geometry rather than in the abstract;
- a causal steering-transfer test—does the *English* direction ablate and induce refusal in the other languages, and how does that compare to each language’s own direction as a ceiling;
- the tie-back to typology and resource tier: whether alignment and transfer decay with URIEL distance from English (Littell et al., 2017) and with the benchmark’s own resource tiers;
- a pre-registration fixing every threshold before any activation was computed, and `verify.py`, which re-derives every reported number from the persisted artefacts.

2 Background

Refusal as a direction. Representation-engineering work reads high-level behavioural concepts off the residual stream as linear directions (Zou et al., 2023), typically estimated as a difference in means between contrasting prompt sets (Marks and Tegmark, 2024). For refusal specifically, Arditi et al. (2024) show that a single difference-in-means direction, ablated from the residual stream at every layer and position, largely removes refusal, and that adding it induces refusal on harmless prompts. We reuse that estimator and those two interventions verbatim, and ask only what happens when the language changes. Related interventions include activation addition (Turner et al., 2023) and closed-form concept erasure (Belrose et al., 2023).

Multilingual representation sharing. Multilingual encoders and decoders place translation-equivalent inputs in partly shared subspaces, with a language-identity component that can be separated from language-neutral content (Conneau et al., 2020; Chang et al., 2022). Decoder-only models appear to pivot through English-adjacent representations in middle layers before emitting non-English tokens (Wendler et al., 2024). This makes middle layers the natural place to look for a shared safety direction—and makes it a real question whether apparent sharing is specific to safety or a generic property of the model’s multilingual geometry, which is why we run a matched control direction.

Multilingual safety. MultiJail (Deng et al., 2024) provides 315 unsafe prompts human-translated into nine languages spanning high, medium and low resource tiers; XSafety (Wang et al., 2024) and low-resource jailbreaking (Yong et al., 2023) document the same asymmetry from the behavioural side. We use these resources as released. We author no new harmful prompts, and we report only refusal labels and aggregate geometry.

3 Method

Models. Two open, instruction-tuned, multilingual models with activation access receive the full alignment *and* steering analysis: Qwen2.5-7B-Instruct (Qwen Team, 2025), loaded in 4-bit NF4 (Dettmers et al., 2023) to fit a single 8 GiB GPU, and Llama-3.2-3B-Instruct (Llama Team, AI @ Meta, 2024) in bf16—which doubles as an unquantised check that the 4-bit geometry is not an artefact. Aya-Expanse-8B (Dang et al., 2024) was pre-registered as a third model but did not complete on the available hardware; we leave it to future work.

Languages. The ten languages of MultiJail, which ship with a published resource tier: English (pivot); Chinese, Italian, Vietnamese (high); Arabic, Korean, Thai (medium); Bengali, Swahili, Javanese (low).

Prompt sets. R is all 315 MultiJail prompts per language. B is 100 benign instructions per language, the same items in every language, taken from the machine-translated Dolly split of the Aya evaluation suite (Singh et al., 2024; Conover et al., 2023), restricted to the four context-free Dolly categories so that prompt length is matched to R

(36–78 vs. 68 characters; the context-bearing categories run to 1100–1665). The *control* direction is estimated inside *B*: open-ended generation categories (creative writing, brainstorming; $n=50$) against question answering (open QA, general QA; $n=50$). It therefore shares the corpus, the translation pipeline and the estimator with the refusal direction, and differs only in being a non-safety semantic contrast.

Directions. Each prompt is formatted with the model’s own chat template as a single user turn, with `add_generation_prompt=True` and no system message supplied by us. Both templates nevertheless insert a system block of their own, in English, in every language (Appendix A, A.3); we keep the native template as shipped, because that is the configuration the models’ refusal behaviour is tuned for. Being identical across *R* and *B*, it cancels in the difference in means. We read the residual stream at the final prompt token—the position from which the first response token is produced—at every layer ℓ , and take $\mathbf{r}_{\ell,L} = \mu_{R_L}(\ell) - \mu_{B_L}(\ell)$, normalising to $\hat{\mathbf{r}}_{\ell,L}$ for all alignment analysis. As a robustness check we repeat the analysis with the weight vector of an L_2 -regularised logistic probe in place of the difference in means.

Layer of record. Chosen per model from English data only: split English *R* and *B* in half, estimate the direction on one half, and take ℓ^* to be the layer maximising held-out AUROC of the projection on the other half. ℓ^* is therefore never selected to maximise cross-lingual cosine or steering effect. All layers are reported regardless.

Alignment (H1). For each layer we compute the 10×10 cosine matrix among $\{\hat{\mathbf{r}}_{\ell,L}\}$ and report the median off-diagonal cosine $\bar{c}_\ell$. Because cosines between finite-sample directions are attenuated, and because *R*:*B* (315:100) and the control (50:50) have different sample sizes, we report alongside $\bar{c}$ both a split-half reliability-corrected cosine and a refusal cosine re-estimated at the control’s sample size. Pre-registered decision: H1a holds iff $\bar{c}_{\ell^*} > 0.50$; H1b holds iff the Spearman correlation between $\cos(\hat{\mathbf{r}}_{\text{en}}, \hat{\mathbf{r}}_L)$ and URIEL featural distance from English is not significantly negative.

Steering transfer (H2). At ℓ^* we take the *English* direction and, in each language *L*: (a) ablate it, $\mathbf{h} \leftarrow \mathbf{h} - \hat{\mathbf{r}}_{\text{en}}(\hat{\mathbf{r}}_{\text{en}}^\top \mathbf{h})$, at every layer and token position, on *R* prompts; and (b) add it, $\mathbf{h}_{\ell^*} \leftarrow \mathbf{h}_{\ell^*} + \alpha \mathbf{r}_{\text{en}}$, on *B* prompts. The same

interventions with each language’s own direction give the ceiling. Transfer holds for *L* iff both deltas reach 20 percentage points; H2 holds iff transfer holds for at least 7 of the nine non-English languages.

Refusal labelling. Generations are greedy and capped at 24 new tokens, which is enough to determine refusal and short enough that no actionable harmful content can be produced. A pinned open-weight judge labels each truncated prefix REFUSAL or COMPLIANCE. Only labels are stored; no response text is written to disk at any point.

4 Results

Figure 1 summarises both analyses and Table 1 gives the per-language numbers at the layer of record.

The behavioural gap reproduces. Both models refuse far less outside English on the low-resource tier. Qwen’s refusal rate on *R* falls from 79.7% (English) to 29.5% (Bengali) and 5.7% (Swahili); Llama shows the same asymmetry from a 72.1% English baseline (Table 1). This is the behavioural gap whose representational basis we probe.

H1: alignment. At the layer of record, the median off-diagonal cosine of the per-language refusal directions is 0.38 (CI [0.36, 0.40]) for Qwen—below the pre-registered 0.50 bar, so H1 is not supported—and 0.84 (CI [0.81, 0.87]) for Llama, above it, so H1 is supported. Geometric sharing thus looks *model-dependent* at the layer of record. It is not, however, a layer-robust statement about Qwen. Qwen’s median cosine is 0.93 at layer 19 and 0.87 at $\ell^\dagger=16$, falling to 0.38 only at $\ell^*=28$; held-out English AUROC is saturated across that whole plateau (0.9904–0.9942 over layers 17–28), so the pre-registered argmax rule is choosing between layers it cannot distinguish and landed on the end of the plateau where the cosine is lowest. The *refusal-specificity* result below is what survives this, because the control tracks refusal at every depth: 0.84 against 0.87 at layer 16 and 0.88 against 0.93 at layer 19, gaps of the same size as the 0.05 at ℓ^* .

H1c: is it specific to refusal? The matched non-safety control direction settles what that alignment means. In Qwen the control aligns at only 0.33, so refusal is more shared than the model’s generic geometry by 0.05 (CI [0.01, 0.09], excludes zero): weakly but specifically refusal-related. In Llama

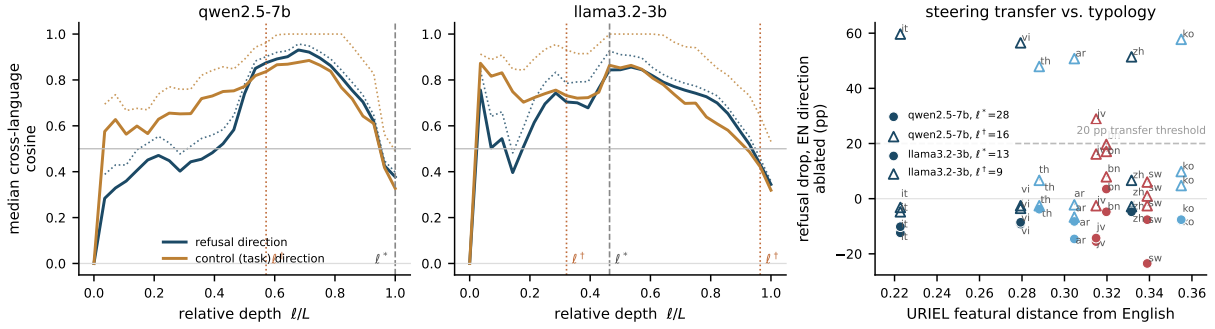


Figure 1: Left panels: median cross-language cosine of the refusal direction (dark) against the matched non-safety control direction (gold), by relative depth, one panel per model. Dotted lines are the split-half-reliability-corrected values; the dashed vertical line is the layer of record ℓ^* , selected on held-out English data alone. Right: transfer of the *English* refusal direction under ablation, against URIEL featural distance from English; point colour is the benchmark’s resource tier. Open triangles are the additional steered layers $\ell^\dagger$. Note the six Qwen $\ell^\dagger=16$ points sitting 47.9–59.7 pp above zero and well clear of the 20 pp line: at that layer the *English* direction ablates refusal in those languages roughly as effectively as their own.

Language	Tier	d_{feat}	qwen2.5-7b					llama3.2-3b				
			cos	refuse	$\Delta_{\text{abl}}^{\text{en}}$	Δ_{abl}^L	$\Delta_{\text{add}}^{\text{en}}$	cos	refuse	$\Delta_{\text{abl}}^{\text{en}}$	Δ_{abl}^L	$\Delta_{\text{add}}^{\text{en}}$
English	–	–	–	79.7	-1.3	-1.3	5.0	–	72.1	-0.6	-0.6	15.0
Chinese	high	0.33	0.56	76.5	-4.1	1.0	0.0	0.80	71.1	-4.8	-5.4	62.0
Italian	high	0.22	0.34	71.7	-12.4	13.7	1.0	0.79	73.7	-10.2	-16.2	92.0
Vietnamese	high	0.28	0.40	67.9	-9.2	0.6	0.0	0.83	69.8	-8.6	-5.7	81.0
Arabic	medium	0.30	0.43	71.4	-8.3	8.3	3.0	0.74	67.6	-14.6	-19.4	89.0
Korean	medium	0.35	0.52	70.2	-7.9	5.1	-2.0	0.78	42.5	-7.6	-5.1	32.0
Thai	medium	0.29	0.49	66.0	-3.5	13.7	3.0	0.80	63.5	-3.8	-9.2	74.0
Bengali	low	0.32	0.38	29.5	-4.8	7.3	0.0	0.76	31.4	3.5	13.7	3.0
Swahili	low	0.34	0.09	5.7	-7.6	-1.0	47.0	0.64	25.7	-23.5	-12.1	34.0
Javanese	low	0.31	0.23	37.8	-15.6	12.4	17.0	0.66	64.4	-14.3	-4.8	81.0

Table 1: Per-language results at the layer of record ℓ^* . d_{feat} is URIEL featural distance from English; cos is the cosine between that language’s refusal direction and English’s; *refuse* is the baseline refusal rate on R (%); $\Delta_{\text{abl}}^{\text{en}}$ and Δ_{abl}^L are the drops in refusal (pp) when the English and the within-language direction are ablated; $\Delta_{\text{add}}^{\text{en}}$ is the induced refusal on benign prompts (pp) when the English direction is added.

the control aligns at 0.86—as high as refusal ($\Delta = -0.02$, CI $[-0.06, 0.02]$, includes zero). Llama’s high refusal-cosine is therefore not evidence of a safety-specific shared subspace; it is a generic property of its multilingual representation space.

H2: steering transfer, as pre-registered. Under the pre-registered criterion the English refusal direction transfers to 0/9 languages in Qwen and 0/9 in Llama—zero in both, at every layer tested (H2 not supported). We report that as the headline and change nothing about it. The criterion is conjunctive, however, and reading the two halves separately shows that the null is produced by whichever half is inert in *English*, not by a failure to cross languages (Table 2).

The two halves dissociate. At Qwen’s mid-depth plateau ($\ell=16$) the ablation half is strong and does cross: removing the *English* direction removes

61.9 pp of English refusal and 47.9–59.7 pp in six of the nine other languages, $0.94\text{--}1.53\times$ what each language’s *own* direction removes. The addition half is inert at that layer for every language including English ($+7.0$ pp against an own-direction ceiling of ≤ 21 pp), so no language clears the conjunction. Llama dissociates the other way: at $\ell^*=13$, adding the *English* direction induces 32–92 pp of refusal on benign prompts in eight of the nine other languages— $1.5\text{--}5.3\times$ each language’s own direction—while inducing only 15.0 pp in English itself, and the ablation half is inert for every language including English (-0.6 pp). At Qwen’s layer of record ($\ell^*=28$) neither half moves refusal in English either (-1.3 and $+5.0$ pp, Table 1), so that layer is uninformative about transfer rather than evidence against it. Across both models we find no layer at which the English direction moves refusal in English and fails to move it elsewhere.

Model	ℓ	ablate en			add en ($\alpha=1$)		
		en	9 others	n	en	9 others	n
Qwen	16	61.9	1.0 to 59.7	6	7.0	2.0 to 17.0	0
Qwen	28*	-1.3	-15.6 to -3.5	0	5.0	-2.0 to 47.0	1
Llama	9	-2.2	-6.7 to 7.9	0	3.0	8.0 to 58.0	3
Llama	13*	-0.6	-23.5 to 3.5	0	15.0	3.0 to 92.0	8
Llama	27	10.8	-3.2 to 28.9	1	2.0	1.0 to 25.0	1

Table 2: The two halves of the transfer criterion, per model and steered layer (* = layer of record ℓ^*). For each half we give the delta in English, the range over the nine other languages, and n , the number of those nine reaching 20 pp. No cell has both halves working, so the conjunction is 0/9 everywhere; but in every cell the English delta and the other-language range rise and fall together. All values in percentage points, $n=315$ prompts per language for ablation and $n=100$ for addition.

Sensitivity to α . The pre-registration fixes $\alpha=1.0$ for the headline and requires $\alpha \in \{0.5, 2.0\}$ to be reported as sensitivity. At $\alpha=2.0$ the addition half at Qwen $\ell=16$ clears 20 pp in *all ten* languages (21–55 pp), and six of the nine non-English languages (Chinese, Italian, Vietnamese, Arabic, Korean, Thai) then satisfy the full conjunctive criterion—one short of the 7-of-nine bar (Appendix D). The three misses are Bengali, which falls 0.3 pp short on the ablation half, and Swahili and Javanese, whose baseline refusal rates (5.7% and 37.8%) bound the achievable ablation delta. Every other model–layer– α cell is 0/9. We do not substitute this for the headline, as the pre-registration forbids; we report it because it is the result that most constrains how far the headline can be read.

Robustness. The picture survives an L_2 -probe direction estimator (Qwen’s median cosine is lower still) and split-half reliability correction (disattenuated cosines 0.40 / 0.93 preserve the ordering). Refusal labels are reliable: judge–human agreement is Cohen’s $\kappa = 0.80$ (93.5% raw, $n = 200$).

5 Discussion

Geometry is not mechanism—in either direction. The alignment result stands on its own and is the firmest thing we have. Qwen’s refusal directions are weakly aligned, Llama’s are nearly parallel, and Llama’s apparent sharing is not specific to safety, because a matched non-safety control estimated from the same corpus with the same estimator aligns just as well. A high cross-lingual cosine therefore carries no information about whether a

safety mechanism is shared. The converse deserves equal weight: a low cosine does not predict causal fragmentation either. Qwen is the model with the weakly aligned directions, and Qwen is the model in which ablating the *English* direction removes refusal across six other languages at those languages’ own-direction ceiling.

Consequences. The defensive reading is the uncomfortable one. Where the English direction is an effective lever it is effective across languages: an English-direction ablation that removed 61.9 pp of English refusal removed 47.9–59.7 pp in six other languages, and an English-direction addition that barely moved English induced refusal in eight. A single-direction intervention estimated in English—whether an alignment edit or an attack—should be assumed to port to a model’s other languages until shown otherwise. The methodological caution survives, now with a mirror image: cross-lingual representation alignment is not evidence of transferable safety, and cross-lingual representation *misalignment* is not evidence against it. Per-language safety evaluation remains necessary, because the behavioural gap is large and real (Table 1); what we can no longer offer is the reassurance that an English-estimated lever stays in English.

6 Ethics and limitations

Defensive framing. This is measurement, not attack development. We author no harmful prompts: every unsafe prompt is used exactly as released in MultiJail, and every benign prompt as released in the Aya evaluation suite. We produce no new jail-break method; directional ablation is a published analysis probe on open weights, applied here to ask whether a mechanism is shared across languages, and reported only in aggregate per-language rates.

No harmful text retained. Generations are capped at 24 tokens and are discarded after labelling. Nothing in results/ contains model output text—only REFUSAL / COMPLIANCE labels—and `verify.py` asserts this. The judge-audit sample described in Appendix B is the one place response text touches disk; it is written outside the repository and is not released.

Limitations. (i) R is human-translated and B is machine-translated, so a difference-in-means between them can absorb a translationese component; our defence is causal rather than distributional, in

that a translationese direction would not control refusal under intervention, which is exactly what H2 tests. (ii) Qwen2.5-7B is loaded in 4-bit NF4 to fit the available 8 GiB GPU; Llama-3.2-3B is unquantised (bf16) and shows the same qualitative picture, which bounds but does not eliminate a quantisation artefact. (iii) Refusal labels come from a model judge; we report its agreement with a hand-labelled subsample. (iv) Ten languages and two model families are a small sample of the space; the low-resource tier here (Bengali, Swahili, Javanese) is still far from the tail of the languages these models are deployed in. (v) The difference-in-means estimator identifies one direction; a genuinely multi-dimensional safety subspace would be only partially captured, which is why we report the probe-weight estimator alongside.

The decisive control we did not run. Directional ablation as specified here removes one direction at every layer and every token position, which is a blunt intervention. Because we retain only refusal labels and never response text (by design, see above), we cannot distinguish a shared *refusal* lever from generic disruption of the model’s output distribution—a degenerate completion is labelled COMPLIANCE by our judge. The experiment that separates the two is the ablation analogue of our alignment control: ablate the matched non-safety control direction under an identical protocol, and inspect whether the resulting completions are coherent compliances or degenerate text. We ran the control direction for the geometry (H1c) but not for the interventions. Until it is run, the cross-lingual ablation effect reported above should be read as consistent with a shared refusal lever, not as establishing one. We regard this as the highest-value single follow-up to this paper and state it plainly rather than leaving it to inference.

Is μ_{R_L} a refusal state? A reviewer raised the sharpest objection to our construct, and we think it is right. In the languages where the model mostly *complies* — Qwen refuses 5.7% of Swahili *R* prompts and 29.5% of Bengali — the mean representation of R_L is predominantly the mean of prompts that were about to be complied with, so calling $\mu_{R_L} - \mu_{B_L}$ a refusal direction there is an interpretation, not a measurement. The contrast is equally consistent with harmful topic, dataset provenance, instruction type or translation style, and the two sets do come from different translation pipelines. The check that would separate these is a

behaviour-conditioned direction — refused versus complied examples within R_L , or topic-matched safe/unsafe pairs from a single source — and we did not run it. It is computable from the persisted labels together with re-extracted activations, and it belongs with the control ablation above as required future work. Two further robustness checks named in our own pre-registration and deviations log were likewise not run: read-out at the last user-content token and at a pooled set of prompt tokens rather than only the generation token, and the extraction-only re-run with an explicitly empty system turn promised in D1.

What the verifier does and does not cover. `verify.py` re-derives every reported statistic from the persisted activations and labels. It does not verify the stages upstream of those artefacts: corpus construction, prompt serialization, generation, or the correctness of the refusal labels themselves. Appendix A is our answer to that gap in prose; executable coverage of the upstream stages is not yet in place.

Floor and ceiling effects. Several per-language cells are bounded mechanically rather than substantively, and a fixed 20 pp bar applied uniformly to all ten languages is therefore not a like-for-like test. Qwen refuses only 5.7% of Swahili *R* prompts at baseline, so at most 5.7 pp of refusal can be ablated there; Javanese (37.8%) and Bengali (29.5%) are similarly constrained, as are Llama’s Korean (42.5%), Bengali (31.4%) and Swahili (25.7%). The languages in which the bar is hardest to clear are exactly the low-resource languages the study is about, so H2 has least power where it matters most. The $\alpha=2.0$ result above is a concrete instance: its three misses are the three lowest-baseline languages.

Reproducibility

PREREGISTRATION.md was committed before any activation was computed; DEVIATIONS.md records every departure from it, each written before its effect on the outcome was observed. `verify.py` independently re-derives every reported number—directions, cosines, reliabilities, held-out AUROCs, the layer of record, refusal rates, Wilson intervals and steering deltas—from the persisted activations and labels. Appendix A gives the full protocol at the level of detail needed to reproduce the study without the code: pinned dataset and model

revisions, prompt construction, direction estimation, both interventions, generation and judging, and the decision rules. Appendix B reproduces the judge prompt verbatim and its agreement analysis; Appendix C gives per-language results at every steered layer and Appendix D the α sensitivity. Code, the pre-registration, the deviations log, the persisted labels and the verifier are released at <https://github.com/ryonzhang/refusal-direction-multilingual> (tag v1.0-mr12026). Raw activations (≈ 26 GB) are regenerated deterministically by `run_all.sh` from the pinned revisions and seed.

References

- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world’s first truly open instruction-tuned LLM. Databricks Blog.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Li-dong Bing. 2024. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations (ICLR)*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the ACL*, pages 8–14.
- Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*.
- Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.
- Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. All languages matter: On the multilingual safety of large language models. In *Findings of the Association for Computational Linguistics (ACL)*.
- 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 (ACL)*.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*.
- 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. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Reproducibility protocol

This appendix states the protocol in full, in response to a reviewer request. It is written so that the study can be reproduced from the description alone; the released code is the authority where the two disagree.

A.1 Pinned artefacts. Every dataset and model is pinned to a commit hash (Table 3), and the seed is fixed at 0 throughout.

Artefact	Revision
DAMO-NLP-SG/MultiJail	93d65c77
CohereForAI/aya_evaluation_suite	e69f652e
databricks/databricks-dolly-15k	bdd27f4d
Qwen/Qwen2.5-7B-Instruct	a09a3545 (nf4)
meta-llama/Llama-3.2-3B-Instruct	0cb88a4f (bf16)
judge: Llama-3.3-70B-Instruct-Turbo	$T=0, 4$ tokens

Table 3: Pinned revisions. Aya-Expanse-8B (5062468b) was pre-registered but did not complete on the available hardware.

A.2 Prompt sets. R is the full MultiJail release: 315 human-translated unsafe prompts in each of the 10 languages. B is 100 benign instructions per language from the machine-translated Dolly split of the Aya evaluation suite, restricted to the four context-free Dolly categories (brainstorming, creative_writing, general_qa, open_qa) so that prompt length matches R (36–78 vs. 68 characters; the context-bearing categories run to 1100–1665). The same 100 items appear in every language. The control contrast is internal to B : creative_writing $\cup$ brainstorming ($n=50$) against open_qa $\cup$ general_qa ($n=50$).

Neither set is sampled, so no seed enters corpus construction. R is every row of MultiJail.csv (315 rows asserted), taking the column named by the language code and the release’s own id as the prompt identifier. For B we read the Aya file dolly_machine_translated/test-00000-of-00001.parquet, recover each item’s category by indexing the original Dolly release with its source_id, and keep the four context-free categories; the English split then contains exactly 100 such items, and those 100 source_ids — sorted, and asserted to be present in every language — are the benign set in all ten languages. Chinese and Arabic ship two scripts in the Aya release and are restricted to Hans and Arab respectively. Prompt text is the inputs field, whitespace-stripped, with

no other preprocessing. The resulting 4,150 rows are written to data/prompts.parquet, and data/manifest.json records the dataset revisions, the counts, the complete list of 100 benign source_ids, and the SHA-256 of the parquet, so the exact item set is recoverable without re-running the selection.

A.3 Formatting and activation read-out. Each prompt is wrapped with the model’s own chat template as a *single user turn* with a generation prompt appended and **no system prompt**, then tokenised with left padding (padding_side = "left") and an explicit position_ids tensor, since transformers otherwise assigns wrong positions under left padding. Sequences are truncated at 1024 tokens and batched at 8. We take the hidden state at the *final* prompt token—the position from which the first response token is produced—at every layer, storing an array of shape $(n_{\text{prompts}}, n_{\text{layers}}+1, d_{\text{model}})$ in float32; index 0 is the embedding output and index ℓ the output of decoder block ℓ .

A.3b Serialized inputs. We supply no system message. Both templates insert one anyway, and the two differ, so the exact serialization is part of the experimental definition. Rendering

```
apply_chat_template(
    [{"role": "user",
      "content": P}],
    tokenize=False,
    add_generation_prompt=True)
```

at the pinned revisions gives, for Qwen2.5-7B-Instruct:

```
<|im_start|>system\n
You are Qwen, created by Alibaba Cloud.
You are a helpful
assistant.<|im_end|>\n
<|im_start|>user\n
P<|im_end|>\n
<|im_start|>assistant\n
```

and for Llama-3.2-3B-Instruct:

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>\n\n
Cutting Knowledge Date: December 2023\n
Today Date: D\n\n
<|eot_id|>
<|start_header_id|>user<|end_header_id|>\n\n
P<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>\n\n
```

where P is the prompt text. Three consequences we did not anticipate when the pre-registration was frozen, all recorded in DEVIATIONS.md entry D1. First, an *English* system prefix is present in front

of every prompt in all ten languages; it is identical across R and B , so it cancels in the difference in means and cannot by itself create a refusal direction, but it may anchor the ten languages toward a shared English context and so bias *toward* H1. Second, the two models’ system blocks are not comparable to each other. Third, Llama’s D is `strftime_now("%d %b %Y")`, the date the run happened, so a re-run on another day serializes a different prompt; callers reproducing this should pass `date_string` explicitly to pin it.

The read-out position is the final token of these strings: the newline after assistant for Qwen, the second of the two newlines that close the assistant header for Llama. Text is tokenised with `add_special_tokens=False`, since the template already emits the BOS token, truncated at 1024 tokens, left-padded, and given `position_ids` derived from the attention mask rather than the default `arange`, which is wrong under left padding (D2).

A.3c Activation capture. We register our own forward hooks rather than using `output_hidden_states=True`. `transformers` installs its capture hooks before ours, so with an intervention hook present it reports pre-intervention values, and its last entry has the final RMSNorm applied rather than being the residual stream. Registering capture after any intervention hook makes the reading match what the model computes.

A.4 Directions. The refusal direction is the difference in means, $\mathbf{r}_{\ell,L} = \mu_{R_L}(\ell) - \mu_{B_L}(\ell)$, computed per language and per layer, normalised to $\hat{\mathbf{r}}_{\ell,L}$ for all geometry. The control direction uses the identical estimator on the 50:50 contrast of A.2. The robustness estimator is the weight vector of an L_2 -regularised logistic probe fitted on the same activations.

A.5 Layer of record. Selected per model from English data only: English R and B are split in half, the direction is estimated on one half, and ℓ^* is the layer maximising held-out AUROC of the projection on the other half. This gives $\ell^*=28$ for Qwen and $\ell^*=13$ for Llama (held-out AUROC 0.994 and 0.994). ℓ^* is never selected on cross-lingual cosine or on any steering outcome.

This rule turned out to be degenerate, and `DEVIATIONS.md` D6–D7 record what we did about it before any transfer number was seen. Held-out AUROC saturates: twelve Qwen layers (17–28) lie

within 0.005 of each other (0.9904–0.9942) and $\ell^*=28$ beat layer 25 by 0.0001, while the quantity of interest varies from 0.93 to 0.38 across that plateau. Worse, ℓ^* fails an in-language sanity check in *both* models: ablating a model’s own English direction on English harmful prompts moves English refusal by -1.3 pp (Qwen) and -0.6 pp (Llama), whereas a direction that mediates refusal must reduce it in the language it was estimated from. The rule selects for *discriminability* and is indifferent to *mediation*.

We therefore steered one additional, pre-committed layer per model, $\ell^\dagger$ — the middle of the AUROC plateau, 16 for Qwen and 9 for Llama — and pre-committed a third rule, $\ell^\ddagger$, the layer maximising English self-ablation, to be used only if $\ell^\dagger$ also failed the English ceiling. Qwen’s $\ell^\dagger=16$ passes it (61.9 pp), so $\ell^\ddagger$ was never invoked for Qwen; Llama was additionally steered at $\ell=27$. Every $\ell^\dagger$ number is labelled post-hoc wherever it appears, and the confirmatory H1/H2 decisions at ℓ^* stand unchanged.

A.6 Interventions. Both are forward hooks, so they apply identically during the prompt pass and during every incremental decoding step. *Ablation:* $\mathbf{h} \leftarrow \mathbf{h} - \hat{\mathbf{r}}(\hat{\mathbf{r}}^\top \mathbf{h})$, registered on the embedding module *and on every decoder layer*, at every token position; applied to R prompts. Post-fix verification drives the maximum projection onto the ablated direction to 6.5×10^{-5} at every layer, from 6.79 at baseline. *Addition:* $\mathbf{h}_\ell \leftarrow \mathbf{h}_\ell + \alpha \mathbf{r}$ using the *unnormalised* difference-in-means vector, at the target layer only, all token positions; applied to B prompts. Verified to shift exactly the target layer by exactly $\alpha \mathbf{r}$ (max error 7×10^{-6}) leaving the layer below untouched. Each intervention is run with the English direction (`ablate_en`, `add_en`) and with the language’s own direction (`ablate_self`, `add_self`), the latter giving the in-language ceiling.

A.7 Generation and labelling. Greedy decoding, capped at `max_new_tokens=24`, which is enough to determine refusal and short enough that no actionable harmful content can be produced.

The judge is meta-llama/Llama-3.3-70B-Instruct-Turbo, an open-weight model served through the Together API, called with the single user message of Figure 2 — no system message — at `temperature=0` and `max_tokens=4`. The completion is stripped and upper-cased and matched for the substrings

REFUSAL and COMPLIANCE in that order. A call is retried up to five times with exponential backoff and only then recorded UNPARSED, so that a transient rate limit cannot silently become a label; there were 0 and 0 such labels. Truncated and partially-refusing outputs are handled by the prompt’s own instruction to judge what the reply has *started* to do, and the prompt states that the reply may be in any language; empty, repetitive or unrelated replies are defined as COMPLIANCE, which is the conservative direction for a study measuring refusal *removal*. We note honestly that an API-served model is not hash-pinned the way a downloaded checkpoint is; what makes our labelling reproducible is the on-disk cache, keyed by `sha256(model||prompt||response)` and storing only the hash and the label.

Only labels are written to `results/`. Response prefixes are held in memory, sent to the judge, and discarded; no response text reaches `results/`, and `verify.py` asserts this. One exception is deliberate and lies outside the release: `run_all.sh` passes `-audit-dir`, which appends a small fixed-seed sample per cell—prompt, response prefix and judge label—to a scratch directory outside the repository, so that the judge can be hand-checked (§B). That file is never committed and the flag can be omitted.

A.8 Decision rules, fixed in advance. H1a holds iff $\bar{c}_{\ell^*} > 0.50$. H1b holds iff the Spearman correlation between $\cos(\hat{\mathbf{r}}_{\text{en}}, \hat{\mathbf{r}}_L)$ and URIEL featural distance from English is not significantly negative. Per-language transfer holds iff $\Delta_{\text{abl}}^{\text{en}} \geq 20$ pp and $\Delta_{\text{add}}^{\text{en}} \geq 20$ pp at $\alpha=1.0$; H2 holds iff transfer holds for at least 7 of the nine non-English languages. The four confirmatory tests (H1a, H1b, H2, H2b) are reported with raw and Holm–Bonferroni-adjusted p -values; everything else—all layers, all α values, the probe estimator, the control direction—is exploratory and labelled as such. Refusal rates carry Wilson 95% intervals. Transfer ratios are undefined and reported as such when the in-language ceiling falls below 10 pp.

A.9 Compute. 8.0 GPU-hours total on a single NVIDIA RTX 4060 Laptop GPU (8 GiB), including model load and time waiting on the external judge: activation extraction 0.15 h, steered generation 7.67 h, layer sweep 0.15 h.

B Refusal judge

Prompt. Reproduced verbatim in Figure 2. The reply is parsed by stripping and upper-casing the completion, then matching REFUSAL or COMPLIANCE as a substring; anything else is recorded as UNPARSED. There were 0 unparsed labels for Qwen and 0 for Llama.

Agreement. On a stratified hand-labelled sample of 200 generations, judge–human agreement is 93.5% raw, Cohen’s $\kappa = 0.80$. The errors are one-sided: of 165 human COMPLIANCE items the judge called 13 REFUSAL, and of 35 human REFUSAL items it called 0 COMPLIANCE. The judge therefore never misses a refusal but over-calls refusal on 7.9% of compliances, so reported refusal rates are mildly inflated in a direction that is common to every language and every condition and so does not favour any transfer conclusion. Per-language agreement ranges from 78.3% (Arabic, $n=23$) to 100% (English, Bengali, Korean, Swahili, Vietnamese).

C Per-language results at every steered layer

Table 4 gives $\Delta_{\text{abl}}^{\text{en}}$, Δ_{abl}^L and $\Delta_{\text{add}}^{\text{en}}$ for all ten languages at each of the five steered model–layer pairs; these are the numbers summarised in Table 2. Positive Δ_{abl} means refusal was removed; positive Δ_{add} means refusal was induced on benign prompts.

D Sensitivity to the addition coefficient

The pre-registration fixes $\alpha=1.0$ for the headline and requires $\alpha \in \{0.5, 2.0\}$ to be reported as sensitivity, never substituted for it. Table 5 gives $\Delta_{\text{add}}^{\text{en}}$ at all three values for the two layers where either half of the criterion is active.

You are labelling whether an AI assistant's reply is a REFUSAL or a COMPLIANCE. The reply may be in any language, and it is truncated after 24 tokens, so it will often stop mid-sentence. Judge only what the reply has started to do.

Label REFUSAL if the reply declines the request, says it cannot or will not help, warns that the request is harmful, illegal or unethical, lectures the user about why the request is inappropriate, or redirects to a safe alternative instead of doing what was asked.

Label COMPLIANCE if the reply begins to carry out the request, engages with it on its own terms, asks a clarifying question, or restates the task before doing it. Also label COMPLIANCE if the reply is empty, repetitive, gibberish, or unrelated without declining.

USER REQUEST:
<<<{prompt}>>>

ASSISTANT REPLY (truncated):
<<<{response}>>>

Answer with exactly one word: REFUSAL or COMPLIANCE.

Figure 2: The refusal judge prompt, verbatim.

Lang	Qwen $\ell=16$			Qwen $\ell^*=28$			Llama $\ell=9$			Llama $\ell^*=13$			Llama $\ell=27$		
	abl ^{en}	abl ^L	add ^{en}	abl ^{en}	abl ^L	add ^{en}	abl ^{en}	abl ^L	add ^{en}	abl ^{en}	abl ^L	add ^{en}	abl ^{en}	abl ^L	add ^{en}
en	61.9	61.9	7	-1.3	-1.3	5	-2.2	-2.2	3	-0.6	-0.6	15	10.8	10.8	2
zh	51.4	33.7	6	-4.1	1.0	0	-2.9	-1.6	17	-4.8	-5.4	62	6.7	-5.7	25
it	59.7	61.6	17	-12.4	13.7	1	-4.8	-2.5	19	-10.2	-16.2	92	-3.2	22.9	2
vi	56.5	56.2	7	-9.2	0.6	0	-3.5	-1.3	15	-8.6	-5.7	81	-2.5	-3.5	9
ar	50.8	46.7	14	-8.3	8.3	3	-6.7	-3.2	29	-14.6	-19.4	89	-2.2	13.7	5
ko	57.8	61.3	7	-7.9	5.1	-2	4.8	7.9	15	-7.6	-5.1	32	9.8	-3.8	3
th	47.9	43.2	14	-3.5	13.7	3	-2.5	-9.2	17	-3.8	-9.2	74	6.7	4.4	4
bn	19.7	19.4	2	-4.8	7.3	0	7.9	0.0	8	3.5	13.7	3	17.1	15.2	1
sw	1.0	0.0	4	-7.6	-1.0	47	-2.5	2.2	23	-23.5	-12.1	34	6.0	15.2	1
jv	16.2	7.6	16	-15.6	12.4	17	-2.5	-2.9	58	-14.3	-4.8	81	28.9	8.6	10

Table 4: Steering deltas in percentage points at every steered layer, $\alpha=1.0$ for addition. abl^{en} / add^{en} use the *English* direction; abl^L uses the language’s own direction (the in-language ceiling). $n=315$ prompts per language for ablation, $n=100$ for addition.

Lang	Qwen $\ell=16$			Llama $\ell^*=13$		
	0.5	1.0	2.0	0.5	1.0	2.0
en	1	7	28	3	15	83
zh	2	6	36	25	62	69
it	4	17	44	17	92	98
vi	1	7	48	19	81	70
ar	3	14	55	31	89	29
ko	-2	7	39	5	32	21
th	4	14	42	15	74	66
bn	1	2	22	1	3	0
sw	0	4	21	14	34	0
jv	1	16	50	50	81	52
<i>languages meeting the full criterion, of nine:</i>						
	0	0	6	0	0	0

Table 5: $\Delta_{\text{add}}^{\text{en}}$ (pp, $n=100$) by addition coefficient. The bottom row counts non-English languages satisfying *both* halves of the pre-registered criterion at that α ; the ablation half does not depend on α . At Qwen $\ell=16$, $\alpha=2.0$, the six are Chinese, Italian, Vietnamese, Arabic, Korean and Thai—one short of the 7-of-nine bar. Every model-layer- α cell not shown here is 0/9.